

## Regression Models for Ordinal Data

By PETER McCULLAGH

*University of Chicago, Chicago, Illinois 60637, U.S.A.†*

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, February 13th, 1980, Professor D. R. Cox in the Chair]

### SUMMARY

A general class of regression models for ordinal data is developed and discussed. These models utilize the ordinal nature of the data by describing various modes of stochastic ordering and this eliminates the need for assigning scores or otherwise assuming cardinality instead of ordinality. Two models in particular, the proportional odds and the proportional hazards models are likely to be most useful in practice because of the simplicity of their interpretation. These linear models are shown to be multivariate extensions of generalized linear models. Extensions to non-linear models are discussed and it is shown that even here the method of iteratively reweighted least squares converges to the maximum likelihood estimate, a property which greatly simplifies the necessary computation. Applications are discussed with the aid of examples.

**Keywords :** COMPLEMENTARY LOG–LOG TRANSFORM; GENERALIZED EMPIRICAL LOGIT TRANSFORM; LINK FUNCTION; LOCATION PARAMETER; LOG–LINEAR MODEL; LOGIT LINEAR MODEL; MULTIVARIATE GENERALIZED LINEAR MODEL; ORDERED CATEGORIES; PROPORTIONAL HAZARDS; PROPORTIONAL ODDS; SCALE PARAMETER; SCORES; SURVIVOR FUNCTION

### 1. INTRODUCTION

It is widely recognized that the types of data as well as the class of problems that a statistician is likely to encounter vary greatly with the field of research. Consequently, methods that are useful in one area or discipline may be of little use or interest to researchers in another area. In the physical sciences, for example, the overwhelming proportion of data is essentially quantitative although possibly measured on an arbitrary scale. In the social sciences and to a lesser extent in the biological sciences, qualitative data are more common. These qualitative measurements, whether subjective or objective, usually take values in a limited set of categories which may be on an ordinal or on a purely nominal scale. Intermediate types of scales are also possible but the purpose of this paper is to investigate structural models appropriate to measurements on a purely ordinal scale.

For a discussion of the classification of scale types and their relevance to the statistical procedures employed, see Stevens (1951, 1958, 1968) who distinguishes nominal, ordinal, interval and ratio scales. Even this list however is incomplete since only a partial order may exist among the categories. More complex order structures arise when a bivariate response is observed, the categories for each margin being ordinal. One possibility investigated by Anscombe (1970) for modelling bivariate ordinal responses is to develop models based on the so-called cross-ratio distributions (Pearson, 1913; Plackett, 1965). This paper, however, is devoted solely to the case where there is a single response measured on an ordinal scale, there being possibly multiple explanatory factors or covariates.

Motivation for the proposed models is provided by appeal to the existence of an underlying continuous and perhaps unobservable random variable. In bioassay this latent variable usually corresponds to a “tolerance” which is assumed to have a continuous distribution in the population. Tolerances themselves are not directly observable but increasing tolerance is manifest through an increase in the probability of survival. The categories are envisaged as contiguous intervals on the continuous scale, the points of division being denoted in this paper

† Present address : Department of Mathematics, Imperial College, London.

by  $\theta_1, \dots, \theta_{k-1}$ . In many cases where, for convenience of tabulation, data are grouped in this way, the points of division are known, but in the case of qualitative data such information is usually absent. Throughout this paper the cut points  $\{\theta_j\}$  are assumed unknown. Ordinality is therefore an integral feature of such models and the imposition of an arbitrary scoring system for the categories is thereby avoided.

All the models advocated in this paper share the property that the categories can be thought of as contiguous intervals on some continuous scale. They differ in their assumptions concerning the distributions of the latent variable (e.g. normality (after suitable transformation), homoscedasticity etc.). It may be objected, in a particular example, that there is no sensible latent variable and that these models are therefore irrelevant or unrealistic. However, the models as introduced in Sections 2.1 and 3.1 make no reference to the existence of such a latent variable and its existence is not required for model interpretation. If such a continuous underlying variable exists, interpretation of the model with reference to this scale is direct and incisive. If no such continuum exists the parameters of the models are still interpretable in terms of the particular categories recorded and not those which might have obtained had the defining criteria  $\{\theta_j\}$  been different. Quantitative statements of conclusions are therefore possible in both cases although more succinct and incisive statements are usually possible when direct appeal to a latent variable is acceptable.

## 2. THE PROPORTIONAL ODDS MODEL

### 2.1. General

In all of the problems considered here, it is important to distinguish clearly between response variables, on the one hand, and explanatory factors or covariates, on the other. For further discussion of this point see Section 7.2. Suppose that the  $k$  ordered categories of the response have probabilities  $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \dots, \pi_k(\mathbf{x})$  when the covariates have the value  $\mathbf{x}$ . In the case of two groups  $\mathbf{x}$  is an indicator variable or two-level factor indicating the appropriate group. Let  $Y$  be the response which takes values in the range  $1, \dots, k$  with the probabilities given above, and let  $\kappa_j(\mathbf{x})$  be the odds that  $Y \leq j$  given the covariate values  $\mathbf{x}$ . Then the proportional odds model specifies that

$$\kappa_j(\mathbf{x}) = \kappa_j \exp(-\boldsymbol{\beta}^T \mathbf{x}) \quad (1 \leq j < k), \quad (2.1)$$

where  $\boldsymbol{\beta}$  is a vector of unknown parameters. The ratio of corresponding odds

$$\kappa_j(\mathbf{x}_1)/\kappa_j(\mathbf{x}_2) = \exp\{\boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1)\} \quad (1 \leq j < k) \quad (2.2)$$

is independent of  $j$  and depends only on the difference between the covariate values,  $\mathbf{x}_2 - \mathbf{x}_1$ .

Since the odds for the event  $Y \leq j$  is the ratio  $\gamma_j(\mathbf{x})/\{1 - \gamma_j(\mathbf{x})\}$ , where  $\gamma_j(\mathbf{x}) = \pi_1(\mathbf{x}) + \dots + \pi_j(\mathbf{x})$ , the proportional odds model is identical to the linear logistic model

$$\log[\gamma_j(\mathbf{x})/\{1 - \gamma_j(\mathbf{x})\}] = \theta_j - \boldsymbol{\beta}^T \mathbf{x} \quad (1 \leq j < k) \quad (2.3)$$

with  $\theta_j = \log \kappa_j$ , so that the difference between corresponding cumulative logits is independent of the category involved. Note in particular that when there are only two response categories, (2.3) is equivalent to the usual linear logistic model for binary data (Cox, 1970) and in this particular case it is also equivalent to a log-linear model. In general, however, when the number of categories exceeds 2, the linear logistic model (2.3) does not correspond to a log-linear structure.

### 2.2. An Example

As an initial example we take a two-sample problem where the response variable has three ordered categories. Model (2.3) reduces to

$$\begin{aligned} \lambda_{1j} &= \log\{\gamma_{1j}/(1 - \gamma_{1j})\} = \theta_j - \frac{1}{2}\Delta \\ &\quad (1 \leq j < k) \\ \lambda_{2j} &= \log\{\gamma_{2j}/(1 - \gamma_{2j})\} = \theta_j + \frac{1}{2}\Delta \end{aligned} \quad (2.4)$$

where  $\gamma_{ij}$  is the  $j$ th cumulative probability for the  $i$ th group and  $\lambda_{ij}$  is its logistic transform. The difference between corresponding logits,  $\lambda_{2j} - \lambda_{1j}$ , is the same constant,  $\Delta$ , for all  $j$ .

To illustrate an application of (2.4) we use the data in Table 1 from Holmes and Williams (1954) who classify 1398 children aged 0–15 years according to their relative tonsil size and

TABLE 1  
*Tonsil size of carriers and non-carriers of Streptococcus pyogenes*

	<i>Present but not enlarged</i>	<i>Enlarged</i>	<i>Greatly enlarged</i>	<i>Total</i>
Carriers	19	29	24	72
Non-carriers	497	560	269	1326
Total	516	589	293	1398

whether or not they were carriers of *Streptococcus pyogenes*. Our perspective in examining these data is to investigate the nature and direction of possible effects of *Streptococcus pyogenes* on tonsil size. Consequently, tonsil size is regarded as the dependent variable, presence or absence of *Streptococcus pyogenes* being regarded as a possible explanatory factor. Certainly this distinction is in keeping with possible biological mechanisms : if there is a causal relationship between the two variables it is almost certainly in the direction indicated rather than the reverse.

At least as a preliminary investigation of the adequacy of the linear logistic or proportional odds model it is recommended that the empirical logistic transforms and their differences be computed as shown in Table 2. The first sample logit for carriers is the log contrast of 19 versus

TABLE 2  
*An analysis of the tonsil size data*

Logits for carriers	– 1.009	0.683
Logits for non-carriers	– 0.511	1.367
Carriers minus non-carriers	– 0.498	– 0.684

29 + 24. To avoid zeros and to reduce bias it is advisable to add  $\frac{1}{2}$  to both numerator and denominator so that  $-1.009 = \log(19.5/53.5)$ . Similarly,  $0.683 = \log(48.5/24.5)$  is the log contrast for not enlarged and enlarged versus greatly enlarged. For non-carriers, the corresponding transforms are  $-0.511$  and  $1.367$  yielding differences on the logit scale of  $0.498$  and  $0.684$  respectively. From the practical viewpoint it is probably sufficient to note that these two values have the same sign and are of approximately the same magnitude. In essence then our conclusion is that the odds of having greatly enlarged tonsils are 1.8 times as large for carriers as for non-carriers and that the odds of having normal-sized tonsils are 1.8 times as large for non-carriers as for carriers. Here I have used  $1.8 = \exp \frac{1}{2}(0.498 + 0.684)$  taking an equally weighted average although this combination can be improved as indicated below. For a similar problem, Tukey (1977) gives essentially the same analysis under the name of flogs.

We now investigate some of the finer details of parameter estimation and model verification. In particular, it would be advantageous to obtain an efficient estimate of  $\Delta$  together with error estimates or confidence intervals for the common odds-ratio,  $\exp(\Delta)$ .

2.3. *A Generalized Empirical Logit Transform*

Let the cell counts be  $\{n_{ij}\}$  with row totals  $n_{i.}$ ,  $n_{.j}$  and column or category totals  $\{n_{.j}\}$ . The cumulative row sums are  $R_{ij}$  so that  $n_i = R_{i.}$  is the  $i$ th row total. Under the assumption of multinomial sampling in each row, the marginal distribution of  $R_{ij}$  conditional only on the row

total  $n_i$  is binomial with index  $n_i$  and parameter  $\gamma_{ij}$  satisfying (2.4). Hence the  $j$ th sample logit

$$\tilde{\lambda}_{ij} = \log \{(R_{ij} + \frac{1}{2}) / (n_i - R_{ij} + \frac{1}{2})\}$$

has expectation  $\lambda_{ij} + O(n_i^{-2})$  (Cox, 1970, p. 33; Plackett, 1974, p. 3). Hence, for any fixed weights  $\{w_j\}$  with  $\sum w_j = 1$ , the linear combination  $Z_i = \sum_j w_j \tilde{\lambda}_{ij}$  has expectation given by

$$E(Z_1) = -\frac{1}{2}\Delta + \sum w_j \theta_j + O(n_1^{-2}),$$

$$E(Z_2) = \frac{1}{2}\Delta + \sum w_j \theta_j + O(n_2^{-2}),$$

so that  $E(Z_2 - Z_1) = \Delta + O(n_1^{-2}, n_2^{-2})$ . For a similar estimator of  $\Delta$ , Clayton (1974) derived a formula for the weights  $\{w_j\}$  which minimize the asymptotic variance of  $Z_2 - Z_1$  when  $\Delta = 0$ . The weights are given by

$$w_j \propto \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}), \quad (2.5)$$

where  $\gamma_j$  is the common value of  $\gamma_{1j}$  and  $\gamma_{2j}$  under the hypothesis that  $\Delta = 0$ . Using these weights, the asymptotic variance of  $\Delta = Z_2 - Z_1$  was shown to be

$$\text{var}(\tilde{\Delta}) = \left\{ \frac{n_1 n_2}{n} \sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \right\}^{-1} + O(\Delta^2). \quad (2.6)$$

When  $k = 2$ , the expression  $\sum \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1})$  reduces to the familiar formula  $p(1 - p)$  for the binomial variance. When  $k = 3$ , the weights are proportional to  $\pi_1$  and  $\pi_3$  respectively,  $\pi_2$  being a measure of the correlation or information common to both  $\lambda_{i1}$  and  $\lambda_{i2}$ .

There are many equivalent forms of the summation in (2.6). The following are a few.

$$\begin{aligned} \text{(i)} \quad & \sum_{j=1}^{k-1} \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}); \quad \text{(ii)} \quad \sum_{j=1}^k \pi_j(1 - \gamma_j - \gamma_{j-1})^2; \quad \text{(iii)} \quad \sum_{j=1}^{k-1} \gamma_j \gamma_{j+1} \pi_{j+1}, \\ \text{(iv)} \quad & \sum_{j=1}^{k-1} (1 - \gamma_j)(1 - \gamma_{j-1}) \pi_j; \quad \text{(v)} \quad \frac{1}{3} - \frac{1}{3} \sum_{j=1}^k \pi_j^3. \end{aligned}$$

These expressions are related to the intrinsic accuracy of the logistic function and also to the loss of information about  $\Delta$  incurred by grouping the data. Expression (v) shows that for fixed  $k$  the variance is smallest when all categories have equal probability. For continuous distributions, the analogue is obtained by replacing  $\gamma$  by  $F(x)$ ,  $\pi$  by  $dF(x)$  and the summation becomes an integral. In this limit, all are equal to  $\frac{1}{3}$ .

A further slight problem is that the weights (2.5) are not obtainable directly and must therefore be estimated from the data. Clayton used weights obtained by substituting parameter estimates derived from category totals into (2.5). Simulation results by McCullagh (1977) indicate that, for a wide range of conditions, these weights do not produce noticeable bias in  $\tilde{\Delta}$ . The variance estimator (2.6) with  $\gamma_j$  estimated by  $R_{.j}/n$  can however seriously underestimate the true variance when  $|\Delta| > 1$ .

The quantity  $Z_i$  with weights given by

$$w_j \propto R_{.j}(n_{.j} - R_{.j})(n_{.j} + n_{.j+1})$$

is called the generalized empirical logit transform for the  $i$ th group. For the data of Table 1, the weights  $w_j$ , being proportional to category 1 and category 3 totals, are  $w_1 = 0.638$ ,  $w_2 = 0.362$ , yielding  $\tilde{\Delta} = 0.565$  with standard error 0.225. The parameter  $\Delta$  provides strong evidence that tonsil size tends to be larger in the carrier group than in the non-carrier group. Normal approximations for significance tests are best done on the logistic rather than the odds-ratio scale since the distribution of  $\tilde{\Delta}$  is likely to be more nearly symmetric than that of  $\exp(\tilde{\Delta})$ .

To check the adequacy of the linear logistic model, all parameters in (2.4) were estimated by maximum likelihood giving the following estimates and standard errors.

$$\hat{\Delta} = 0.603 \pm 0.225; \quad \hat{\theta}_1 = -0.810 \pm 0.116; \quad \hat{\theta}_2 = 1.061 \pm 0.118.$$

The residual deviance or likelihood ratio  $\chi^2$  statistic is  $G^2 = 0.302$  on one degree of freedom indicating a good fit. Details of maximum likelihood estimation are given in the Appendix.

The qualitative conclusion that tonsil size tends to be larger in the carrier than in the non-carrier group could have been obtained by a variety of other methods including the Wilcoxon test and tests based on partitioning Pearson's or the likelihood ratio  $\chi^2$  statistic; see, for example, Armitage (1955) and the discussion in Section 7.1. The quantitative conclusion that the odds for greatly enlarged tonsils are 1.8 times greater in the carrier than in the non-carrier group and that the odds for normal tonsils are 1.8 times greater in the non-carrier group can only be obtained through a parametric model. The great advantage of the quantitative approach is best seen in more complex examples with more structure in the explanatory variables.

### 3. THE PROPORTIONAL HAZARDS MODEL

#### 3.1. General

The hazard function or instantaneous risk function  $\lambda(t; \mathbf{x})$ , of major importance in the analysis of survival data, is defined to be the instantaneous failure probability at time  $t$  conditional on survival up to time  $t$ . For an individual with covariate  $\mathbf{x}$  the proportional hazards model is

$$\lambda(t; \mathbf{x}) = \lambda_0(t) \exp(-\boldsymbol{\beta}^T \mathbf{x}), \quad (3.1)$$

where  $\lambda_0(t)$  is the hazard function when  $\mathbf{x} = \mathbf{0}$  and  $\boldsymbol{\beta}$  is a vector of unknown parameters. Details of the use of this model in the analysis of survival data are given by Cox (1972). In the present context we note simply that the survivor function  $S(t; \mathbf{x})$ , being the probability of surviving beyond time  $t$  given covariate  $\mathbf{x}$ , satisfies

$$-\log \{S(t; \mathbf{x})\} = \Lambda_0(t) \exp(-\boldsymbol{\beta}^T \mathbf{x}), \quad (3.2)$$

where  $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ . Hence, for two individuals with covariates  $\mathbf{x}_1$  and  $\mathbf{x}_2$  respectively the survivor functions satisfy

$$\log \{S(t; \mathbf{x}_1)\} / \log \{S(t; \mathbf{x}_2)\} = \exp \{\boldsymbol{\beta}^T (\mathbf{x}_2 - \mathbf{x}_1)\}. \quad (3.3)$$

In other words, the ratio of log survivor functions, like the ratio of the hazard functions, depends only on the difference between the covariate values  $\mathbf{x}_2 - \mathbf{x}_1$  and is constant for all  $t$ .

For discrete data, the proportional hazards model (3.2) becomes

$$-\log \{1 - \gamma_j(\mathbf{x})\} = \exp(\theta_j - \boldsymbol{\beta}^T \mathbf{x}), \quad (3.4)$$

where  $1 - \gamma_j(\mathbf{x})$  is the complementary probability or the probability of "survival" beyond category  $j$  given covariate values  $\mathbf{x}$ . To obtain the appropriate linear structure analogous to the linear logistic model we write (3.4) in the more convenient form

$$\log [-\log \{1 - \gamma_j(\mathbf{x})\}] = \theta_j - \boldsymbol{\beta}^T \mathbf{x}, \quad (3.5)$$

the transformation to linearity being called the complementary log-log transform. Note in particular that when there are only two groups, so that the covariate  $\mathbf{x}$  takes on only two values,  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , the difference between corresponding complementary log-logs is the constant  $\boldsymbol{\beta}^T (\mathbf{x}_2 - \mathbf{x}_1)$  and is independent of the category involved. In this respect the properties of the complementary log-log model parallel those of the proportional odds or logit model.

#### 3.2. An Example

For an application of the proportional hazards model we turn to Table 638 of the *Statistical Abstract of the U.S.* (1975) which gives the income distribution in constant (1973) dollars for families in four geographic regions and in various years from 1960 to 1974. For illustrative purposes we use only the years 1960 and 1970. These data for the Northeast region of the U.S. are given in Table 3 where the data are expressed in percentages. Rounding errors occasionally force the totals to differ slightly from 100 per cent. It is far from clear *a priori* that

TABLE 3  
Family income distribution in constant (1973) dollars for Northeast U.S.

Year	Income group (000's)						
	0-3	3-5	5-7	7-10	10-12	12-15	15+
1960	6.5	8.2	11.3	23.5	15.6	12.7	22.2
1970	4.3	6.0	7.7	13.2	10.5	16.3	42.1

the proportional hazards model should provide a better description of these data than the proportional odds model in Section 2. We therefore examine both the empirical logit transformation and the empirical complementary log-log transformation of the data.

In this particular example we are not concerned with fitting a model in the conventional statistical sense. It is sufficient to point out that the sampling variation in the data is likely to be quite complex but the relative variation in the data is probably small. In particular, assumptions such as multinomial variation or even “proportional to multinomial” are probably quite untenable. Furthermore, the data for one year are not independent of the data for another year because the same individuals may be involved in both samples.

Table 4 gives an analysis of the income data based on the logit model of Section 2. The logit differences, being all of the same sign, indicate a strict stochastic ordering between the two

TABLE 4  
An analysis of family income data based on logits

	Category						
	1	2	3	4	5	6	7
Logits for 1960	-2.67	-1.76	-1.05	-0.02	0.62	1.25	
Logits for 1970	-3.10	-2.16	-1.52	-0.79	-0.34	0.32	
Differences	0.43	0.40	0.47	0.77	0.96	0.93	

distributions. In fact the odds for earning more than \$x were greater in 1970 than in 1960 for all x in the range \$3000 to \$15000. However, the ratio of corresponding odds is not constant but tends to increase with x. This tentative conclusion is reinforced when similar data for other areas of the U.S. are seen to display the same pattern (McCullagh, 1979).

Table 5 gives the corresponding analysis of the income data based on the proportional “hazards” or complementary log-log model. Thus the entries for 1960 are  $\log\{-\log(1-0.065)\}$ ,  $\log\{-\log(1-0.147)\}$ , etc. while those for 1970 are  $\log\{-\log(1-0.043)\}$ ,

TABLE 5  
An analysis of family income data based on complementary log-logs

	Category						
	1	2	3	4	5	6	7
Complementary log-log (1960)	-2.70	-1.84	-1.20	-0.38	0.05	0.41	
Complementary log-log (1970)	-3.12	-2.22	-1.62	-0.98	-0.62	-0.14	
Differences	0.42	0.38	0.42	0.60	0.67	0.55	

$\log \{ -\log(1 - 0.103) \}$ , etc. The differences between complementary log–logs for 1960 and 1970 are relatively constant with median value 0.49. Certainly these differences are more stable than the corresponding differences on the logit scale. The conclusion, therefore, is that if  $p_1(x)$  is the proportion of the population in the Northeast earning more than \$x in 1960 and  $p_2(x)$  is the corresponding proportion in 1970 then

$$\log p_1(x) = \exp(0.49) \log p_2(x), \quad (3.6)$$

at least for  $x$  in the range \$3000 to \$15000. Of course (3.6) is at best an approximation to reality but it is a simple and convenient description of the change in income distribution and may be sufficiently accurate for many purposes. McCullagh (1979) has examined the corresponding data for three other areas of the U.S. and in all cases has found that differences on the complementary log–log scale for the period 1960 to 1970 are more stable than differences on the logit scale. However, the size of the observed difference is substantially smaller in the West than in any other area of the U.S.

With hindsight, it is hardly surprising that the proportional hazards model should do better than the proportional odds model when applied to income distribution data. Econometricians frequently use the Pareto distribution to describe the tails of income distributions and the Weibull distribution is sometimes used in the centre. These two distributions, although quite different in shape, share the proportional hazard property (3.1) and (3.2). A great many other pairs of distributions satisfy the proportional hazards model (3.1). In fact, for any 1960 income distribution  $\{\pi_{1j}\}$  and for any value of  $\beta^T(x_2 - x_1)$  there exists a corresponding distribution  $\{\pi_{2j}\}$  for the 1970 incomes satisfying (3.4). Thus (3.4) makes no assumptions about the shape of income distributions but merely specifies a relationship between two distributions.

This example illustrates the power of the quantitative or parametric approach as opposed to the qualitative or non-parametric approach based on statistical tests. The major systematic component in the data is explained by (3.6). Any attempt to assess the adequacy of (3.6) will almost certainly yield a very large test statistic leading us to reject the model. However, this does not mean that (3.6) is not a useful description of the data. The single parameter accounts for roughly 90 per cent of the variation in the data. Other systematic components are undoubtedly present but their effect is small. In particular, there may also be a tendency towards uniformity of income. Such an effect can be detected by the non-linear models described in Section 6.1 but in this particular example the dominant effect is given by the relationship (3.6).

#### 4. PROPERTIES OF RELATED LINEAR MODELS

##### 4.1. *Stochastic Ordering*

The proportional odds and the proportional hazards models have the same general form namely

$$\text{link} \{ \gamma_j(x) \} = \theta_j - \beta^T x, \quad (4.1)$$

where “link” is the logit or complementary log–log function. Any other monotone increasing function mapping the unit interval  $(0, 1)$  onto  $(-\infty, \infty)$  can be used as a link function. In particular, the inverse normal function  $\Phi^{-1}(\gamma)$ , the inverse Cauchy function,  $\arctan \{ \pi(\gamma - 0.05) \}$ , and the log–log function,  $\log(-\log(\gamma))$  are possible candidates, although parameter interpretation is not generally so straightforward as with the proportional odds or proportional “hazards” models. Models of this form with a probit link function have been used by Aitchison and Silvey (1957), Ashford (1959), Gurland, Lee and Dolan (1960) and Finney (1971). The logit link function is preferred by Snell (1964), Simon (1974) and Bock (1975).

The parameters  $\{\theta_j\}$  are generally of little interest but are usually referred to as “cut points” on the logistic, probit, complementary log–log or other scale. The regression parameter  $\beta$  describes how the log odds or other quantity of interest is related to the covariates  $x$ . All models

of the form (4.1) describe strict stochastic ordering. Thus if we take two groups of subpopulations with covariates  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , it follows from (4.1) that

$$\text{link} \{ \gamma_j(\mathbf{x}_1) \} - \text{link} \{ \gamma_j(\mathbf{x}_2) \} = \boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1) = \Delta.$$

Hence, since “link” is a monotone function it follows that either

$$\gamma_j(\mathbf{x}_1) > \gamma_j(\mathbf{x}_2) \quad \text{for all } j,$$

or (4.2)

$$\gamma_j(\mathbf{x}_1) < \gamma_j(\mathbf{x}_2) \quad \text{for all } j,$$

according as  $\Delta > 0$  or  $\Delta < 0$ .

Thus all linear models of the form (4.1) are qualitatively similar and for any given data set the fits are often indistinguishable. Selection of an appropriate link function should therefore be based primarily on ease of interpretation. For this reason the linear logistic, log-log and complementary log-log models are preferred to the probit and inverse Cauchy models. For further discussion of the question of interpretability and model selection see Section 7.1.

#### 4.2. *Reversibility and Invariance*

A general property of all log-linear models that do not use “scores” is that they are permutation invariant. That is to say that the categories of the response can be permuted in an arbitrary way without affecting the fit or the values of the parameters. While this is an appealing property for responses on a nominal scale it is entirely inappropriate for ordinal data. A more appealing requirement for ordinal data is that the model should in some sense be invariant under a reversal of category order but not under arbitrary permutations. These ideas underlie the concept of palindromic invariance (McCullagh, 1978) but the force of this requirement depends heavily on the particular application.

Of the models discussed here, the logistic, probit and inverse Cauchy are invariant under a reversal of category order since the parameter  $\boldsymbol{\beta}$  merely changes sign and the  $\{\theta_j\}$  reverse sign and order. The complementary log-log model and its counterpart, the log-log model, are not so invariant. Depending on the application, this lack of invariance may or may not be seen as a flaw in the model.

An appealing requirement for any model is that it should be parameterized in such a way that the form of the systematic relation should apply under varying conditions and should, as far as possible, be consistent with known physical or biological laws. This means, for example, that to measure the difference between two proportions, the logistic scale is preferable to the probability scale since a constant difference is a logical possibility on the logistic scale but is logically impossible on the probability scale. The logistic scale is not unique in this respect but it does have other advantages. In the present context, varying conditions could mean a redefinition of the response categories, grouping or merging of the categories or the splitting of categories. Hence the parameter or parameters of interest should not depend *for their interpretation* on the actual response categories involved although the estimate will in general be affected. This property permits testing the consistency of various sources of information and, if warranted, combining information from the separate sources. All the models advocated in this paper share the above property: log-linear models such as those described in Section 7.1 do not.

#### 4.3. *Similar Rank Tests*

Under the hypothesis  $\boldsymbol{\beta} = \mathbf{0}$  in (4.1) the marginal totals for the response are sufficient for the nuisance parameters  $\{\theta_j\}$  regardless of the link function. Similar tests of this null hypothesis are therefore based on the conditional distribution of the data given the marginal totals for the response. Uniformly most powerful similar tests do not exist because there is no reduction by sufficiency away from the null hypothesis. The efficient score, however, gives a test with



maximum power locally and its value is unaffected by conditioning. For the linear logistic model the components of the efficient score are weighted cross-products of the covariates with the average rank for the response category, the weights being the cell counts. For the two sample problem this is exactly the Wilcoxon average rank statistic. In general, when  $\beta$  is scalar, the test is locally most powerful similar for one-sided alternatives. When  $\beta$  and hence the efficient score are vector valued, reduction by invariance leads to a generalization of the Wilcoxon test which is locally most powerful invariant similar for the hypothesis  $\beta = 0$ . When  $k = 2$  this test is equivalent to the conditional test given by Cox (1970, p. 45) and in this special case it is uniformly most powerful similar. Other link functions lead to different locally most powerful similar rank tests. Algebraic details in the general case are given by McCullagh (1977).

Exact probability calculations for significance tests are based on conditional distribution of the test statistic given the marginal totals. Except in the simplest of cases, this calculation will be extremely difficult and the asymptotically equivalent likelihood ratio statistic is suggested as a simple approximation although the associated test is not similar. However, the likelihood ratio has the great advantage that it provides an approximate test when components of  $\beta$  are regarded as nuisance parameters or when tests of the non-null hypothesis  $\beta = \beta_0 \neq 0$  are required. For these problems, no similar test seems possible except for the linear logistic model with  $k = 2$ .

Despite the fact that exact similar tests of certain interesting null hypotheses are possible via rank tests and tests based on parameter estimates are only approximate, rank tests are not put forward as an alternative to model building and parameter estimation. The main thrust of this paper is towards quantitative interpretation and description. Rank tests do not provide the parameter estimates required for this purpose. Furthermore, in many problems such as the example in Table 3, significance tests are irrelevant.

5. MORE COMPLEX COVARIATE STRUCTURE

5.1. *A Linear Regression Example*

We use the data in Table 6 taken from Maxwell (1961, p. 70) who tabulates the frequency of disturbed dreams among 223 boys aged 5–15. For an alternative analysis of the same data see Nelder and Wedderburn (1972) who use a log-linear model with linear scores. Again, in this

TABLE 6  
*Frequency of disturbed dreams among boys aged 5–15*

Age (yr)	Degree of suffering from disturbed dreams				Total
	Very severe	3	2	Not severe	
5–7	7	3	4	1	21
8–9	13	11	15	10	49
10–11	7	11	9	23	50
12–13	10	12	9	28	59
14–15	3	4	5	32	44
Total	40	41	42	100	223

example, it is important to distinguish between the response or dependent variable which is the frequency or severity of disturbed dreams, age being a possible explanatory factor or covariate. This asymmetry is also in keeping with the likely causal relationship between the two variables.

The simplest method of analysis is to use the generalized empirical logistic transform as described in Section 2.3. The category totals are 40, 41, 42, 100 giving weights 0.180, 0.290, 0.530 and the quantity  $\sum \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1})$  is estimated as 0.297. For the purpose of examining contrasts among the transformed values,  $\{Z_i\}$  as defined in Section 2.3, the variance of  $Z_i$  can be taken to be  $(0.297n_i)^{-1}$  as in (2.6) although it is possible to improve a little on this

approximation. Table 7 gives the transformed values  $\{Z_i\}$  together with the simple variance estimate  $(0.297n_i)^{-1}$ . Although the transformed values are not strictly monotone decreasing with age, the relationship is nevertheless very marked.

A weighted linear regression of  $Z$  on negative age, using age-category mid-points yields a regression coefficient of 0.217 with standard error 0.047. The weighted residual sum of squares is

TABLE 7  
*Generalized empirical logistic transform and variances for disturbed dreams data*

	Age				
	5-7	8-9	10-11	12-13	14-15
Transform ( $Z$ )	0.2045	0.5119	-0.3964	-0.3741	-1.4181
Variance	0.1603	0.0687	0.0673	0.0571	0.0765

6.43 on 3 degrees of freedom corresponding to a significance level of about 10 per cent giving reason to doubt the adequacy of a linear relationship while emphasizing that the major source of variation is due to the linear term.

A similar analysis by maximum likelihood involves fitting the two models

$$\log \{\gamma_{ij}/(1 - \gamma_{ij})\} = \theta_j - \beta x_i \quad (5.1)$$

and

$$\log \{\gamma_{ij}/(1 - \gamma_{ij})\} = \theta_j - \alpha_i, \quad (5.2)$$

where  $x_i$  is the age-category mid-point and  $\gamma_{ij} = \gamma_j(x_i)$ . The quantity  $\{\alpha_i\}$  is a factor associated with rows and the usual problems associated with intrinsic aliasing apply here.

The coefficient  $\beta$  in (5.1) is estimated to be 0.219 with standard error 0.0495 in close agreement with the earlier analysis, and the likelihood ratio  $\chi^2$  statistic is  $G_1^2 = 12.42$  on 11 degrees of freedom. For model (5.2) the parameters  $\{\alpha_i\}$  are estimated to be  $(-0.615, -0.720, 0.077, 0.058, 1.201)$  with the convention that  $\Sigma \alpha_i = 0$ , indicating possible departures from linearity similar to the  $\{Z_i\}$  in Table 7. The likelihood ratio goodness of fit statistic for (5.2) is  $G_2^2 = 7.15$  on 8 degrees of freedom. The difference  $G_1^2 - G_2^2 = 5.27$  on 3 degrees of freedom is a test for deviations from the regression line corresponding to a significance level of about 15 per cent.

The conclusion therefore is that, to a close approximation, the odds for disturbed or severely disturbed dreams among boys decreases by a factor of 0.80 per year from 5 to 15 years. Here I have used  $0.80 = \exp(-0.219)$ . There is inconclusive evidence to suggest that this decrease may not be uniform over the 10-year period. Qualitatively similar conclusions would be obtained by an analysis on the probit or other suitable scale.

## 5.2. Factorial Arrangements

Model (4.1) permits the most general factorial or nested structure for the explanatory variables or classifications. The detailed analysis of such an example is beyond the scope of this paper. See, however, McCullagh (1979) who analyses the mouse depletion data of Kastenbaum and Lamphiear (1959). An alternative analysis of the same data is given by Whittaker and Aitkin (1978).

One disadvantage of the present framework is that while use is made of order in the dependent variable, there is no corresponding way of using order in the levels of explanatory factors when such order exists. However, there are several alternative methods. One is to use scores such as the age category mid-points in the previous example and partition the relevant  $\chi^2$  statistic into linear, quadratic and higher order components. However, such scores may not be

given; in which case the method of isotonic regression (Barlow *et al.*, 1973) can be used to estimate the factor values (say  $\{\alpha_i\}$  in (5.2)) subject to the monotonicity property  $\hat{\alpha}_1 \leq \hat{\alpha}_2 \leq \dots \leq \hat{\alpha}_s$ . Details are beyond the scope of this paper.

6. PARAMETER ESTIMATION IN THE GENERAL MODEL

6.1. Non-linear Models

The general linear model (4.1) describes a mode of strict stochastic ordering among responses. While this corresponds to by far the most common pattern observed in data, other patterns also occur. A good example is the quality of vision data for men and women in Table 8 taken from Stuart (1953). On transforming to percentages it is clear that women are relatively more concentrated in the middle categories while the men have higher proportions in two extreme categories. Since (4.1) has an obvious interpretation in terms of shifted distributions on an underlying continuum, the most natural generalization is to relax the assumption of constant “variance” or scale parameter on that continuum. We therefore introduce the multiplicative model

$$\text{link}(\gamma_{ij}) = (\theta_j - \beta^T x_i)/\tau_i, \tag{6.1}$$

where the quantity  $\beta^T x_i$  is called the “location” for the  $i$ th row and  $\tau_i$  is called the “scale” for the  $i$ th row. In general, such a model is appropriate only when the number of response categories is three or more. Since, in (6.1), we have one scale parameter associated with each row of the table, we say that the model is saturated in scale parameters. It is also saturated in location parameters if  $\dim(\beta) \geq t - 1$  where  $t$  is the number of rows. To make the scale parameters identifiable it is convenient to impose a constraint such as  $\tau_1 = 1$  or  $\sum \log \tau_i = 0$ . The latter convention is adopted here and is particularly appropriate when we consider unsaturated scale models satisfying

$$\log \tau_i = \tau^T(x_i - \bar{x}), \tag{6.2}$$

where  $\tau$  is a vector of unknown parameters to be estimated. Furthermore, the estimates of  $\log \tau_i$  are likely to be more nearly symmetrically distributed than those of  $\tau_i$ , a property which greatly improves approximations based on normality.

For the data in Table 8 we find location and scale differences on the logit scale to be  $\hat{\Delta} = 0.061$  with standard error 0.041 and  $\log(\hat{\tau}_1/\hat{\tau}_2) = 0.272$  with standard error 0.025. Fitted values from this model correspond very closely to the observed values. It is clear that the major

TABLE 8  
*Quality of right eye vision in men and women*

	Vision quality				Total
	Highest	2	3	Lowest	
Men	1053	782	893	514	3242
Women	1976	2256	2456	789	7477

difference between the two groups is described by the scale parameters. In fact the ratio of corresponding logits rather than their difference is almost constant for this particular data set. Similar conclusions apply when we compare the corresponding data for left eyes. Essentially the same conclusions could be reached by an analysis on the probit rather than the logit scale.

6.2. Maximum Likelihood Estimation

The problem of obtaining maximum likelihood estimates is considerably simplified by noting that the linear systematic structure (4.1) together with multinomial variation comprise a

multivariate generalized linear model. Univariate generalized linear models have been discussed by Nelder and Wedderburn (1972) who have shown that parameter estimates can be obtained by iteratively reweighted least squares. We show here that a similar algorithm can be used even for the non-linear model (6.1) with scale parameters satisfying (6.2).

The contribution from a single multinomial observation  $(n_1, \dots, n_k)$  to the likelihood function is  $\pi_1^{n_1} \dots \pi_k^{n_k}$  with the probabilities  $\pi_j$  satisfying (4.1) or, more generally, (6.1). Since we are dealing with cumulative probabilities, we define

$$\begin{aligned} R_1 &= n_1, & Z_1 &= R_1/n, \\ R_2 &= n_1 + n_2, & Z_2 &= R_2/n, \\ &\vdots & &\vdots \\ R_k &= \sum n_j = n; & Z_k &= R_k/n = 1. \end{aligned}$$

In terms of the parameters of the cumulative transformation, the likelihood can be written as the product of  $k-1$  quantities

$$\left\{ \left( \frac{\gamma_1}{\gamma_2} \right)^{R_1} \left( \frac{\gamma_2 - \gamma_1}{\gamma_2} \right)^{R_2 - R_1} \right\} \left\{ \left( \frac{\gamma_2}{\gamma_3} \right)^{R_2} \left( \frac{\gamma_3 - \gamma_2}{\gamma_3} \right)^{R_3 - R_2} \right\} \dots \left\{ \left( \frac{\gamma_{k-1}}{\gamma_k} \right)^{R_{k-1}} \left( \frac{\gamma_k - \gamma_{k-1}}{\gamma_k} \right)^{R_k - R_{k-1}} \right\}.$$

These factors are respectively the probability given  $R_2$  that the first two cells divide in the ratio  $R_1 : R_2 - R_1$ ; the probability given  $R_3$  that the proportion in cell 3 relative to cells 1 and 2 combined is  $R_3 - R_2 : R_2$  and so on for the other components.

It is convenient to define

$$\phi_j = \log \{ \gamma_j / (\gamma_{j+1} - \gamma_j) \} = \text{logit} (\gamma_j / \gamma_{j+1})$$

and

$$g(\phi) = \log \{ 1 + \exp(\phi) \} = \log \{ \gamma_{j+1} / (\gamma_{j+1} - \gamma_j) \},$$

whence the log likelihood is

$$l = n[ \{ Z_1 \phi_1 - Z_2 g(\phi_1) \} + \{ Z_2 \phi_2 - Z_3 g(\phi_2) \} + \dots + \{ Z_{k-1} \phi_{k-1} - g(\phi_{k-1}) \} ]. \quad (6.3)$$

A univariate generalized linear model contains only one of the above components. The following relationships follow from (6.3).

$$\begin{aligned} E(Z_j | Z_{j+1}) &= Z_{j+1} g'(\phi_j) = Z_{j+1} \gamma_j / \gamma_{j+1}, \\ E(Z_j) &= \gamma_j, \\ \text{var}(Z_j | Z_{j+1}) &= Z_{j+1} g''(\phi_j) / n, \\ \text{var}(Z_j) &= \gamma_j (1 - \gamma_j) / n. \end{aligned}$$

Details of the general fitting method are not of great interest and are relegated to the Appendix. Experience with an interactive computer package at the University of Chicago shows that the Newton-Raphson method with Fisher scoring, as described in the Appendix, converges rapidly even when the initial estimates are poor. Of course, the initial values assigned to  $\{\theta_j\}$  must be monotone increasing. While uniqueness of the maximum is not guaranteed in general, I have not yet seen an example with multiple maxima. Generally, about 4 or 5 cycles are required to produce accuracy to four significant digits in all parameters.

### 6.3. Existence and Uniqueness of Maximum Likelihood Estimates

Uniqueness of maximum likelihood estimates depends on (1) the concavity of the likelihood function and (2) identifiability of the model. Identifiability is a problem which applies to linear models generally and is intimately related to the rank of the design matrix. Assuming that the

problem of identifiability has been eliminated either by the imposition of appropriate constraints or by the use of generalized inverse matrices, there remains the problem of establishing concavity of the likelihood function in the reduced parameter space.

Strict concavity and hence uniqueness of the maximum is assured if the negative matrix of second derivatives,  $\mathbf{H}$ , is positive definite with eigenvalues bounded away from zero. The weaker result, that the probability of a unique maximum tends to one, follows on demonstrating that, as the sample size increases, the probability tends to one that  $\mathbf{H}$  is positive definite. This latter result follows from the fact that  $\mathbf{H}$  has positive definite expectation  $\mathbf{A}$  since, from the Appendix,  $\mathbf{A} = \Sigma \mathbf{Q}_i \mathbf{Q}_i^T$ . Here,  $\mathbf{Q}_i$  is a full rank matrix of order  $p \times (k-1)$  where  $p$  is the total number of non-aliased parameters in the model. Since the random component of  $\mathbf{H}$  is of smaller order than  $\mathbf{A}$ , it follows that  $\mathbf{H}$  is almost surely positive definite either as  $n_i \rightarrow \infty$  in fixed proportion or as the number of multinomial samples tends to infinity with  $\{n_i\}$  bounded and  $p$  fixed.

This result, while reassuring, is of limited practical value. In practice, problems occasionally arise with infinite parameter values associated with patterns of zeros in sparse data arrays. Furthermore, convergence problems not associated with patterns of zeros can arise in small samples when the non-linear models are used. Non-convergence in this case is usually a reliable indication that the model being fitted is inappropriate.

## 7. ALTERNATIVE MODELS AND THE SCOPE OF POSSIBLE INFERENCES

### 7.1. *The Log-linear Model*

The general linear model (4.1) is not the only structural equation which describes strict stochastic order. One currently popular alternative is to use scores in a log linear model in order to partition the interaction statistic into two components, one describing strict stochastic order and the other, higher order components of the interaction. The idea goes back to Yates (1948) who used essentially the same method to partition Pearson's  $\chi^2$  statistic and thus obtain a more powerful test when the categories of the response are ordered. In the context of log-linear models, see, for example, Nelder and Wedderburn (1972), Haberman (1974), Bock (1975) and Goodman (1979).

To clarify the discussion we restrict attention to the two sample problem where the response has  $k$  ordered categories, scores  $1, \dots, k$  being attached to the response categories. The log-linear model with main effects only implies that the log-cross-ratio for all groups of 4 adjacent cells is zero while inclusion of a linear  $\times$  linear term implies that this log-cross-ratio is constant over the entire table. It is not difficult to see that this model defines a form of stochastic ordering, in the sense of (4.2), for the rows of the table. For this reason the fit produced by such a model is often not very different from that produced by (4.1) especially when the number of categories is small (say 4 or fewer). Because it describes a form of stochastic ordering, the test statistic associated with the linear  $\times$  linear interaction term is more powerful against the anticipated departures than, say, Pearson's or the likelihood ratio  $\chi^2$  test.

However, as a descriptive statistic, the cross-ratio for *adjacent cells* has little to recommend it even in cases where the model fits well which it often does. The major disadvantage is that the scope of possible inferences based on such a statistic is limited to the actual categories used in the sample. If we were to fit such a model to the income distribution data of Table 3, the cross-ratio parameter could be interpreted only with reference to the income grouping 0-3, 3-5, 5-7 and so on. Different, but no less arbitrary, groupings would produce entirely different cross-ratios which, in general, would not remain constant over the table. This leads directly to the second objection, namely lack of invariance under grouping of adjacent response categories.

On the other hand, conclusions based on the proportional odds or proportional hazards models can often be stated without reference to the categories used in the sample. For example, the conclusion from the income distribution data is that, to a close approximation, the proportions of households in the Northeast,  $p_1(x)$ ,  $p_2(x)$  earning more than \$ $x$  in 1960, 1970 are related by  $\log \{p_1(x)\} / \log \{p_2(x)\} = 1.63$ . Essentially the same conclusions

would have been reached had the categories been defined differently. Of course it would be dangerous to extend these conclusions much beyond the range of values of  $x$  in the sample.

This contrast in approaches highlights an important difference between testing and modelling. Inclusion of the linear  $\times$  linear or other low-dimensional interaction term in a log-linear model often produces a good fit and a test statistic with reasonably good power against anticipated departures from the null model. However, it is equally clear from the limited scope of possible conclusions or inferences, and from invariance properties that are often inappropriate to the specific problem, that real data must rarely behave in this way. More incisive and quantitative inferences can be made through the proportional odds or proportional hazards models, whichever is more appropriate. The choice between these and log-linear models should be based primarily on the interpretability of parameters and the scope of possible inferences rather than on goodness-of-fit statistics which rarely differentiate between the various models.

In one of the few attempts to relate the parameters of a log-linear model to an underlying continuous process, Andrich (1979) derives the same model for ordinal responses as Nelder and Wedderburn (1972), Haberman (1974) and Goodman (1979). He considers an underlying continuum with  $k-1$  thresholds  $\{\theta_j\}$ . For each threshold an independent decision is made as to whether or not the given object exceeds that threshold. Independence means that the order in which the decisions are made is unimportant. This procedure obviously leads to possible nonsensical allocations such as assigning the object simultaneously to two different categories. Eliminating such nonsense by redistribution over the logically possible outcomes leads to a log-linear model with “linear response  $\times$  explanatory variables” interaction term.

Certain aspects of this derivation deserve comment. In particular the assumption of independence is highly counter-intuitive. It is unreasonable to require that decisions made at the later thresholds be made independently of earlier decisions, particularly when the later decisions may lead to nonsensical allocations in view of decisions made earlier.

Finally we note that, for the log-linear model with linear response  $\times$  explanatory variable interaction term, the uniformly most powerful similar test for the presence of this type of interaction is based on the statistic  $T = \sum n_{ij} x_i s_j$ , where  $x$  is the covariate vector and  $s_j = j$  is the integer score attached to the  $j$ th response category. Conditional on both margins, the distribution of  $T$  depends only on the interaction parameter. Note in particular that if RIDIT scores (Bross, 1958) are used instead of integer scores for the response categories, one obtains the Wilcoxon test or its generalization as the uniformly most powerful similar test. Taken with the comments of Section 4.3, this result strongly suggests that, unless there are gross differences in the marginal probabilities for the various response categories, there is likely to be little difference in terms of the fitted values between the log-linear model with “linear response  $\times$  explanatory variable” interaction, and the corresponding linear logistic model.

## 7.2. *Symmetric versus Asymmetric Models*

Log-linear models for multi-factor contingency tables are symmetric in that no one factor takes precedence over any other. No factor is considered dependent and the others explanatory: all are treated on an equal footing. Logit linear models for binary data, on the other hand, are asymmetric: one binary factor (the response) is selected for special treatment and the other factors are explanatory. The analogous procedures for continuous data are correlation analysis which is symmetric and regression analysis which is not. Oddly enough, the logit linear model for binary data is a special case of a log-linear model. It is tempting, therefore, to discard the logit linear model for binary data as merely a special case of the more general log-linear model. It seems to me that this would be a serious mistake because the philosophies behind the two approaches are quite different. Furthermore, the natural extension of logit linear models described in this paper is also asymmetric but does not, except in the special case  $k = 2$ , correspond to a log-linear model.

The choice of dependent variable is usually easy as in the case of the income distribution data, the quality of vision data and the disturbed dreams data. In the case of the tonsil size data we attempt to model the biological process involved, in which case the most natural causal relationship is for *Streptococcus pyogenes* to affect tonsil size. Consequently tonsil size is dependent, carrier *versus* non-carrier being regarded as explanatory. On the other hand, regarded as a classification problem and not as a model of the underlying biological process, we might try to classify individuals as carriers or non-carriers on the basis of tonsil size. That is to say that, on the basis of this sample, we construct a model for  $\text{prob}(\text{carrier} | \text{tonsil size})$  as a function of tonsil size and use the parameters of such a model to classify future individuals as carriers or non-carriers. From an epidemiological viewpoint this would mean that the size of the epidemic in the sampled population must be the same as the size in the population for which inference is required—often a most unreasonable assumption. The corresponding allocation problem in the case of the quality of vision data would be to classify individuals as male or female on the basis of vision quality, which is not the most reliable indicator but an indicator nonetheless.

Not every data set falls into the category of “single response, multiple explanatory factor”. Occasionally we have a bivariate or higher dimensional response of either ordinal or nominal type in each margin. This is a multivariate problem and, at least when the categories are on a nominal scale, a log-linear model might be appropriate. On the other hand, when we wish to study the dependence of one variable on the levels of the others we have a univariate problem. Even if the details of the algebra or computation are identical, explanation of the model and conclusions is very much easier in the univariate case.

## 8. AN ANALYSIS OF RESIDUALS

### 8.1. General

The analysis of residuals from multinomial response models is complicated by several factors. Firstly, the residuals, however standardized, must take one of a limited number of possible values which causes problems in rankit or related plots when the cell counts are small. Secondly, it is far from clear in general that *cell* residuals are the relevant quantities to examine. What is important in the data is not a cell count *per se* but, rather, how the count for a cell or group of cells varies relative to another cell or group of cells. For binary data we would normally use one residual associated with two cells (“successes” and “failures”), whereas two residuals would usually be produced if the corresponding log-linear model were used. Clearly, since these residuals are negatively correlated in pairs, only one residual is necessary. Central to this discussion is the concept of an observation. For binary data an observation is the *proportion* of “successes” or equivalently the *number* of “successes” *relative* to the *number* of “failures”. The usual standardized residual, in an obvious notation, is  $(p - \hat{p})/(\hat{p}\hat{q}/n)^{1/2}$  which is the signed square root of the contribution to Pearson’s chi-squared statistic. An analogous residual could be defined by using the contribution to the likelihood ratio statistic instead.

For multinomial responses we propose to use as a residual the contribution to the likelihood ratio statistic from each multinomial sample. This so-called residual is always positive and thus does not indicate the “direction” of departure of the observed values from the fitted values. Of course, in a bivariate or multivariate problem direction of departure cannot be specified merely by a sign. As the second stage in an analysis of residuals, cell residuals can be computed to examine the precise nature of the observed discrepancy. Different cell residual patterns in a single multinomial sample indicate different inadequacies in the model. We illustrate some of these by example.

### 8.2. Example

We use the data in Table 9 from Bradley, Katti and Coons (1962) which gives the response frequency of judges in a taste-testing experiment. The five possible responses are on an ordered

TABLE 9  
Response frequency in a taste-testing experiment

Treatment	Response category					Total
	Terrible	2	3	4	Excellent	
1	9	5	9	13	4	40
2	7	3	10	20	4	44
3	14	13	6	7	0	40
4	11	15	3	5	8	42
5	0	2	10	30	2	44
Total	41	38	38	75	18	210

scale from terrible (1) to excellent (5), while the rows represent five unordered treatments. As a first attempt to describe these data we use the proportional odds model with a five-level treatment factor. Using this model the “residuals” associated with the five treatments are respectively 2.6, 3.7, 2.7, 23.3 and 16.8 giving a total deviance or likelihood ratio statistic of 49.1 on 12 degrees of freedom. Examination of cell residuals for treatments 4 and 5 reveals the following pattern : + + – – + for treatment 4 and – – + + – for treatment 5. Clearly, a model allowing unequal scale parameters is more appropriate. Hence the model

$$\text{logit}(\gamma_{ij}) = (\theta_j - \alpha_i)/\tau_i$$

was fitted giving row “residuals” of 1.3, 1.1, 0.9, 11.6, 0.6 and a total deviance of 21.3 on 8 degrees of freedom. This is clearly an improvement over the linear model although treatment 4 appears to be an outlier. A summary table such as Table 10 is often a useful aid to interpretation.

TABLE 10  
Logistic summary statistics for the taste-testing data

Treatment	Location ( $\hat{\alpha}$ )	Scale ( $\hat{\tau}$ )	“Residual”
1	0.07	1.25	1.86
2	0.56	1.05	5.30
3	– 1.11	0.90	1.97
4	– 0.50	1.66	11.62
5	0.98	0.51	0.59
			21.34

Examination of cell residuals reveals that, for treatment 4, many judges reacted extremely, giving either very low or very high scores. This diversity of opinion is also partly reflected in the large scale value for this observation. The anomalous reaction of the judges to treatment 4 was also noted by Snell (1964) who used a similar model but did not allow for unequal scale parameters. In conclusion, treatment 5 has the most favourable overall response and the consensus among judges is also highest ( $\tau_5 = 0.51$ ) for this treatment. Treatment 3 rates worst overall and the consensus for this is also high ( $\tau_3 = 0.90$ ). As we might have expected, agreement among judges is greatest when the treatment is either very good or very bad. This conclusion is not an artefact of the observations piling up in the end categories.

ACKNOWLEDGEMENTS

Much of the work for this paper was done as part of the author’s Ph.D. thesis at Imperial College, London, where financial support was provided by the Health and Safety Executive. I



wish to thank F. J. Anscombe, A. C. Atkinson, D. R. Cox, L. A. Goodman, S. J. Haberman, W. H. Kruskal, R. L. Plackett, D. L. Wallace and referees for helpful discussions and constructive comments on earlier versions of this paper.

Support for this research was provided in part by National Science Foundation Grant No. SOC76-80389.

## REFERENCES

- AITCHISON, J. and SILVEY, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, **44**, 131–140.
- ANDRICH, D. A. (1979). A model for contingency tables having an ordered response classification. *Biometrics*, **35**, 403–415.
- ANSCOMBE, F. J. (1970). Computing in statistical Science with A.P.L. Unpublished manuscript.
- ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.
- ASHFORD, J. R. (1959). An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics*, **15**, 573–581.
- BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1973). *Statistical Inference under Order Restrictions*. New York : Wiley.
- BRADLEY, R. A., KATTI, S. K. and COONS, I. J. (1962). Optimal scaling for ordered categories. *Psychometrika*, **27**, 355–374.
- BOCK, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York : McGraw-Hill.
- BROSS, I. D. J. (1958). How to use RIDIT analysis. *Biometrics*, **14**, 18–38.
- CLAYTON, D. G. (1974). Some odds-ratio statistics for the analysis of ordered categorical data. *Biometrika*, **61**, 525–531.
- COX, D. R. (1970). *The Analysis of Binary Data*. London : Chapman and Hall.
- (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc. B*, **34**, 187–220.
- FINNEY, D. J. (1971). *Probit Analysis* (3rd edition) London : Cambridge University Press.
- GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Ass.*, **74**, 537–552.
- GURLAND, J., LEE, I. and DAHM, P. A. (1960). Polychotomous quantal response in biological assay. *Biometrics*, **16**, 382–398.
- HABERMAN, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics*, **30**, 589–600.
- HOLMES, M. C. and WILLIAMS, R. E. O. (1954). The distribution of carriers of *Streptococcus pyogenes* among 2413 healthy children. *J. Hyg. Camb.*, **52**, 165–179.
- KASTENBAUM, M. A. and LAMPHEAR, D. E. (1959). Calculation of chi-square to calculate the no three factor interaction hypothesis. *Biometrics*, **15**, 107–115.
- MCCULLAGH, P. (1977). Analysis of ordered categorical data. Ph.D. Thesis, University of London.
- (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika*, **65**, 413–415.
- (1979). The use of the logistic function in the analysis of ordinal data. *Bull. Int. Statist. Inst.*, to appear.
- MAXWELL, A. E. (1961). *Analysing Qualitative Data*. London : Methuen.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A*, **135**, 370–384.
- PEARSON, K. (1913). Note on the surface of constant association. *Biometrika*, **9**, 534–537.
- PLACKETT, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Ass.*, **60**, 516–522.
- (1974). *The Analysis of Categorical Data*. London : Griffin.
- SIMON, G. (1974). Alternative analyses for the singly ordered contingency table. *J. Amer. Statist. Ass.*, **69**, 971–976.
- SNELL, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics*, **20**, 592–607.
- Statistical Abstract of the United States* (1975). Washington, D.C.; U.S. Bureau of the Census.
- STEVENS, S. S. (1951). Mathematics, measurement and psychophysics. In *Handbook of Experimental Psychology* (S. S. Stevens, ed.) New York : Wiley.
- (1958). Problems and methods of psychophysics. *Psychol. Bull.*, **55**, 177–196.
- (1968). Measurement, statistics and the schemaparc view. *Science*, **161**, 849–856.
- TUKEY, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison Wesley.
- WHITTAKER, J. and AITKIN, M. (1978). A flexible strategy for fitting complex log-linear models. *Biometrics*, **34**, 487–495.
- YATES, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, **35**, 176–181.

## APPENDIX

### *Fitting the General Non-linear Model*

The general non-linear model can be written

$$Y_j = \text{link}(\gamma_j) = \beta^* \mathbf{T} \mathbf{X}_j^* \exp(\mathbf{T}^T \mathbf{U}),$$

where  $\beta^* = (\theta_1, \theta_2, \dots, \theta_{k-1}, \beta_1, \dots, \beta_p)$  is the vector of parameters in the location model;  $\mathbf{X}_j^* = (0, \dots, 1, \dots, 0, \mathbf{X})$ , where the 1 occurs in position  $j$  and  $\mathbf{U}_k$ , the vector of parameters in the scale model is normalized so that  $\sum_i \mathbf{U}_i = \mathbf{0}$ . Let  $\Psi^T = (\beta^{*T}, \tau^T)$  be the complete parameter vector and  $w = \exp(\tau^T \mathbf{U})$ .

The derivative of the log-likelihood with respect to  $\beta^*$  is

$$\frac{\partial l}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{dY_j} X_{jr}^* + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} X_{j+1,r}^* \right\}.$$

Substituting  $V_j = \partial \gamma_j / \partial \phi_j$  and  $\partial \phi_j / \partial \gamma_{j+1} = (-\gamma_j / \gamma_{j+1}) V_j^{-1}$  we obtain

$$\frac{\partial l}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial l}{\partial \phi_j} V_j^{-1} q_{jr},$$

where

$$q_{jr} = \left\{ \frac{d\gamma_j}{dY_j} X_{jr}^* - \frac{\gamma_j}{\gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} X_{j+1,r}^* \right\}.$$

Similarly the expected second derivative is

$$A_{rs} = -E \left( \frac{\partial^2 l}{\partial \beta_r^* \partial \beta_s^*} \right) = nw^2 \sum_j V_j^{-1} q_{jr} q_{js}.$$

For the scale parameter  $\tau$  the derivatives are

$$\frac{\partial l}{\partial \tau_r} = \sum \frac{\partial l}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{dY_j} Y_j U_r + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} Y_{j+1} U_r \right\},$$

which reduce to

$$\frac{\partial l}{\partial \tau_r} = U_r \sum_j \frac{\partial l}{\partial \phi_j} V_j^{-1} q_j,$$

where

$$q_j = Y_j \frac{d\gamma_j}{dY_j} - \frac{\gamma_j}{\gamma_{j+1}} Y_{j+1} \frac{d\gamma_{j+1}}{dY_{j+1}}.$$

Similarly the expected second derivatives are

$$-E \left( \frac{\partial^2 l}{\partial \tau_r \partial \tau_s} \right) = n U_r U_s \sum_j V_j^{-1} q_j^2,$$

and the mixed second derivatives are

$$-E \left( \frac{\partial^2 l}{\partial \beta_r^* \partial \tau_s} \right) = nw U_s \sum_j V_j^{-1} q_j q_{jr}.$$

The negative matrix of expected second derivatives is therefore positive semi-definite for all admissible parameter values.

The Taylor series expansion for  $\partial l / \partial \Psi$  gives

$$\frac{\partial l}{\partial \Psi} = \frac{\partial l}{\partial \Psi} \bigg|_{\Psi = \hat{\Psi}} + \mathbf{A}(\delta \Psi) + \dots,$$

where  $\delta \Psi = \hat{\Psi} - \Psi$  and  $\mathbf{A}$  is the negative expected matrix of second derivatives. Hence, updated values of  $\Psi$  are obtained by iterating on the equation

$$\mathbf{A} \Psi_{n+1} = \mathbf{A} \Psi_n + \frac{\partial l}{\partial \Psi} = \mathbf{b}.$$

Writing  $W = \log w = -\tau^T U$  we have

$$\begin{aligned} (\mathbf{A}\Psi)_r &= nw^2 \sum_j V_j^{-1} q_{jr} \sum_s q_{js} \beta_s - nw \sum_s U_s \tau_s \sum_j V_j^{-1} q_{jr} q_j \\ &= nw(1+W) \sum_j V_j^{-1} q_j q_{jr} \end{aligned}$$

for  $r \leq \dim(\beta^*)$ . Similarly, for  $s > \dim(\beta^*)$ ,

$$\begin{aligned} (\mathbf{A}\Psi)_s &= nW U_s \sum_j V_j^{-1} q_j^2 + nU_s \sum_j V_j^{-1} q_j^2 \\ &= n(1+W) U_s \sum_j V_j^{-1} q_j^2. \end{aligned}$$

The corresponding elements of  $\mathbf{b}$  are given by

$$(\mathbf{b})_r = nw \sum_j V_j^{-1} q_{jr} \left\{ q_j(1+W) + Z_j - \frac{\gamma_j}{\gamma_{j+1}} Z_{j+1} \right\}$$

and

$$(\mathbf{b}) = nU_s \sum_j V_j^{-1} q_j \left\{ q_j(1+W) + Z_j - \frac{\gamma_j}{\gamma_{j+1}} Z_{j+1} \right\}.$$

Finally, all the above equations represent the contribution to the log-likelihood and its derivatives from a single multinomial observation and the total contribution is the sum of the individual contributions.

#### DISCUSSION OF DR MCCULLAGH'S PAPER

Professor D. J. BARTHOLOMEW (London School of Economics): Dr McCullagh's paper contains an interesting new development which will be of particular interest to social scientists who often have to deal with ordered data. The author is right to insist that the form of the analysis should match the data. For too long social scientists have had to manage with methods designed for use in the natural sciences in which the variables are well defined and readily capable of measurement. In the last decade or two the balance has been somewhat restored by the development by such things as the log-linear model. Tonight's paper represents a further step along that road. The full potentialities of the author's method have yet to be realized. In his examples only one covariate is considered and in all cases but one this is categorical. It will be interesting to see analyses involving more covariates. My remarks, however, are directed to generalizations in a different direction, the aim being to set the methods given here in a broader context.

In essence the problem is to compare grouped frequency distributions which, in the examples of the paper, appear as the rows of the tables. Their special feature lies in the fact that the values of the group boundaries are the same for each row even though they are not (or need not) be known. The aim is to explain the differences between the row distributions in terms of one or more covariates. The following treatment includes some of the author's results as special cases.

Suppose there is a random variable  $\Theta$  underlying the response categories with distribution function  $F_\Theta(\theta)$ . Then for each row we can estimate  $F$  at those values of  $\theta$  which correspond to the group boundaries,  $\theta_1, \theta_2, \dots, \theta_{k-1}$  say. The ranking of the  $\theta$ 's is known but not their actual values. We suppose that the distribution of  $\Theta$  has the same form for all rows but that the parameters vary from row to row. Let  $F$  have the form

$$F_\Theta(\theta) = \psi\left(\frac{m(\theta) - a}{b}\right), \quad (-\infty < m(\theta) < \infty),$$

where  $m(\theta)$  is any monotonic function of  $\theta$ ,  $\psi(\cdot)$  is a distribution function with range  $(-\infty, +\infty)$  and  $-\infty < a, < \infty$ ,  $b > 0$ . This family includes many commonly occurring distributions and embraces a wide variety of shapes.

Let

$$p_i = \psi\left(\frac{m(\theta_i) - a}{b}\right)$$