J. R. Statist. Soc. B (1980),
42, *No.* 2, *pp.* 109–142

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 $\theta_1, ..., \theta_{k-1}$. 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 $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), ..., \pi_k(\mathbf{x})$ 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 $1, ..., k$ with the probabilities given above, and let $\kappa_j(\mathbf{x})$ be the odds that $Y \leqslant j$ given the covariate values $\mathbf{x}$. Then the proportional odds model specifies that

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

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

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

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

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

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

$$\begin{aligned} \lambda_{1j} &= \log\{\gamma_{1j}/(1-\gamma_{1j})\} = \theta_j - \tfrac{1}{2}\Delta \\ & \qquad\qquad (1 \leqslant j < k) \\ \lambda_{2j} &= \log\{\gamma_{2j}/(1-\gamma_{2j})\} = \theta_j + \tfrac{1}{2}\Delta \end{aligned} \tag{2.4}$$

where γ_{ij} is the jth cumulative probability for the ith group and λ_{ij} is its logistic transform. The difference between corresponding logits, $\lambda_{2j}-\lambda_{1j}$, is the same constant, Δ, 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

	Present but not enlarged	*Enlarged*	*Greatly enlarged*	*Total*
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 $-1{\cdot}009 = \log(19{\cdot}5/53{\cdot}5)$. Similarly, $0{\cdot}683 = \log(48{\cdot}5/24{\cdot}5)$ 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 for having normal-sized tonsils are 1·8 times as large for non-carriers as for carriers. Here I have used $1{\cdot}8 = \exp\frac{1}{2}(0{\cdot}498+0{\cdot}684)$ 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 Δ 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_1, n_2 and column or category totals $\{n_{\cdot j}\}$. The cumulative row sums are R_{ij} so that $n_i = R_{ik}$ is the ith 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 γ_{ij} satisfying (2.4). Hence the jth sample logit

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

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

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

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

so that $E(Z_2 - Z_1) = \Delta + O(n_1^{-2}, n_2^{-2})$. For a similar estimator of Δ, 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

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

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

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

When $k = 2$, the expression $\Sigma\gamma_j(1-\gamma_j)(\pi_j+\pi_{j+1})$ reduces to the familiar formula $p(1-p)$ for the binomial variance. When $k = 3$, the weights are proportional to π_1 and π_3 respectively, π_2 being a measure of the correlation or information common to both $\tilde{\lambda}_{i1}$ and $\tilde{\lambda}_{i2}$.

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

(i) $\sum_{j=1}^{k-1}\gamma_j(1-\gamma_j)(\pi_j+\pi_{j+1})$; (ii) $\sum_{j=1}^{k}\pi_j(1-\gamma_j-\gamma_{j-1})^2$; (iii) $\sum_{j=1}^{k-1}\gamma_j\gamma_{j+1}\pi_{j+1}$,

(iv) $\sum_{j=1}^{k-1}(1-\gamma_j)(1-\gamma_{j-1})\pi_j$; (v) $\frac{1}{3} - \frac{1}{3}\sum_{j=1}^{k}\pi_j^3$.

These expressions are related to the intrinsic accuracy of the logistic function and also to the loss of information about Δ 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 γ by $F(x)$, π 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 γ_j estimated by $R_{.j}/n$ can however seriously underestimate the true variance when $|\Delta| > 1$.

The quantity Z_i with weights given by

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

is called the generalized empirical logit transform for the ith group. For the data of Table 1, the weights w_j, being proportional to category 1 and category 3 totals, are $w_1 = 0{\cdot}638$, $w_2 = 0{\cdot}362$, yielding $\tilde{\Delta} = 0{\cdot}565$ with standard error 0·225. The parameter Δ 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.

$$\hat{\Delta} = 0{\cdot}603 \pm 0{\cdot}225; \quad \hat{\theta}_1 = -0{\cdot}810 \pm 0{\cdot}116; \quad \hat{\theta}_2 = 1{\cdot}061 \pm 0{\cdot}118.$$

The residual deviance or likelihood ratio χ^2 statistic is $G^2 = 0{\cdot}302$ 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 χ^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

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

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

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

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

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

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

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

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

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

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 $\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x}_2 - \mathbf{x}_1)$ 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.

	Income group (000's)						
Year	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 $\log\{-\log(1-0{\cdot}065)\}$, $\log\{-\log(1-0{\cdot}147)\}$, etc. while those for 1970 are $\log\{-\log(1-0{\cdot}043)\}$,

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

$\log\{-\log(1-0{\cdot}103)\}$, 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

$$\log p_1(x) = \exp(0{\cdot}49)\log p_2(x), \tag{3.6}$$

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 $\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x}_2-\mathbf{x}_1)$ 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

$$\text{link}\{\gamma_j(\mathbf{x})\} = \theta_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \tag{4.1}$$

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, $\arctan\{\pi(\gamma-0{\cdot}05)\}$, 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 $\boldsymbol{\beta}$ describes how the log odds or other quantity of interest is related to the covariates $\mathbf{x}$. All models

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

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

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 $\boldsymbol{\beta}$ is scalar, the test is locally most powerful similar for one-sided alternatives. When $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta} = \mathbf{0}$. 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 $\boldsymbol{\beta}$ are regarded as nuisance parameters or when tests of the non-null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0 \neq \mathbf{0}$ 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

	Degree of suffering from disturbed dreams				
Age (yr)	*Very severe*	3	2	*Not severe*	*Total*
5–7	7	3	4	7	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 $\Sigma\gamma_j(1-\gamma_j)(\pi_j+\pi_{j+1})$ 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 $(0{\cdot}297n_i)^{-1}$ 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 $(0{\cdot}297n_i)^{-1}$. 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

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

and

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

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 β 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 χ^2 statistic is $G_1^2 = 12{\cdot}42$ 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 $G_2^2 = 7{\cdot}15$ on 8 degrees of freedom. The difference $G_1^2 - G_2^2 = 5{\cdot}27$ 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 $0{\cdot}80 = \exp(-0{\cdot}219)$. 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 χ^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 $\hat{\alpha}_1 \leqslant \hat{\alpha}_2 \leqslant \ldots \leqslant \hat{\alpha}_5$. 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

$$\text{link}(\gamma_{ij}) = (\theta_j - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i)/\tau_i, \tag{6.1}$$

where the quantity $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i$ is called the "location" for the ith row and τ_i is called the "scale" for the ith 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 $\dim(\beta) \geqslant t-1$ 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 $\Sigma \log \tau_i = 0$. The latter convention is adopted here and is particularly appropriate when we consider unsaturated scale models satisfying

$$\log \tau_i = \boldsymbol{\tau}^{\mathrm{T}}(\mathbf{x}_i - \bar{\mathbf{x}}), \tag{6.2}$$

where $\boldsymbol{\tau}$ 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 τ_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 $\hat{\Delta} = 0{\cdot}061$ with standard error 0·041 and $\log(\hat{\tau}_1/\hat{\tau}_2) = 0{\cdot}272$ 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				
	Highest	2	3	*Lowest*	*Total*
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 $(n_1, ..., n_k)$ to the likelihood function is $\pi_1^{n_1} ... \pi_k^{n_k}$ with the probabilities π_j satisfying (4.1) or, more generally, (6.1). Since we are dealing with cumulative probabilities, we define

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

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

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

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

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

and

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

whence the log likelihood is

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

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

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

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, $\mathbf{A} = \Sigma \mathbf{Q}_i \mathbf{Q}_i^{\mathrm{T}}$. 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 $n_i \to \infty$ 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 χ^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 $1, \ldots, k$ 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 × 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 × linear interaction term is more powerful against the anticipated departures than, say, Pearson's or the likelihood ratio χ^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 $\log\{p_1(x)\}/\log\{p_2(x)\} = 1{\cdot}63$. 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 × 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 × 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 × explanatory variable interaction term, the uniformly most powerful similar test for the presence of this type of interaction is based on the statistic $T = \Sigma n_{ij} \mathbf{x}_i s_j$, where $\mathbf{x}$ is the covariate vector and $s_j = j$ is the integer score attached to the jth 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 × 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 prob (carrier | 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 $(p-\hat{p})/(\hat{p}\hat{q}/n)^{\frac{1}{2}}$ 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

	Response category					
Treatment	*Terrible*	2	3	4	*Excellent*	*Total*
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 ($\tau_5 = 0{\cdot}51$) for this treatment. Treatment 3 rates worst overall and the consensus for this is also high ($\tau_5 = 0{\cdot}90$). 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 Univeristy 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 *Steptococcus pyogenes* among 2413 healthy children. *J. Hyg. Camb.*, **52**, 165–179.

KASTENBAUM, M. A. and LAMPHIEAR, 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 schemaparic 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

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

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

The derivative of the log-likelihood with respect to $\boldsymbol{\beta}^*$ 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 $\boldsymbol{\tau}$ 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

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

and the mixed second derivatives are

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

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

The Taylor series expansion for $\partial l / \partial \boldsymbol{\psi}$ gives

$$\frac{\partial l}{\partial \boldsymbol{\psi}} = \left. \frac{\partial l}{\partial \boldsymbol{\psi}} \right|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}} + \mathbf{A}(\delta \boldsymbol{\psi}) + ...,$$

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

$$\mathbf{A}\boldsymbol{\psi}_{n+1} = \mathbf{A}\boldsymbol{\psi}_n + \frac{\partial l}{\partial \boldsymbol{\psi}} = \mathbf{b}.$$

Writing $W = \log w = -\boldsymbol{\tau}^{\mathrm{T}}\mathbf{U}$ we have

$$(\mathbf{A}\boldsymbol{\psi})_r = nw^2 \sum_j V_j^{-1} q_{jr} \sum_s q_{js}\beta_s - nw \sum_s U_s \tau_s \sum_j V_j^{-1} q_{jr} q_j$$
$$= nw(1+W) \sum V_j^{-1} q_j q_{jr}$$

for $r \leqslant \dim(\boldsymbol{\beta}^*)$. Similarly, for $s > \dim(\boldsymbol{\beta}^*)$,

$$(\mathbf{A}\boldsymbol{\psi})_s = nWU_s \sum V_j^{-1} q_j^2 + nU_s \sum V_j^{-1} q_j^2$$
$$= n(1+W)\, U_s \sum V_j^{-1} q_j^2.$$

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 Θ underlying the response categories with distribution function $F_\Theta(\theta)$. Then for each row we can estimate F at those values of θ which correspond to the group boundaries, $\theta_1, \theta_2, \ldots, \theta_{k-1}$ say. The ranking of the θ's is known but not their actual values. We suppose that the distribution of Θ 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 θ, $\psi(.)$ 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

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

then

$$m(\theta_i) = a + b\psi^{-1}(p_i).$$

Now suppose we wish to compare any pair of distributions indexed by j ($= 1, 2$). Then for given i, $m(\theta_i)$ will have the same (but unknown) value for both distributions. Hence, by subtraction,

$$0 = a_1 - a_2 + b_1\psi_1^{-1}(p_{i1}) - b_2\psi_2^{-1}(p_{i2}).$$

Next suppose that $b_1 = b_2 = b$. Without loss of generality we may take $b = 1$ and then

$$a_1 - a_2 = \psi_2^{-1}(p_{i1}) - \psi_1^{-1}(p_{i2}) \quad (i = 1, 2, \ldots). \tag{1}$$

If this is constant for all i we can reasonably introduce covariates into the model by supposing that $a_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$ so that the left-hand side of (1) becomes $\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2)$. Since p_{ij} ($j = 1, 2$) can be estimated for all i the estimated values of the right-hand side of (1) can be used to fit the model.

Alternatively, if $a_1 = a_2$ then

$$b_1/b_2 = \frac{\psi_2^{-1}(p_{i2})}{\psi_1^{-1}(p_{i1})} \quad \text{or} \quad \log b_1 - \log b_2 = \log\psi_2^{-1}(p_{i2}) - \log\psi_1^{-1}(p_{i1}) \quad (i = 1, 2, \ldots). \tag{2}$$

If the right-hand side of (2) is constant for all i we can postulate $\log b_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j$ and fit the model using the estimated values of the difference of logarithms. Note that the method does not depend on knowing $m(\theta)$; this is as it should be since only the ranks of the x_i's are known.

The question now arises: can we find a distribution function ψ such that the differences in (1) or (2) are constant for all i?

Suppose we take

$$1 - F_\Theta(\theta) = \left\{1 + \frac{1}{u}\exp\left(\frac{m(\theta) - a}{b}\right)^{-1}\right\}^{-u} \quad \begin{pmatrix} u > 0, & -\infty < a < \infty \\ b > 0, & -\infty < m(\theta) < \infty \end{pmatrix}.$$

This includes the logistic ($u = 1$, $m(\theta) = \theta$), the exponential ($u = \infty$, $m(\theta) = \log\theta$) the log-logistic ($u = 1$, $m(\theta) = \log\theta$), Pareto Type II ($u > 1$, $m(\theta) = \log\theta$) and many other distributions. For (1) we have

$$a_1 - a_2 = \mathrm{logit}(1 - p_{2i})^{1/u} - \mathrm{logit}(1 - p_{1i})^{1/u} \tag{3}$$

and for (2)

$$b_1/b_2 = \{\log u - \mathrm{logit}(1 - p_{2i})^{1/u}\}/\{\log u - \mathrm{logit}(1 - p_{1i})^{1/u}\}. \tag{4}$$

When $u = 1$, (3) gives the author's log-odds model; in the limit as $u \to \infty$ we have the proportional hazards model. In addition we have a range of intermediate cases and by estimating u we are allowing the best model to be selected by the data. Equation (4) leads to a parallel class of models. Models with values of u between the two extremes are not so easy to interpret and one may well wish to follow the author in using whichever of $u = 1$ and $u = \infty$ gives the better fit.

In one case, at least, the family of functions ψ can be restricted *a priori* by invoking the reversibility considerations touched on in Section 4.2. The question here is: is the nature of the underlying variable such that one would want the conclusions to be the same if the columns of the table were reversed? The answer to this, I suggest, does not depend on the purpose of the analysis but on the arbitrariness or otherwise of the direction of measurement of the latent variable. Thus, for example, with a variable such as political hue it is completely arbitrary whether we measure from left to right or *vice versa*. If we require reversibility then, in the location case, ψ must be such that

$$\psi^{-1}(p_2) - \psi^{-1}(p_1) = \psi^{-1}(1 - p_1) - \psi^{-1}(1 - p_2)$$

for all p. In other words, ψ must be the distribution function of a symmetrical random variable and this, in turn, implies that $u = 1$ in the above family. In other words, the log-odds model is appropriate.

This conclusion renders it somewhat surprising that the log-odds model should have been fitted to Table 1 where the variable is a measure of size for which the direction of the scale of measurement does not seem to be arbitrary. I have therefore investigated this example using the more general model with the following results for the relevant logit differences of (3):

$u = 1$ (*McCullagh*)	$u = 3$	$u = \infty$ (*Proportional hazards*)
−0·498	−0·455	−0·472
−0·684	−0·462	−0·373

It could be maintained that the proportional hazards model (with $u = \infty$) was at least as good as the log-odds model but that the intermediate case $u = 3$ is better than either.

In spite of their incompleteness I hope that these remarks will help to reveal some of the richness implicit in the author's approach. In proposing the vote of thanks I warmly welcome this paper and look forward in hopeful anticipation to the next.

Dr P. M. E. ALTHAM (Statistical Laboratory, Cambridge University): I congratulate Dr McCullagh on an important, stimulating and very practical paper. I like especially his emphasis on quantitative interpretation and description, rather than on significance tests alone. Another feature of his approach to ordered categories that I find appealing is the fact that his models behave nicely when adjacent categories are pooled. Invariance considerations are important here; in a psychological experiment, for example, a subject may be asked to give a response taking one of five ordered values, and the experimenter may later decide to group these into three classes. As Dr McCullagh points out, it would be difficult to handle (and probably difficult also to interpret) log-linear models for such data.

I have three main comments to make.

(1) The idea of the generalized "residuals" given in Section 8.2 is an interesting one, and it would be good to know more about the distribution of these. I think they are really only useful if the sample sizes for the various rows of the table are roughly similar, as is indeed the case in the numerical examples given. If one had rather heterogeneous sample sizes, then interpreting large "residuals" for given rows might be tricky, since if the model does not fit, the expected value of any row "*residual*" will increase as the corresponding row *total* increases.

(2) Ordered categorical data have been extensively considered in psychological experiments, in particular in *signal detection theory*, as some members of the audience may well know. Suppose a subject is asked to discriminate, in an experiment on auditory perception, between two tones, called A and B say, and he can give one of k responses, ranging from "certainly A" (response 1) through "possibly A" to "certainly B" (response k).

	Response 1 . . . . . k	
signal A presented	n_{11} ... n_{1k}	n_1
signal B presented	n_{12} ... n_{2k}	n_2

Assume that the signals A, B are presented, independently, to the subject n_1 and n_2 times, and the data of responses are recorded in the $2 \times k$ contingency table above.

One of the statistical problems is to estimate the "separation" between signals A and B, and for this problem Dr McCullagh's methods might be very suitable. A graphical technique the psychologists find useful is to plot the *cumulative* probabilities against each other, yielding a graph of $(k+1)$ points; this graph being the ROC (receiver operating characteristic) curve. I mention this since to plot such a graph might be found more generally useful, if the number of categories k is not too small.

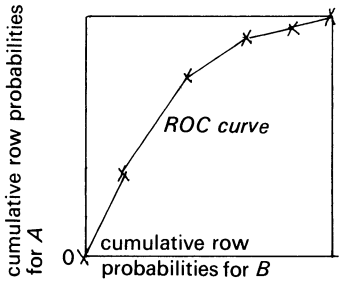

FIG. D1.

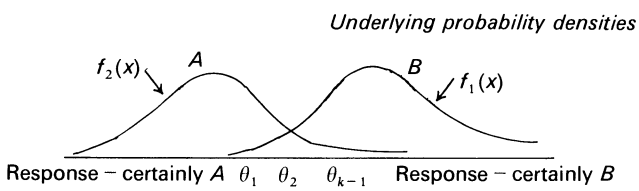

FIG. D2.

If our underlying model is that the cumulative row probabilities are

$$F_2(\theta_1) \ldots F_2(\theta_{k-1}), 1 \text{ for } A \quad \text{and} \quad F_1(\theta_1) \ldots F_1(\theta_{k-1}), 1 \text{ for } B,$$

where $\theta_1, \ldots, \theta_{k-1}$ are the subject's unknown "threshold" values, and F_1, F_2 are distributions, then,

neglecting sampling variation, our ROC curve is a graph of

$$y = F_2(\theta_i) \text{ against } x = F_1(\theta_i) \text{ as } \theta_i \text{ varies, together with the points } (0,0) \text{ and } (1,1).$$

In practice this graph often looks roughly *concave*—and this feature is telling us something about suitable choices for $F_2(.), F_1(.)$ in our model. It can in fact be proved (Altham, 1973) that the graph $y = F_2(F_1^{-1}(x)), 0 \leqslant x \leqslant 1$, is concave if and only if the likelihood ratio $f_2(x)/f_1(x)$ is a decreasing function of x. If the two underlying distributions differ only by a shift parameter Δ say, so that $f_2(x)/f_1(x) = f(x)/f(x-\Delta)$, then it is known that the monotone likelihood ratio property is closely connected with the property

$$f'(x)/f(x) \text{ is a decreasing function of } x. \tag{i}$$

In signal detection theory, models for which (i) holds have some rather nice properties, regardless of the actual form of $f(.)$; see Thomas and Myers (1972) and Altham (1973). Many of the "link" functions suggested by Dr McCullagh correspond to probability densities $f(.)$ for which the property (i) holds; I wonder if he could exploit this feature to get any optimality results? Incidentally, maximum likelihood estimation of Δ, and a scale parameter σ also, has been considered in the psychological context by Grey and Morgan (1972).

(3) The results of Section 2.3 may readily be generalized to the case of an arbitrary underlying distribution function $F(.)$. If the cumulative row frequencies are $\{n_1 c_{1j}, n_2 c_{2j}\}, j = 1, ..., k$, and if our model is that $E(c_{1j}) = F(\theta_j - \frac{1}{2}\Delta)$, $E(c_{2j}) = F(\theta_j + \frac{1}{2}\Delta)$, then if we put $\lambda_{1j} = F^{-1}(c_{1j})$, $\lambda_{2j} = F^{-1}(c_{2j})$, and take $\tilde{\Delta} = \sum_{j=1}^{k-1} w_j(\lambda_{2j} - \lambda_{1j})$ as our weighted estimator, and choose (w_j) to minimize $\text{var}(\tilde{\Delta})$ when $\Delta = 0$, we find that

$$w_j \propto f(\theta_j)\left[\frac{f(\theta_j)-f(\theta_{j-1})}{F(\theta_j)-F(\theta_{j-1})} - \frac{f(\theta_{j+1})-f(\theta_j)}{F(\theta_{j+1})-F(\theta_j)}\right]$$

for $j = 1, ..., k-1$, where $\theta_0 \equiv -\infty$, $\theta_k \equiv \infty$.

In this case

$$\text{var}\,\tilde{\Delta} = \text{function of } (n_1, n_2)\left\{\sum_{j=1}^{k} \frac{(f(\theta_j)-f(\theta_{j-1}))^2}{[F(\theta_j)-F(\theta_{j-1})]}\right\}^{-1}$$

and for large k, the summation becomes

$$\int_{-\infty}^{\infty} \left(\frac{f'(\theta)}{f(\theta)}\right)^2 f(\theta)\, d\theta,$$

which is not at all surprising, when we consider the Cramér–Rao lower bound for estimating a location parameter Δ given two independent random samples from densities $f(x-\frac{1}{2}\Delta), f(x+\frac{1}{2}\Delta)$ respectively.

It is clear that I have found the paper very interesting, and I have much pleasure in seconding the vote of thanks to Dr McCullagh.

The vote of thanks was passed by acclamation.

Dr J. A. ANDERSON (University of Newcastle): I would also like to congratulate Dr McCullagh on his presentation of a stimulating and useful paper, containing a very acceptable combination of theory and application.

The range of examples given was impressive and illustrated some of the practical problems very nicely. All the examples followed the traditional pattern of examining the dependence of an ordered variable on one other variable, whereas it is clear from (2.1) and (4.1) that the dependence may be on arbitrary linear functions of variables. This includes multiple regression and the analysis of designed experiments with ordered response variables. A word of warning is necessary, before these applications are tried. As the number of parameters increases, the chance of the sparse data problem occurring (see the end of Section 6) also increases. This is particularly marked with designed experiments. For example, I tried to analyse a randomized blocks experiment with 20 blocks, 5 treatments and 4 ordered categories of response. The estimates of six of the β-parameters were unbounded. This will certainly occur when all the responses in a block or treatment are the same and in either extreme category (1 or k). This does *not* correspond to an inappropriate model; it is a sparse data problem similar to those noted for binary responses by Anderson (1974).

A problem in discriminant analysis is how to distinguish between ordered groups. For example, we may wish to predict the degree of response (Y) to a treatment (good, poor or indifferent) given certain initial information ($\mathbf{x}$) on a patient. This paper provides suitable models for $\Pr(Y = y \mid \mathbf{x})$. Anderson and Philips (in a paper not yet published, 1980) have shown that this approach is feasible in a back pain prognosis study with six ordered, response categories and up to nine predictor variables. More work is needed on the estimation of Y. The classical approach of choosing $\hat{Y} = s$ to maximize over t the estimated probabilities $\{\Pr(Y = t \mid \mathbf{x})\}_{\text{est}}$ is unsatisfactory because it gives equal losses to errors from $Y = s$ to $Y = t$ $(t \neq s)$, irrespective of how far t is from s. Unfortunately, there is no general loss structure that can be agreed for this estimation problem. An *ad hoc* estimate for Y based directly on $\hat{z} = \hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}$ is perhaps preferable. An obvious choice is $\hat{y} = s$ if $\hat{\theta}_{s-1} \leqslant \hat{z} < \hat{\theta}_s$.

There are many unsolved problems in this area; Dr McCullagh's paper provides a good step forward and a stimulus to further work.

Professor M. AITKIN (University of Lancaster): I would like to make three points on this interesting and useful paper. The first is a general one concerning the relation between the proportional odds or hazard models and the underlying distribution of a latent variable. This point was discussed in the Introduction but disappears from view in the body of the paper. We may use the distribution of the latent variable, if it is convenient, to simplify the computation of the maximum likelihood estimates. In the simple probit model, the EM algorithm may be used for parameter estimation instead of the general Newton method with Fisher scoring. In the multinomial models considered in this paper, the cut points θ_j must be estimated together with the regression coefficient estimates, and this makes the EM algorithm much less convenient, though still possible.

The second point concerns the proportional odds and proportional hazards models fitted to the income distribution. Since income is measured on a continuous scale, there is a real income distribution, which has been grouped in Table 3. Why not examine the c.d.f. in the usual way by probability plotting? The failure of the proportional odds model to fit simply reflects the fact that the income distributions are not logistic with the same scale parameter but different location parameters, as is clear from a normal probability plot. A lognormal model is clearly wrong, but a Weibull plot shows a very close fit with nearly equal shape parameters, so that a proportional hazard model will fit well. It is quite possible, however, that the proportional hazard model would *not* fit, because of a close fit to two Weibulls with different shape parameters. In this case it is not useful to know that a proportional hazard model does not fit, but it is useful to know that the distributions are Weibull.

The final point concerns Table 9. If we assume an interval scale for the response category (i.e. the cut points are *fixed* at 1·5, 2·5, ..., 4·5), then it is a simple matter to fit grouped normal distributions using the EM algorithm, with different location and scale parameters for each treatment. The parameter estimates and goodness of fit of the models are shown in Table D1. The fit is worse for the grouped normal than for the logistic model in Table 10 because of the strong metric assumption, but the parameter estimates and conclusions are very similar from both models. Neither model can fit the U-shaped response distribution for Treatment 4.

TABLE D1

Means, standard deviations and deviances for data of Table 9

Treatment	*Mean*	*S.D.*	*Deviance*
1	2·89	1·54	5·8
2	3·01	1·26	8·2
3	2·00	1·27	5·4
4	2·52	1·96	9·8
5	3·72	0·55	5·3
		Total	34·5

Mr J. BURRIDGE (Imperial College and Post Office Telecommunications): Several speakers have mentioned that they have experienced little difficulty in fitting the linear models described by Dr McCullagh in tonight's paper. This is probably because the log-likelihood is concave with respect to the

parameters for most of the models considered in his paper. Thus the problem of multimodality of the likelihood does not arise (Section 6.2) and maximum likelihood estimates are generally unique except when the data matrix contains certain patterns of zeros as noted in the paper (Section 6.3). The result can be seen as follows:

For the linear model the log-likelihood is

$$L = \sum_{ij} n_{ij} \ln F(\theta_{j-1} - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i,\ \theta_j - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i),$$

where

$$F(u, v) = \int_u^v f(\varepsilon)\, d\varepsilon$$

and $f(\varepsilon)$ is the "error" density which is assumed given. A recent result (Burridge, 1980) shows that $\ln F$ is concave with respect to (u, v) if $\ln f$ is concave with respect to ε. Thus L is concave with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta})$. For example, this applies to the models

$$\text{logit: } f(\varepsilon) = \exp(\varepsilon)\{1+\exp(\varepsilon)\}^{-2},$$
$$\text{probit: } f(\varepsilon) = \exp(-\varepsilon^2/2)/\sqrt{(2\pi)},$$
$$\text{and complementary log-log: } f(\varepsilon) = \exp\{\varepsilon - \exp(\varepsilon)\},$$
$$\text{but } \textit{not} \text{ to Cauchy: } f(\varepsilon) = 1/\{\pi(1+\varepsilon^2)\}.$$

Unfortunately, the result does not apply to the non-linear models considered in tonight's paper so we have to rely on the results for *expected* second derivatives given in Section 6.3 of the paper.

Dr J. T. Smith (Queen's University, Kingston, Ontario): I would like to add my congratulations and thanks to Dr McCullagh for his lucid treatment of the ordinal data problem. My comments concern the choice of link function for the analysis of grouped income data.

Dr McCullagh recommends the logit and complementary log-log functions for ease of interpretation. In Section 3.2 he notes that the latter function is supported by the apparent fit of the proportional hazards model to the given illustrative data and by the precedent of using the Weibull and Pareto distributions to fit the body and upper tail, respectively, of grouped income data. If historical precedent is to be invoked, there are grounds for favouring proportional odds, at least for the body of the data (Fisk, 1961). Where experience suggests that other parametric distributions are preferred, for example, lognormal, gamma (Salem and Mount, 1974) or Burr (Singh and Maddala, 1976), what are the counterparts of the proportional odds and proportional hazards properties by which one may select an appropriate link function? It may be helpful to make the connection between the generalization of the proportional odds and proportional hazards properties, rewritten from equation (4.1) as

$$\text{link}\{F(t;\ \mathbf{x})\} = \text{link}\{F_0(t)\} - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x},$$

and the property sometimes exploited by econometricians in fitting parameteric distributions, namely

$$\text{link}\{F(t;\ \theta_1, \theta_2)\} = \theta_1 + \theta_2\, h(t),$$

where h is a monotone function (for example, scale and shape parameters of the log-logistic distribution are estimated by weighted regression of the logit on $\log t$).

In studies of income data, the issue is often not whether the distribution as a whole has changed, for inflation and other factors usually guarantee that it has, but by how much some aspect of the distribution, such as relative inequality, has changed. What are the consequences of choosing a particular link function for inferring changes in measures of concentration such as the Gini index?

Dr W. J. R. Eplett (University of Birmingham): The main comment I should like to make concerns the robustness of the estimators for $\boldsymbol{\beta}$ in the proportional odds model (Section 2.1). These estimators are defined using the linear combinations

$$Z_i = \sum_j w_j \bar{\lambda}_{ij},$$

where $\tilde{\lambda}_{ij}$ is defined in Section 2.3. For given $\varepsilon > 0$, define

$$\tilde{\lambda}_{ij}^{(r)}(\varepsilon) = \log[\{(1-\varepsilon)\tilde{\gamma}_{ij} + (2n_i)^{-1}\}/\{1-(1-\varepsilon)\tilde{\gamma}_{ij} + (2n_i)^{-1}\}] \quad (1 \leqslant j < r)$$
$$= \log[\{(1-\varepsilon)\tilde{\gamma}_{ij} + \varepsilon + (2n_i)^{-1}\}/\{1-(1-\varepsilon)\tilde{\gamma}_{ij} - \varepsilon + (2n_i)^{-1}\}] \quad (r \leqslant j \leqslant k),$$

where

$$\tilde{\gamma}_{ij} = R_{ij}/n_i,$$

the cumulative proportions for the ith explanatory variable. When $\varepsilon = 0$, $\tilde{\lambda}_{ij}^{(r)}(\varepsilon) = \tilde{\lambda}_{ij}$ and for $\varepsilon > 0$ it represents the effect upon $\tilde{\lambda}_{ij}$ of a proportion ε of contamination in the rth category of the response variable. The effect of this contamination upon the value of Z_i is assessed through

$$Z_i^{(r)}(\varepsilon) = \Sigma w_j \tilde{\lambda}_{ij}^{(r)}(\varepsilon).$$

In order to assess the influence of the contamination we examine the derivative

$$\mathrm{IF}(Z_i, r) = \frac{d}{d\varepsilon} Z_i^{(r)}(\varepsilon)\big|_{\varepsilon=0} = -\sum_{j<r} w_j/(1-\tilde{\gamma}_{ij}) + \sum_{j \geqslant r} w_j/\tilde{\gamma}_{ij}$$

which is the dominant term of the Taylor series expansion for $Z_i^{(r)}(\varepsilon)$ in powers of ε provided of course that ε is small. Here n_i is assumed to be sufficiently large so that the term $(2n_i)^{-1}$ can be ignored; this term is the result of a modification to the natural estimator anyway (Cox, 1970, p. 33).

It should be stressed that it is a worthwhile exercise discussing the sensitivity towards contamination of estimators for $\boldsymbol{\beta}$ since when the response variable is of the kind found in Tables 1 and 9, for instance, a certain amount of misclassification appears inevitable. Deciding upon tonsil size or the taste of a particular brand of tea must be quite difficult, and indeed fairly arbitrary, in some instances. The influence function studies the effect of this variation upon the estimator for $\boldsymbol{\beta}$.

In order to simplify the discussion, consider the case where there are two explanatory categories. Then the estimator for Δ given by Section 2.3 is $\hat{\Delta} = Z_2 - Z_1$. Define

$$\hat{\Delta}^{(r,r')}(\varepsilon) = Z_2^{(r')}(\varepsilon) - Z_1^{(r)}(\varepsilon),$$

which describes the effect upon the estimator of contamination in the rth category of the response variable for explanatory variable 1 and contamination in category r' of the response variable for explanatory variable 2. The influence function for the estimator is then given by

$$\frac{d}{d\varepsilon}\hat{\Delta}^{(r,r')}(\varepsilon)\big|_{\varepsilon=0} = \mathrm{IF}(Z_2, r') - \mathrm{IF}(Z_1, r)$$
$$= \sum_{j<r} w_j\{(1-\tilde{\gamma}_{1j})^{-1} - (1-\tilde{\gamma}_{2j})^{-1}\} + \sum_{j \geqslant r'} w_j\{\tilde{\gamma}_{2j}^{-1} - \tilde{\gamma}_{1j}^{-1}\}$$
$$- \sum_{r \leqslant j < r'} w_j\{\tilde{\gamma}_{1j}^{-1} + (1-\tilde{\gamma}_{2j})^{-1}\}$$

provided $r \leqslant r'$, with a similar expression in the case where $r > r'$.

Let me try and interpret this expression for the influence function. Provided Δ itself is reasonably small, the first two terms of the right-hand side are not going to contribute significantly to the expression. The third term can be expected to be large in certain cases. The worst cases for the effect of contamination of the data upon the estimator occur when the contamination occurs in a low category of the response variable for one of the explanatory variables and in a high category of the response variable for the other explanatory variable (using the ordering of the categories of the response variable). The effect of contamination is particularly pronounced if any of the categories in the third term involve small γ_j (near 0) or large γ_j (near 1). This seems to shed some light upon the choice of the weights. They should damp down these extreme cases and this appears to be achieved to some extent by the optimal weights defined by (2.5) through the term $\gamma_j(1-\gamma_j)$ appearing there. It may be that different weights which damp down the extremes even more might be appropriate in some cases.

The influence function when more than two explanatory variables are involved is rather more complicated, but the basic feature of sensitivity to contamination at the extremes remains.

Finally I should very much like to thank Dr McCullagh for a most interesting paper.

Dr A. C. ATKINSON (Imperial College, London): Like other speakers I congratulate Dr McCullagh on a paper which contains an exemplary blend of statistical theory and analysis of data. Also like several other speakers, I have three points to make.

(1) There has been some discussion this evening of derivations of the competing models. In an incorrectly titled paper Aitchison and Bennett (1970) derived a model for ordinal data using the normal distribution. I wonder whether a parallel derivation using the logistic distribution would be illuminating.

(2) The eye data in Table 8 is famous and seemingly inexplicable. However, Heim (1970) has a chapter with the resonant title "The Mediocrity of Women" in which she discusses the finding that, for many characteristics, men tend to be more extreme than women, even if the means of the two sexes are the same. Her main concern is with intelligence, but it would be fascinating if the same property of higher male variance is really to be found in such a basic physiological characteristic as eye-sight.

(3) Several speakers have suggested that Dr McCullagh was prudent not to have entitled his paper "Ordinal Data from Designed Experiments" and have stressed the problems which arise when including, for example, block effects in the logistic model. However, in McCullagh (1977) an analysis is given of matched pairs in which empirical Bayes methods are used with a parameter for each pair. This model leads to appreciable, if entertaining, numerical problems. Empirical Bayes methods were not mentioned tonight. I wonder what Dr McCullagh's current thinking is about them.

Mr IAN PLEWIS (University of London Institute of Education): Dr McCullagh's proportional odds model considers $\log(\Sigma_{j=1}^{m}\pi_j/1-\Sigma_{j=1}^{m}\pi_j)$, $m=1,\ldots,k-1$. Fienberg and Mason (1979), in an analysis of ordered educational levels, suggest continuation odds i.e. $\log\{(\Sigma_{j>m}\pi_j)/\pi_m\}, m=1,\ldots,k-1$, which is the odds of getting beyond a certain level of education, m, given that one got as far as m. The two models have similar aims and I wonder how one should choose between them.

The following contributions were received in writing, after the meeting.

Professor A. AGRESTI (University of Florida, Gainesville): Dr McCullagh is to be congratulated for an interesting and important article. Too much of the literature on categorical data analysis makes no distinction between nominal and ordinal scales. Also, I believe there has been too much emphasis on testing goodness-of-fit compared to model building and parameter estimation.

Dr McCullagh makes an important point in Section 7 in arguing that the results of fitting a particular model should be robust to the definition of categories of the response variable. However, I believe he maligns too harshly the log-linear model in this respect. If categories are redefined or if certain categories are grouped together, results are not invariant because the equal-interval scores $1, 2, \ldots, R$ are no longer appropriate. The cross-ratio per unit distance on the response (not necessarily for adjacent categories) would still be a meaningful measure, however, when the cores have been reassigned to reflect the new grouping. In other words, one could assume in the two-sample cases that there is a parameter θ such that the odds ratio for adjacent categories having s_1 and s_2 is $\theta^{s_2-s_1}$. The estimated value of θ should be approximately the same no matter how the response categories are defined, if scores are assigned in a sensible manner. Of course, one important advantage of the models McCullagh discusses is that it is unnecessary to assign scores to the response categories.

For the proportional odds model with values $x_1 < x_2$ on a single covariate x, note that $\kappa_j(x_1)/\kappa_j(x_2)$ is the ratio of the probabilities of concordance and discordance for the 2×2 table obtained by dichotomizing the response at the jth category. This ratio is assumed to be constant for all cutpoints, $j = 1, 2, \ldots, k-1$. Now,

$$P_c = \sum_{i<j} \pi_i(x_1)\pi_j(x_2) / [\sum_{i<j} \pi_i(x_1)\pi_j(x_2) + \sum_{i>j} \pi_i(x_1)\pi_j(x_2)]$$

is the overall probability of concordance obtained using the full set of k response categories, for a randomly selected pair of observations (one observation at each x_i) which is untied on the response. A simple one-parameter model which I have not seen proposed, but which relates to the traditional analysis of ordinal data through the notions of concordance and discordance of pairs, is a logistic one in P_c; that is, $\log[P_c/(1-P_c)] = \beta(x_2 - x_1)$. I conjecture that this model fits reasonably well in a variety of settings in which models of the "link" form provide decent fits. This model does not assume a constant difference between link values at each cutpoint of the response, though, so it may be more widely applicable than any single model of the "link" form.

Dr D. ANDRICH (University of W. Australia): The imposed ordering of the categories in the proportional odds and other related models may be a virtue in circumstances where the variable and threshold points, such as in survival or income data, are clear. However, I wonder whether this same feature may not be a weakness in cases where neither the variable nor the classification system is so explicit and where, as a result, the categories may not necessarily operate as ". . . contiguous intervals on a continuous scale".

The data on disturbed dreams (Table 6), where the conclusion is that severity of disturbance tends to decrease with age, are of this type. Now assuming that severity of disturbance shows a decrease with age across the categories in general, then one would expect it also to show a decrease with age across any pair of adjacent categories in particular. Instead, conditional on classification in one of the two middle categories, and quite independently of age, the allocation to one of these two categories appears to be random. Thus these two categories appear not to operate consistently with the extreme categories and therefore presumably not consistently with the intended ordering. At best, the supposed distinction between these two categories seems to be of no value. Perhaps at worst, if a different treatment were prescribed corresponding to the different categories, assignment of a subject to one of the two treatments corresponding to the middle categories would also be at random. Rather than being of no interest, the middle cut point in these data indicates a need for further examination. Of course, this further examination would be of a substantive rather than of a statistical kind.

I concur strongly with Dr McCullagh's emphasis on the importance of distinguishing between explanatory and dependent variables and on the importance of using models connecting variables to laws generating data. In some cases, these requirements may include scrutinizing the presumed ordering properties of the variable concerned. Unfortunately, because of the strict ordering of the categories essentially ensuring that $\hat{\theta}_j > \hat{\theta}_{j-1}$, the models of (4.1) with which he has dealt so comprehensively do not expose any anomalies in category ordering. The ordering of categories is a property of these models whether or not it is an actual property of the data. Therefore, may I suggest that models which permit the categories to *reveal* themselves to be ordered or otherwise, rather than those which *constrain* the categories to be ordered, may be more instructive for fully understanding certain types of ordinal variables.

Mr F. J. ARANDA-ORDAZ (Imperial College, London): I want to congratulate the author for the ingenious way in which he has connected different models for the analysis of response data. I wish to make two comments only. The first one is in relation to expression (4.1) in the paper. For binary data it is possible to use a relationship very similar to (4.1) and, for a convenient definition of the link function, provide a quantitative assessment of a scale where a simple decomposition in terms of the systematic component of the model is more plausible. I have obtained good results with two parameterized (one indexing parameter) link functions which allow discrimination between the logistic scale, and symmetric or asymmetric alternatives for the decomposition of the probabilities. To determine the form of the corresponding parameterized link function, a linear scale in the probabilities has been used in the first case and the complementary log-log in the second. It seems that the use of expressions similar to (4.1) may lead to comprehensive comparisons of a broad range of alternative models in a simple way.

My second comment refers to Section 7.2. I warmly endorse Dr McCullagh's opinion about the current tendency to disregard logit models for binary data, just because they may be obtained as a special case of the log-linear model. I personally believe that the different objectives which lead to the construction of logit models justify their use and make them irreplaceable.

Dr V. FAREWELL (Fred Hutchinson Cancer Research Center): This is an excellent paper in both content and presentation and I have only a minor technical comment. In the analysis of epidemiological case-control studies it is advantageous to be able to make inferences on a model for disease incidence conditional on an explanatory variable although the sampling distribution is of explanatory variable conditional on disease status. Although both the continuous proportional hazards model and the discrete logistic model allow such inferences (Prentice and Breslow, 1978; Farewell, 1979) the necessary inversion does not appear to be possible with the complementary log-log model.

A discussion and example of the use of the complementary log-log model with survival data is given in Prentice and Gloeckler (1978).

Professor S. E. FIENBERG (University of Minnesota): This is a thoroughly enjoyable paper. Dr McCullagh presents a clear and well-reasoned argument for the use of his regression-like models for the analysis of ordinal data, and thus adds a valuable tool to the statistician's workbench. My only complaint is that, in attempting to describe the virtues of his approach, he implies that other approaches such as those

associated with loglinear and logit models are less virtuous and have little to recommend them. I would dispute such a suggestion for at least three reasons:

(1) The ordinal nature of some categorical variables is often crucial to the structural organization of categorical data subjected to loglinear analysis, as in triangular arrays (Bishop and Fienberg, 1969), and social mobility tables (Goodman, 1972, 1979a), and age-period-cohort structures (Fienberg and Mason, 1978). The fit of loglinear models to such structures *is not* permutation invariant, despite McCullagh's insistence that it is.

(2) Analyses of multidimensional structures using loglinear and logit techniques can easily take ordinal structure into account, although the results will not necessarily be invariant under a reversal of categories.

(3) The use of ordinal information in McCullagh's models effectively assumes that the underlying latent variable does not possess natural discontinuities or shifts in terms of its relationship with other variables—a matter for empirical investigation.

Let me elaborate briefly on points 2 and 3 using the tonsils data. Instead of McCullagh's model, (2.4), I propose to look at logits for the continuation rates (see Fienberg, 1980, Chapter 6), which correspond here to components of the factorization of the likelihood in Section 6.2:

$$\log\left(\frac{\pi_{11}}{\pi_{12}+\pi_{13}}\Big/\frac{\pi_{21}}{\pi_{22}+\pi_{23}}\right) \quad \text{and} \quad \log\left(\frac{\pi_{12}}{\pi_{13}}\Big/\frac{\pi_{22}}{\pi_{23}}\right).$$

The first of these appears in McCullagh's analysis and is the log-odds ratio for "not enlarged" *vs* "enlarged"; the second is for "greatly enlarged" *vs* "moderately enlarged". We ask if these two log-odds ratios have a common value Δ. The corresponding empirical estimates (without the added $\frac{1}{2}$'s) are $-0{\cdot}514$ and $-0{\cdot}543$, and this model of equality of log-odds ratios fits the data even better than does McCullagh's. Actually, the two models are compatible, and when taken together imply that $\log(\pi_{11}\pi_{22}/\pi_{12}\pi_{21}) = 0$. A direct test of this yields a χ^2_1 value of about 1, suggesting that there is a basic shift in the effect of being a carrier on the underlying latent variable; there appears to be no effect until the tonsil enlargement reaches a threshold.

Finally, I note that McCullagh skips rather blithely over the critical issues of computation, and of existence and uniqueness of estimates. Since many important practical problems involving ordinal data are large and sparse, these topics bear further attention and more detailed exposition.

Professor S. J. HABERMAN (University of Chicago): Dr McCullagh's already strong case for his method of analysis of ordered multinomial responses can be strengthened through some results concerning existence and uniqueness of maximum likelihood estimates. Let Y_i, $1 \leqslant i \leqslant n$, be independent random variables with the integers 1 to $k \geqslant 2$ as their common range. Let $F(\lambda_{ij})$ be the probability that $Y_i \leqslant j$, $1 \leqslant j \leqslant k-1$, $1 \leqslant i \leqslant n$, where F is a known distribution function and the λ_{ij} are unknown parameters. Assume that F has a derivative f such that $\log f$ is strictly concave, and assume that the λ_{ij} satisfy an additive linear model

$$\lambda_{ij} = \theta_j - \sum_{r=1}^{p} \beta_r x_{ir}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant k-1,$$

for known independent variables x_{ir}, $1 \leqslant i \leqslant n$, $i \leqslant r \leqslant p$, unknown parameters θ_j, $1 \leqslant j \leqslant k-1$, and β_r, $1 \leqslant r \leqslant p$, where $\theta_j < \theta_{j'}$ whenever $j < j'$. Let T_{ij} be 1 if $Y_i = j$ and 0 otherwise. It is well known that $\log F$ and $\log(1-F)$ are strictly concave if $\log f$ is strictly concave, and one can show that $\log[F(x)-F(y)]$ is strictly concave for $y < x$. Thus $\hat{\theta}_j$, $1 \leqslant j \leqslant k-1$, and $\hat{\beta}_r$, $1 \leqslant r \leqslant p$, are maximum likelihood estimates of θ_j, $1 \leqslant j \leqslant k-1$, and β_r, $1 \leqslant r \leqslant p$, respectively, if and only if $\hat{\theta}_j < \hat{\theta}_{j'}$ whenever $j < j'$ and the following equations hold:

$$\sum_{i=1}^{n} T_{ij}\,\hat{a}_{ij}/\hat{\pi}_{ij} = \sum_{i=1}^{n} T_{i(j+1)}\,\hat{a}_{ij}/\hat{\pi}_{i(j+1)}, \quad 1 \leqslant j \leqslant k-1,$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ir}\,T_{ij}(\hat{a}_{ij} - \hat{a}_{i(j-1)})/\hat{\pi}_{ij} = 0, \quad 1 \leqslant r \leqslant p.$$

Here for $1 \leqslant j \leqslant k-1$, $\hat{a}_{ij}$ is the density $f(\hat{\lambda}_{ij})$ at the estimate $\hat{\lambda}_{ij} = \hat{\theta}_j - \sum_{r=1}^{p} \hat{\beta}_r x_{ir}$ of λ_{ij}. If $j = 0$ or $j = k$, $\hat{a}_{ij} = 0$. The estimated probability that $Y_i = j$ is $\hat{\pi}_{ij} = F(\hat{\lambda}_{ij}) - F(\hat{\lambda}_{i(j-1)})$, where $\hat{\lambda}_{ik} = \infty$ and $\hat{\lambda}_{i0} = -\infty$. The maximum likelihood estimates exist if and only if $\sum_{i=1}^{n} T_{ij} > 0$, $1 \leqslant j \leqslant k$, and if the conditions $t_{ij} \geqslant 0$, $Y_i = j$,

$t_{ij} \leqslant 0$, $Y_i = j+1$, $1 \leqslant i \leqslant n$; $t_{ij} = \delta_j - \Sigma_{r=1}^{p} \gamma_r x_{ir}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant k-1$, $\delta_1 < \ldots < \delta_{k-1}$, can only hold for some δ_j and γ_r if all the t_{ij} are zero. In particular, existence is assured if the model is applied to a contingency table in which all cells have positive counts. If the estimates exist, they are unique if and only if the λ_{i1}, $1 \leqslant i \leqslant n$, uniquely determine the β_r, $1 \leqslant r \leqslant p$. The arguments required for these results are similar to those of Haberman (1974, pages 309 and 320–321).

In the cumulative logit model, $F(x) = 1/(1+e^{-x})$, so that $\log f(x) = -x - 2\log(1+e^{-x})$ is strictly concave. In the proportional hazard case, $F(x) = 1 - \exp\{-\exp x\}$ and $\log f(x) = x - \exp x$ is strictly concave. One may also use cumulative normal models with $F(x)$ the distribution function $\Phi(x)$ of the standard normal. Unfortunately, the argument just made does not apply to the nonlinear case of Section 6.1.

Dr R. R. HARRIS (University of Exeter): Like most of the speakers at the meeting I was stimulated by this interesting paper, in particular because equation (3.5) has been fundamental to work undertaken at the University of Exeter to the development of a method of analysis for survival data with time-dependent covariates. In this work, undertaken with Dr K. L. Q. Read and a research student, A. A. Noura, we divided the time domain into suitably defined intervals and upon consideration of the proportional hazards model were clearly led, like Dr McCullagh, to equation (3.5). By noting that in the general case (with $k > 2$) the number of parameters required for equation (3.5) is less than the number required for a saturated model for an N (say) $\times k$ contingency table we have, in a paper about to be submitted for publication, used these extra parameters to model violations of the proportional hazards model. More than one covariate may readily be incorporated into the analysis by use of a suitably defined new cross-classification of risk groups and the flexibility available in such a manoeuvre (there is no unique saturated model) allows for various parameterizations with corresponding interpretations, for example hierarchical or ANOVAR-type.

On another aspect of Dr McCullagh's paper, I wonder if he could expand (by example perhaps) on his comment, in Section 4.2, that lack of invariance may or may not be seen as a flaw in the model. In particular, with survival data such as those mentioned above, would Dr McCullagh worry about lack of invariance in the proportional hazard model if it appeared to fit the data?

Dr GARY G. KOCH (University of North Carolina, Chapel Hill): The methodology in this paper represents a very useful strategy for the analysis of a broad class of ordinal data. As the author has indicated, it provides a general framework for expressing a set of response distributions in terms of two types of parameters. One of these (the $\{\theta_j\}$) pertains to common aspects of their form while the other (the $\boldsymbol{\beta}$) involves measures of variation among them. These models account for ordinality by being directed at the cumulative probabilities $\gamma_j(\mathbf{x})$ for which constraints are specified; e.g. the proportional odds model requires

$$\log\{\kappa_j(\mathbf{x}_1)\kappa_{j'}(\mathbf{x}_2)/\kappa_{j'}(\mathbf{x}_1)\kappa_j(\mathbf{x}_2)\} = 0 \quad \text{for all } j,\ j' \quad \text{and} \quad \mathbf{x}_1, \mathbf{x}_2.$$

Under these conditions, sub-population differences can be described by the common value of $\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2)$ for the odds ratio at all outcomes j.

If the cell frequencies n_{ij} are large, then the parameters for the proportional odds, proportional hazards and other related models can be estimated by either maximum likelihood (ML) or weighted least squares (WLS). For Table 1, the WLS estimates (without iteration) are $\hat{\Delta} = 0{\cdot}602 \pm 0{\cdot}224$, $\hat{\theta}_1 = -0{\cdot}809 \pm 0{\cdot}116$ and $\hat{\theta}_2 = 1{\cdot}061 \pm 0{\cdot}118$, which are similar to those obtained by the author via ML (with iteration).

More generally, WLS methods have the advantage of more direct applicability to large samples where the likelihood function is not convenient. These include ratio estimates from sample surveys and correlated marginal distributions from repeated measurements studies (see Koch *et al.*, 1977).

Alternatively, for single responses with product multinomial likelihoods, ML would be preferable to WLS when many of the n_{ij} are small because the asymptotic basis of WLS is directed at the $\{n_{ij}\}$. However, the usage of ML here may require caution because the computations need to yield $\hat{\theta}_j < \hat{\theta}_{j'}$ for all $j < j'$ so that $\hat{\gamma}_{j'}(\mathbf{x}) - \hat{\gamma}_j(\mathbf{x}) > 0$. Although negative values for $\{\hat{\gamma}_{j'}(\mathbf{x}) - \hat{\gamma}_j(\mathbf{x})\}$ can be avoided with the long-linear model, non-monotonicity of the $\{\hat{\theta}_j\}$ may require attention for interpretation (see Andrich, 1979).

Thus, when the n_{ij} are small, mean scores based on successive integers or ranks may still be a reasonable framework for the analysis of ordinal data. This approach has the additional advantage of not involving model constraint assumptions. It also accounts for ordinality through weighted sums of the $\gamma_j(\mathbf{x})$ with integer scores corresponding to equal weights. As a result, the valid use of such mean scores does not strictly require any underlying scaling assumptions.

In summary, the methods discussed in this paper are definitely of interest for ordinal data; but they have some limitations which imply that other approaches or methods may be more appropriate for some situations.

Dr B. J. T. MORGAN (University of Kent): In "signal-detection" theory mentioned by Dr Altham the cut points, $\{\theta_j\}$ are sometimes of particular interest to the psychologist. An example is provided by Craig (1979), who describes how performance measures on a binary discrimination task were taken on subjects, in the morning and in the evening. It was found that whereas efficiency did not alter significantly between times, report confidence, as measured, effectively, by the $\{\hat{\theta}_j\}$, did. This change in report confidence was then shown to be significantly related to the subjects' increase in body temperature between morning and evening.

Professor R. L. PLACKETT (University of Newcastle upon Tyne): This is an interesting and useful paper which should become essential reading for all concerned with the analysis of ordinal data. Dr McCullagh has devoted his sole attention to the combination of one ordinal response with several factors. Here are a few comments on the case of two ordinal responses.

Following Yates (1948), the use of scores for the analysis of association in contingency tables was studied by Williams (1952) in terms of the model

$$p_{ij} = p_{i0}\, p_{0j}(1 + \theta c_i d_j),$$

where p_{i0}, p_{0j} are marginal probabilities and c_i, d_j are given. The proposal of Mantel (1963) to use conditional distributions in this context was brought within the log linear framework by Birch (1965). Suppose that $p_{ij} > 0$ for all i, j and define

$$\lambda_{ab} = \log(p_{ab}\, p_{rs}/p_{as}\, p_{rb}) \quad (a = 1, 2, ..., r-1;\ b = 1, 2, ..., s-1).$$

Birch introduced the model

$$\lambda_{ab} = \beta(c_a - c_r)(d_a - d_s),$$

which implies that the log cross-product ratio corresponding to a pair of successive rows and a pair of successive columns has the form $\beta(c_a - c_{a+1})(d_b - d_{b+1})$.

An isotonic formulation of the problem of testing independence is as follows. Most of the information about $\{\lambda_{ab}\}$ is provided by the distribution of cell frequencies conditional on their marginal totals. We require to test the hypothesis

$$\lambda_{ab} = 0 \quad \text{for all } a, b$$

against the alternative that λ_{ab} is non-decreasing in a and non-decreasing in b.

Dr DARYL PREGIBON (Princeton University): My congratulations to Dr McCullagh for a very useful, timely and well-written paper. I especially applaud the more exploratory data analysis techniques employed in several of the examples. My detailed comments are as follows.

(1) I mildly object to the reference to these models as multivariate generalized linear models (g.l.m.'s). The justification is given by (6.3) but g.l.m.'s specify linear relations via link (ϕ_j) rather than by link (γ_j) as in (4.1).

(2) An alternative class of models for the regression analysis of ordinal data is available. These models have link $(p_j^*) = \theta_j + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ where $p_j^* = \mathrm{pr}\{Y = j \mid Y > j-1\}$. Bartlett (1978) describes such an analysis using the complementary log-log (abbreviated in what follows as $c\log^2$) link, which is essentially equivalent to Dr McCullagh's $c\log^2$ (γ_j) model. Thompson (1977) utilizes the logit (p_j^*) link. The most useful point about these models is that one can exploit the conditional independence and use a standard computing package (say GLIM) to perform the analysis.

(3) I question whether the deviance of 0·302 on one degree of freedom indicates a good fit. Clearly it is not significantly large in comparison with $\chi^2(1)$, but what can one expect when fitting three constants to four observations? A fit of the $c\log^2$ (p_j^*) model to the tonsil data leads to a deviance of 0·0624. The logit (p_j^*) model leads to a deviance of 0·0056. Clearly, we cannot depend much on absolute deviances to guide us in judging the quality of fit, or where to stop.

(4) Some concern is raised about link function selection. Simple methods are available for testing link function adequacy within a specified family. Essentially, one (or more) degree(s) of freedom can be

extracted from the fit of the assumed link which is (are) locally sensitive to departures from that link, within the family. For details, see Pregibon (1980).

(5) I am disappointed that no real regression example was considered. I have data on the postoperative status of 165 patients who underwent open heart surgery, the response being one of: no complications, ischaemia, death due to an infarct. The response is ordinal and associated with each patient is a host of covariates; what I call a real ordinal regression problem. In such cases, (i) I doubt whether the generalized empirical transform would prove useful, and (ii) I strongly believe that the deviance contributions are better candidates for residuals than the individual cell differences. I have fitted the logit (p_j^*) model to these data and am now trying to understand exactly what I have—besides a deviance of 251·2 on 300 d.f. A normal probability plot based on the conditionally independent deviance contributions appears as Fig. D3. No observations seem extreme, and I am not sure that I should be upset by the hole in the centre. I would appreciate any comments the author could offer in such cases.

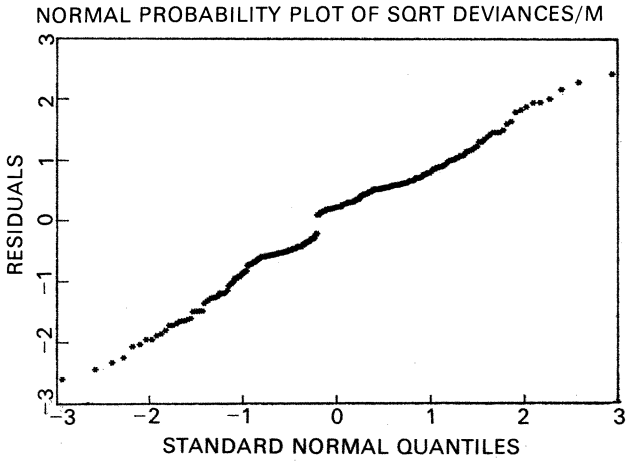

FIG. D3.

Dr J. WAHRENDORF (Krebsforschungszentrum, Heidelberg): I would like to congratulate the author on his very interesting and stimulating paper. His proportional hazards model nicely illustrates that statistical modelling can be very similar in different frameworks. This should remind us of the pragmatism which is always inherent in statistical modelling.

In Section 7.2. the author clearly points out the difference between symmetric and asymmetric models which, I think, is essential and should be more seriously taken into consideration in applications. However, one has to admit that it is not always easy to see whether the variables can be distinguished in this respect and, if so, which can be assumed to be the dependent one; especially since analyses in both directions often give similar results, as can be seen when comparing the analysis of the data from Table 1 of the present paper with the analysis given by Armitage (1955).

In the case of symmetric models, ordinality in both factors can be treated through an approach (Wahrendorf, 1980) which utilizes a one-parameter class of bivariate distributions given by Plackett (1965).

An essential feature of the parametric nature of the methods presented by the author can be seen by comparing the estimated parameters in different independent samples. Let us consider a case-control study where the differences in age distributions between exposed and non-exposed individuals is to be examined for cases and controls. The terms cases, controls, exposure and age may be regarded as substitutes for any other reasonable variables. Separate estimates of, for example, $\bar{\Delta}$ can be calculated for cases and controls. Each of them describes the difference in age distribution between exposed and non-exposed individuals. Comparison of these two parameters on the basis of a normal approximation then allows the analysis of the similarity of these differences for cases and controls. It seems to me that this can open a wide range of applications in case-control studies.

The AUTHOR replied later, in writing, as follows.

I would like to thank the discussants for their useful and encouraging comments. Many points were

raised, only a few of which are dealt with in this reply. Where the same or similar issues were discussed by several contributors these are treated together.

Professor Bartholomew suggests using other and more general link functions possibly incorporating one or more unknown parameters. The same point is covered by Dr Smith, Mr Oranda-Ordaz, Dr Harris and Dr Pregibon. Professor Bartholomew's general class of link functions could be used in a semi-automatic way to let the data choose the appropriate form of link function. Other considerations, however, such as the ease with which quantitative conclusions can be stated and understood, will, in general, strongly favour the extreme cases, i.e. the logistic or complementary log-log link function.

Professor Fienberg's query regarding the existence and uniqueness of maximum likelihood estimates for these models is partially answered in Section 6.3 which deals with large samples only. In the case of linear models, a more complete answer is given by Mr Burridge and Professor Haberman. As Professor Haberman points out, these results do not apply to the non-linear models: in particular, the condition that all cells have positive entries does not guarantee finite estimates in the non-linear model. For these models it is generally the case that infinite parameter estimates indicate either that the data are too sparse or that the model is inappropriate.

Mr Plewis, Professor Fienberg and Dr Pregibon suggest a model for ordinal data based on "continuation ratios". Those familiar with the analysis of survival data will recognize this as the discrete-time proportional hazards model (Cox, 1972, Section 6). This is not a discretized version of the continuous time proportional hazards model so that the complementary log-log model and the analogous continuation ratio model are not the same. Furthermore, the continuation ratio model is not log-linear nor is it of the type (4.1) but, at least in many simple cases, it does satisfy the important property of stochastic ordering (4.2). It is therefore a useful alternative to the models discussed in this paper and is particularly suited to the case where the response categories really are discrete, cannot sensibly be grouped and cannot be thought of as coarse groupings of some finer scale.

Professor Agresti rightly points out that when scores are used in a log-linear model to reflect ordinality, the scores should be altered to take account of subsequent grouping. The trouble is that when the definition of the categories is rather arbitrary, as in most of the examples in this paper, it is impossible to determine the extent of any grouping. One does not know what the appropriate scores are initially. Some justification for preferring ridit scores to integers is given in Section 7.1. Ridit scores also have better invariance properties under grouping of adjacent categories but, regardless of the scores used, concise statement of conclusions is easier with the models (2.3) or (3.5).

Professor Fienberg's assertion that, in certain cases, log-linear models are not permutation invariant does not, I think, contradict an apparent claim to the contrary in Section 4.2. We have, in this paper, restricted attention to a single ordinal response with possibly several explanatory covariates. This precludes triangular structures and bivariate responses. Goodman's (1979a) models, for example, are for bivariate responses with no covariates. When scores are used the log-linear model ceases to be permutation invariant unless, of course, the scores are permuted in the same way as the response categories.

Dr Farewell raises the important question of retrospective sampling. This problem is dealt with in an unpublished paper by Anderson and Philips (1980). The results are algebraically and numerically more complicated than in the case of the logistic model for binary data.

Dr Anderson's discrimination problem raises several important issues. Clearly, the rule which maximizes the posterior probability of correct assignment is inappropriate in most cases. The alternative allocation rule seems preferable but perhaps it would be better for the statistician not to allocate at all but simply to specify the k posterior probabilities or odds for the k categories. If a decision or allocation is absolutely necessary the user, not the statistician, can choose the category with maximum posterior probability or, alternatively, the median category which corresponds to Anderson's alternative allocation rule. This procedure seems satisfactory on general grounds and is in keeping with Bayesian principles.

Dr Anderson, Dr Atkinson and Professor Fienberg discuss the problem of sparse data which usually leads to estimates on the boundary of the parameter space. This is a difficult problem that occurs with sparse data regardless of the model fitted and underlines the need for small sample theory. If the structure of the model is sufficiently simple exact significance tests are available for certain null hypotheses as described in Section 4.3. Further work is clearly required in this area.

Graphical methods, often very helpful for presenting conclusions, are discussed by Dr Altham and Professor Aitkin. In the two sample problem, strict adherence to the straight line principle suggests that the empirical transforms—logistic or complementary log–log—of the cumulative proportions be plotted against each other. If a straight line is obtained, both the slope and the intercept are functions of the parameters β, τ in (6.1). In the case of the income data, there is a real underlying continuous variable,

namely dollars or any monotone transformation thereof. The failure of the logistic model does not imply, *pace* Professor Aitkin, that "the income distributions are not logistic with the same scale parameter but different location parameters". Instead it implies that there is no monotone transformation of the dollar scale that makes the two groups simultaneously logistic in shape and differing only in location. Neither the proportional odds model nor the proportional hazards model makes any reference to the dollar scale. It follows therefore that a failure of the Weibull model does not imply a failure of the complementary log-log model. On the other hand, the failure of the complementary log-log model with location and scale parameters as appropriate implies the failure of the Weibull model. Finally, it should be emphasized that, by refusing to recognize the importance of the dollar scale over arbitrary transformations, we greatly restrict the class of possible conclusions or inferences that might be drawn from these data. For many purposes, depending on the nature of the conclusions to be drawn, it is imperative to make explicit use of the dollar scale, other non-linear scales being irrelevant.

Dr Andrich points out an apparent anomaly in the data of Table 6. If the response is conditional on being in one of the middle categories there is no relation between the conditional response and the covariate. However this is not an anomaly but is predicted by the model. Suppose, for example we have four response categories, two groups and an odds ratio of 6. The probabilities for the two groups might be as follows:

	Response category			
	I	*II*	*III*	*IV*
Group 1	0·20	0·10	0·10	0·60
Group 2	0·60	0·12	0·08	0·20

If the two groups represent the extreme values of some covariate then the probabilities for intermediate values of the covariate will lie between those of groups 1 and 2. The proportions in the extreme categories change by a factor of 3 in each case while those in the middle categories change by only 20 per cent. This kind of variation is seen in Table 6: it is predicted by the model and is not anomalous.

Many other issues were raised that cannot be discussed at length in this reply. A few brief comments are in order. Professor Plackett's suggestion for bivariate responses seems similar to Goodman's (1979) work. Wahrendorf (1980) uses a different approach: neither method can readily take account of covariates. I agree with Dr Altham that further work is required to determine the approximate distribution of the generalized residuals. Dr Eplett's work on robustness seems to rely on there being equal proportionate contamination in each group. It is not immediately clear how these calculations could or should be altered to permit unequal proportionate contamination. Finally, we hope that at least the linear versions of these models will soon become a standard feature of general purpose statistical computing packages.

References in the Discussion

Aitchison, J. and Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika*, **57**, 253–262.

Altham, P. M. E. (1973). A non-parametric measure of signal discriminability. *Brit. J. Math. Statist. Psychol.*, **26**, 1–12.

Anderson, J. A. (1974). Diagnosis by logistic discriminant function: further practical problems and results. *Appl. Statist.*, **23**, 397–404.

Bartlett, N. R. (1978). A survival model for a wood preservative trial. *Biometrics*, **34**, 673–679.

Birch, M. W. (1965). The detection of partial association, II: the general case. *J. R. Statist. Soc.* B, **27**, 111–124.

Burridge, J. (1980). A note on maximum likelihood estimation for regression models using grouped data. *J. R. Statist. Soc.* B, **42**, in press.

Craig, A. (1979). Discrimination, temperature and time of day. *Human Factors*, **21**, 61–68.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27–32.

Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data*, 2nd ed. Cambridge, Mass.: M.I.T. Press.

Fienberg, S. E. and Mason, W. (1978). Identification and estimation of age, period and cohort models in the analysis of discrete archival data. *Sociological Methodology*, 1979, 1–67.

Fisk, P. R. (1961). Estimation of location and scale parameters in a truncated grouped sech square distribution. *J. Amer. Statist. Ass.*, **56**, 692–702.

Goodman, L. A. (1972). Some multiplicative models for the analysis of cross-classified data. *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, **1**, 649–696.

—— (1979a). Multiplicative models for square contingency tables with ordered categories. *Biometrika*, **66**, 413–418.

GREY, D. R. and MORGAN, B. J. T. (1972). Some aspects of ROC curve fitting: normal and logistic models. *J. Math. Psychol.*, **9**, 128–139.
HABERMAN, S. J. (1974). *Analysis of Frequency Data.* Chicago: University of Chicago Press.
HEIM, A. (1970). *Intelligence and Personality.* Harmondsworth: Penguin.
KOCH, G. G. *et al.* (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133–158.
MANTEL, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Ass.*, **58**, 690–700.
PREGIBON, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.*, **29**, 115–123, then p. 114.
PRENTICE, R. L. and BRESLOW, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, **65**, 153–158.
PRENTICE, R. L. and GLOECKLER, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, **34**, 57–67.
SALEM, A. B. Z. and MOUNT, T. D. (1974). A convenient descriptive model of income distribution: the gamma density. *Econometrica*, **42**, 1115–1127.
SINGH, S. K. and MADDALA, G. S. (1976). A function for size distribution of incomes. *Econometrica*, **44**, 963–970.
THOMAS, E. C. and MYERS, J. L. (1972). Implications of latency data for threshold and non-threshold models of signal detection. *J. Math. Psychol.*, **9**, 253–285.
THOMPSON, W. A., JR (1977). On the treatment of grouped observations in life-tables. *Biometrics*, **33**, 463–470.
WAHRENDORF, J. (1980). Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. *Biometrika*, **67**, 15–21.