Regression and Ordered Categorical Variables

Author(s): J. A. Anderson

Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, 1984, Vol. 46, No. 1 (1984), pp. 1-30

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/2345457

J. R. Statist. Soc. B (1984),
46, *No.* 1, *pp.* 1–30

Regression and Ordered Categorical Variables

By J. A. ANDERSON

University of Newcastle upon Tyne

[*Read before the* Royal Statistical Society, *by Professor* R. L. Plackett *on behalf of the late Professor Anderson, at a meeting organized by the* Research Section *on Wednesday, October 5th, 1983*, Professor J. B. Copas *in the Chair*]

SUMMARY

A general approach to regression modelling for ordered categorical response variables, y, is given, which is equally applicable to ordered and unordered y. The regressor variables $\mathbf{x}^T = (x_1, \ldots, x_p)$ may be continuous or categorical. The method is based on the logistic family which contains a hierarchy of regression models, ranging from ordered to unordered models. Ordered properties of the former, the *stereotype* model, are established. The choice between models is made empirically on the basis of model fit. This is particularly important for *assessed*, ordered categorical response variables, where it is not obvious *a priori* whether or not the ordering is relevant to the regression relationship. Model simplification is investigated in terms of whether or not the response categories are distinguishable with respect to $\mathbf{x}$. The models are fitted iteratively using the method of maximum likelihood. Examples are given.

Keywords: ORDERED CATEGORICAL RESPONSE VARIABLE; ASSESSED VARIABLES; LOGISTIC REGRESSION MODELS; STEREOTYPE REGRESSION MODEL; ORDERED PROPERTIES; MAXIMUM LIKELIHOOD ESTIMATION; EXAMPLES

1. INTRODUCTION

Models for the analysis of qualitative, categorical variables are now well established with monographs by Cox (1970), Fienberg (1980) and Plackett (1981). We are concerned here with regression models linking a categorical response variable, y, with a set of predictor or explanatory variables $\mathbf{x}^T = (x_1, \ldots, x_p)$, which may be continuous or discrete. The k categories of y are denoted by $y_1, \ldots, y_k$. Particular emphasis is placed on the case where y is an *ordered* categorical variable and the category with $y = y_i$ is taken to be "lower" than the category with $y = y_j$ if $i < j$. In Section 3, a new approach is presented which unifies regression modelling for *all* categorical response variables, whether ordered or not. This is based on the logistic family which contains a hierarchy of models, ranging from ordered to un-ordered (or nominal) with intermediates. The choice between the models is made on empirical grounds, dependent on model fit. This is particularly important for *assessed* or *judged* ordered response variables, since it is not always clear *a priori* whether the ordering is relevant to the regression relationship.

Several ways have been suggested to allow for an ordered response variable in regression modelling. Any such model will be called an ordered regression model. Existing methods are briefly reviewed in Section 2 and several shortcomings identified. A new ordered model, the *stereotype* model, is introduced in Section 3 as part of the logistic family.

Other objectives of the new approach are to provide a method: (i) suitable for multiple regression, when $p > 1$; (ii) suitable for sparse data; (iii) which allows a parsimonious approach to modelling in the interests of simplifying the description of a relationship. The latter is particularly appropriate for sparse data.

We are mostly concerned with the standard regression situation where $\mathbf{x}$ is fixed and y is the observed random response. However, the methods of Section 3 are equally applicable when

0035–9246/84/46001 $2.00

sampling the joint distribution $(y, \mathbf{x})$ or the conditional distributions $\mathbf{x} \mid y_s, s = 1, \ldots, k$. Hence, depending on the context, either y or $\mathbf{x}$ or both are random variables. Maximum likelihood inference is discussed in Section 4 for all these sampling plans and examples are given in Section 5.

We argue here that the class of regression models currently available for ordered categorical response variables is not wide enough to cover the range of problems that arise in practice. Factors affecting the kind of regression model required are (i) the type of ordered categorical variable, (ii) the observer error process and (iii) the "dimensionality" of the regression relationship. These factors relate to the processes giving rise to the observations and have been rather neglected in the literature. They are now discussed in turn.

There are perhaps two major types of observed ordered categorical variables. A variable of the first type is directly related to a single, underlying continuous variable. For example, McCullagh (1980) refers to "income in dollars": 0-2000, 2001-3000, and so on. Such variables could be termed "grouped continuous". A variable of the second type is generated by an *assessor* who processes an indeterminate amount of information before providing his *judgement* of the grade of the ordered variable. For example, Anderson and Philips (1981) refer to the "extent of pain relief after treatment": *worse*, *same*, *slight improvement*, *moderate improvement*, *marked improvement* or *complete relief*. In principle, there is a single, unobservable, continuous variable related to this ordered scale, but in practice, the doctor making the assessment will use several pieces of information in making his judgement on the *observed* category. For example, he might use severity of pain, kind of pain, consistency in time and degree of disability. We will refer to variables of the second type as "assessed" ordered categorical variables and argue that, in general, a different approach to modelling regression relationships is appropriate for the two types. Assessed, ordered variables occur frequently in the biomedical, social and other sciences.

Differences in the pattern of observer error can be expected over a range of ordered, categorical variables. The errors will probably be greater for assessed ordered variables than for grouped continuous variables. In both cases, uniform error structures across all categories will be rare.

The dimensionality of the regression relationship between y and $\mathbf{x}$ is determined by the number of linear functions required to describe the relationship. If only one linear function is required, the relationship is one-dimensional; otherwise it is multi-dimensional. For example, in predicting k categories of pain relief from predictors $\mathbf{x}$, suppose that different functions $\boldsymbol{\beta}_1^T\mathbf{x}$ and $\boldsymbol{\beta}_2^T\mathbf{x}$ are required to distinguish between the pairs of categories (*worse*, *same*) and (*same*, *better*), respectively. Then the relationship is neither one-dimensional nor ordered with respect to $\mathbf{x}$.

An important concept related to dimensionality is *indistinguishability*. A pair of categories are indistinguishable with respect to $\mathbf{x}$ if $\mathbf{x}$ is not predictive between the two categories. In the above example, suppose that $\mathbf{x}$ is not predictive between *worse* and *same*, but $\mathbf{x}$ is predictive between (*worse* or *same*) and *better*. In this case, *worse* and *same* are *indistinguishable* with respect to $\mathbf{x}$. In the above notation, this would imply that $\boldsymbol{\beta}_1 = \mathbf{0}$. Indistinguishability is an important concept in practice. Possible causes are: (i) intrinsic lack of distinguishability with respect to $\mathbf{x}$; (ii) high observer error for the appropriate categories. Recognition of indistinguishability simplifies the description of complex relationships. Note that the dimensionality of a relationship is usually reduced if some categories are indistinguishable. It is sometimes inadvisable to combine categories indistinguishable with respect to $\mathbf{x}$; it is often sufficient to appreciate their similarity.

McCullagh (1980) argues that the only effect of combining adjoining categories in ordered categorical regression problems should be a loss of efficiency in the estimation of the regression parameters. He uses this almost as an axiom for the acceptability of ordered regression models. However, in some contexts, particularly with assessed ordered variables, this axiom is less attractive. For example, it is well known in questionnaire design that a change in the form of words of a question can lead to an unpredictable change in the pattern of observed responses. Hence, it is not obvious that the change in the form of a question corresponding to the amalgamation of two categories would simply result in the amalgamation of responses in the two categories. Hence, we question the universal application of the combination axiom for judging models, particularly for assessed ordered categorical variables. Related points are made in the discussion of McCullagh's (1980) paper: for example, see Fienberg's contribution. The logistic models introduced in Section 3 do not satisfy the combinability axiom. However, in practice, our standpoint is not so different from McCullagh's since the effect of combining indistinguishable categories in the logistic, and other, methods is negligible.

Although we are given a response variable, y, which is putatively ordered, we have seen that there is no guarantee that an ordered regression model is appropriate. A choice should be made, usually on empirical grounds between models of varying dimensionality and only if the model is one-dimensional should an ordered model be considered. There is no merit in fitting an ordered relationship as a routine, simply because the response variable is ordered. The logistic regression family of Section 3 provides the framework for this choice.

2. EXISTING METHODS FOR ORDERED CATEGORICAL VARIABLES

Plackett (1981, pp.61 *et seq.*) reviews methods for analysing relationships between ordered categorical variables and others. He deals with both regression and correlation models. We are concerned here with regression models and possibly the most popular method to date is the "grouped continuous" approach (Ashford, 1959; McCullagh, 1980). We show below that this method is more appropriate for fitting one-dimensional regression models to grouped-continuous ordered variables. In common with other methods currently available, it is not really flexible enough for the range of problems outlined in the Introduction.

A similar approach to the one advocated here, in Section 3, is the log-linear model for contingency tables when modified for use with ordered variables. Plackett (1981, pp.75 *et seq.*) reviews this material. A simple approach for two-way tables is to assign subjectively selected scores $(c_1 < c_2 < \ldots < c_k)$ to each category of the ordered variable. Then in Plackett's notation, the two-factor, log-linear term is written

$$\lambda_{ab} = \beta_a(c_b - c_k),$$

where the (β_a) are parameters to be estimated. This implicit regression model can be fitted and tested but the subjectively assigned scores are not in keeping with the exploratory nature of the previous Section. Instead, the (c_s) may be regarded as extra parameters to be estimated, with or without the order restriction. Goodman (1981) describes a method of this kind for two-way tables and Haberman (1981) considers in detail the corresponding significance tests. Fienberg (1982) considers this and related material. Estimation problems can become severe when the model is extended to higher dimensions. This approach, when converted into a conditional probability for y given $\mathbf{x}$ and interpreted properly, can give the stereotype model suggested here in Section 3. In the log-linear framework, the (x_j) are restricted to discrete values whereas in Section 3 they may range over continuous intervals. Furthermore, the estimation problems in Section 3 are not so severe.

We now consider the grouped continuous regression model in some detail.

2.1. *The Grouped Continuous Regression Model*

Following McCullagh's (1980) notation, the grouped continuous regression model gives the conditional probability of y given $\mathbf{x}$ as

$$\text{pr}\,(y \leqslant y_s \mid \mathbf{x}) = F(\theta_s - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \quad s = 1, \ldots, k, \tag{1}$$

where $F(\cdot)$ is any convenient cumulative distribution function. The parameters $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ are unknown regression coefficients and the parameters (θ_s) are also unknown $(s = 1, \ldots, k-1)$ and satisfy $\theta_1 < \theta_2 < \ldots < \theta_{k-1}$. For convenience, we usually define $\theta_0 = -\infty$ and $\theta_k = +\infty$.

The model (1) can be taken as a given model to be verified empirically but it is most appropriate when the ordered categories are related monotonically to an unobservable, continuous variable, z, such that y_s is observed if $\theta_{s-1} < z < \theta_s$, $s = 1, \ldots, k$. The parameters (θ_s) are thus the division points of the latent scale, z. Note that we require the actual process of recording y_s to be related to the unobservable scale, z, and hence y is a grouped continuous ordered variable in the terminology of the Introduction. The distribution of y is linked to $\mathbf{x}$ by postulating that the conditional distribution function of z given $\mathbf{x}$ is $F(z - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and model (1) follows immediately. The same function of $\mathbf{x}$, $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ is involved for all categories y_s, thus model (1) gives a one-dimensional, ordered regression model for a grouped continuous ordered response variable.

Model (1) is widely used and is undoubtedly useful in some contexts. McCullagh (1980) gives a good account of many aspects of its use. Anderson and Philips (1981) consider its use in multiple regression and under varying sampling schemes; they also note some difficulties. Thompson and Baker (1981) demonstrate that the model can be fitted using GLIM. However, the application of model (1) to the wider range of problems outlined in the Introduction does not make best use of the available information. It is difficult to interpret the (θ_s) parameters unless the observed y-variable is directly related to a latent variable. The estimates of the (θ_s) are strongly related to the average proportion in the corresponding categories, as recourse to any specified functional form for $F(\cdot)$ indicates. Hence, the (θ_s) parameters are not informative about the closeness of categories. As noted above, the regression relationship is based on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and is firmly one-dimensional.

There does not appear to be any natural way of extending model (1) to vary the assumptions about the way the data are generated. If we use more than one unobservable, continuous variable z in an attempt to model the generating mechanisms of an assessed ordered variable, the resulting model is either computationally impractical or the same as model (1). If we attempt to model multi-dimensional or indistinguishable regression relationships within the framework of a latent, continuous variable, z, a rather unnatural modification of model (1) ensues with severe inferential difficulties. A simple context to see this is the distinguishability example of the Introduction. We wish to test whether *worse* and *same* are indistinguishable with respect to $\mathbf{x}$, while (*worse* or *same*) and *better* are distinguishable with respect to $\mathbf{x}$. This is the null hypothesis, H_0, which is equivalent to a multi-dimensional regression. The alternative hypothesis, H_1, is that all three categories are distinguishable with respect to $\mathbf{x}$. This is a one-dimensional regression model. Under H_1 with model (1), there are $p + 2$ parameters ($\boldsymbol{\beta}$, θ_1 and θ_2) to estimate. Under H_0, one way of proceeding is to amalgamate *worse* and *same* as *worsame* and to model:

$$\mathrm{pr}\,(\textit{worse} \mid \mathbf{x}) = \tau\, \mathrm{pr}\,(\textit{worsame} \mid \mathbf{x}), \tag{2}$$

$$\mathrm{pr}\,(\textit{same} \mid \mathbf{x}) = (1-\tau)\, \mathrm{pr}\,(\textit{worsame} \mid \mathbf{x}), \tag{3}$$

$$\mathrm{pr}\,(\textit{better} \mid \mathbf{x}) = \mathrm{pr}\,(\textit{better} \mid \mathbf{x}). \tag{4}$$

The parameter $\tau (0 < \tau < 1)$ gives the proportion of *worsame* points from *worse*. A two-state model of type (1) gives the probability of *worsame* or *better* given $\mathbf{x}$ and has $p + 1$ parameters. Together with τ, this gives $p + 2$ parameters again. Hence, the models under H_0 and H_1 are from different families and have the same number of parameters. The choice between H_0 and H_1 is not a straightforward inferential problem (Atkinson, 1970) and simple asymptotic methods are not available. This perhaps emphasizes the unnaturalness of multi-dimensional and indistinguishable regression models within the context of model (1). We conclude that a different family of regression models is required to deal with the regression problems outlined in the Introduction.

3. THE STEREOTYPE REGRESSION MODEL FOR ORDERED CATEGORICAL VARIABLES

3.1. *The Qualitative Logistic Regression Model*

The regression models suitable for the problems given in the Introduction need to be more flexible than the grouped continuous model of the previous Section. The ordering of the categories, or subsets of them, with respect to the regression variables is open to question in some cases. Hence, we start with the logistic regression model suitable for a qualitative, categorical response variable (Cox, 1970; Anderson, 1972). This is

$$\operatorname{pr}(y=y_s \mid \mathbf{x}) = \exp(\beta_{0s}^* + \boldsymbol{\beta}_s^{\mathrm{T}}\mathbf{x}) \Big/ \sum_{t=1}^{k} \exp(\beta_{0t}^* + \boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}) \quad (s=1,\ldots,k), \tag{5}$$

where $\beta_{0k}^* = 0$ and $\boldsymbol{\beta}_k = \mathbf{0}$ are introduced to simplify the notation. The parameter vector $\boldsymbol{\beta}_s^{\mathrm{T}} = (\beta_{s1}, \ldots, \beta_{sk})$ gives the regression coefficients for the odds of $y = y_s$ relative to $y = y_k$ $(s = 1, \ldots, k-1)$ and there are $(k-1)$ other parameters β_{0s}^*, $s = 1, \ldots, k-1$. This model is appropriate in many distributional contexts for both discrete and continuous regression variables, x_j, and gives a fairly robust, general approach to discrete regression modelling (Anderson, 1972, 1983a). Note that interaction, quadratic or other transform terms can be introduced into $\boldsymbol{\beta}_s^{\mathrm{T}}\mathbf{x}$ as appropriate for the complexity of the problem.

There is an equivalence between model (5) and the following partially parametric assumptions about the conditional distributions, $f_s(\mathbf{x})$, of $\mathbf{x}$ given $y = y_s$,

$$f_s(\mathbf{x})/f_k(\mathbf{x}) = \exp(\beta_{0s} + \boldsymbol{\beta}_s^{\mathrm{T}}\mathbf{x}) \quad (s = 1, \ldots, k-1). \tag{6}$$

This follows since if we mix the k distributions $\{f_s(\mathbf{x})\}$ in the proportions $\Pi_1, \ldots, \Pi_k$,

$$\operatorname{pr}(y = y_s \mid \mathbf{x}) = \Pi_s f_s(\mathbf{x}) \Big/ \left\{ \sum_{t=1}^{k} \Pi_t f_t(\mathbf{x}) \right\}.$$

Model (5) follows by substituting from (6) and taking $\beta_{0s}^* = \beta_{0s} + \log(\Pi_s/\Pi_k)$. Anderson (1972, 1983a) and Anderson and Blair (1982) give more details of this and the equivalence between models (5) and (6). The relationship between y and $\mathbf{x}$ may be investigated in terms of either the $\{\operatorname{pr}(y = y_s \mid \mathbf{x})\}$ or the $\{f_s(\mathbf{x})\}$ given by the above models. The inferences drawn are the same (Anderson, 1972, 1983a).

We proceed by imposing structure on the $(\boldsymbol{\beta}_s)$ in (5) or (6) to generate an ordered regression model and models suitable for relationships in one-dimension, two-dimensions and so on. Note that the maximum dimension possible is the rank of the $p \times k$ matrix, $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k)$, which is at most $d = \min(p, k-1)$. For example, if $p = 2$, the maximum dimension for the relationship is two.

3.2. *The Stereotype Ordered Regression Model*

Model (5) often gives a good fit to real data, even when the $(\boldsymbol{\beta}_s)$ are restricted to be parallel. This is particularly true when the categories (y_s) are ordered. Taking the $(\boldsymbol{\beta}_s)$ to be parallel,

$$\boldsymbol{\beta}_s = -\phi_s \boldsymbol{\beta} \quad (s = 1, \ldots, k), \tag{7}$$

$\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ and the (ϕ_s) are now the parameters to estimate. Note that $\phi_k \equiv 0$, since $\boldsymbol{\beta}_k = \mathbf{0}$ and we will take $\phi_1 \equiv 1$, to avoid identifiability difficulties. Substitution from (7) into (5) and (6) gives

$$P_s(\mathbf{x}) = \operatorname{pr}(y = y_s \mid \mathbf{x}) = \exp(\beta_{0s}^* - \phi_s \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \Big/ \sum_{t=1}^{k} \exp(\beta_{0t}^* - \phi_t \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \quad (s = 1, \ldots, k) \tag{8}$$

and

$$f_s(\mathbf{x})/f_k(\mathbf{x}) = \exp(\beta_{0s} - \phi_s \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \quad (s = 1, \ldots, k). \tag{9}$$

Since these models for the relationship between y and $\mathbf{x}$ involve only one function of $\mathbf{x}$, $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, they are one-dimensional by definition. There is a growing body of empirical evidence that the $(\boldsymbol{\beta}_s)$ are often parallel. An example of this is given in Section 5.3.

The next step is to order the $(\boldsymbol{\beta}_s)$ given by (7) to obtain an ordered regression relationship. This is achieved by ordering the (ϕ_s) and for definiteness, we take (ϕ_s) as monotone decreasing:

$$1 = \phi_1 > \phi_2 > \ldots > \phi_k = 0. \qquad (10)$$

Monotone increasing (ϕ_s) are dealt with by changing the sign of $\boldsymbol{\beta}$. The ordered regression model (8) subject to constraints (10) will be termed the *stereotype* model. The intuitive introduction of the one-dimensional and the ordered, stereotype models will be amplified in succeeding sections and examples of their use will be given. Inferential problems are discussed in Section 4.2.

We can argue quite generally in favour of a one-dimensional or ordered regression model linking y and $\mathbf{x}$ if the distributions $\{f_s(\mathbf{x})\}$ have one-dimensional or ordered properties with respect to each other: See Section 3.3 for the multivariate normal case. Equivalently, we can argue from the probabilities $\{\mathrm{pr}\,(y = y_s \mid \mathbf{x})\}$. These orderings are quite different from those imposed in the grouped continuous model. There the ordered categories are "given" and are not necessarily ordered with respect to $\mathbf{x}$. Indeed, in the grouped continuous model, it is not necessary to have any regressor variables (x_j). By contrast, in the stereotype approach, we are only interested in regression relationships between y and $\mathbf{x}$.

In a similar way to the above, a two-dimensional regression model can be introduced. Suppose now that

$$\boldsymbol{\beta}_s = -\phi_s \boldsymbol{\beta} - \psi_s \boldsymbol{\gamma}, \quad s = 1, \ldots, k, \qquad (11)$$

where $\boldsymbol{\gamma}^{\mathrm{T}} = (\gamma_1, \ldots, \gamma_p)$. As before, $\phi_k \equiv \psi_k \equiv 0$. Identifiability considerations place further constraints on the parameters. One viable set of constraints is $(\phi_1 = 1, \psi_1 = 0, \phi_2 = 0, \psi_2 = 1)$, but the context may suggest other choices. When substituted into (5) and (6), this form for $(\boldsymbol{\beta}_s)$ gives a two-dimensional regression model, since the model now depends on two functions of $\mathbf{x}$, $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$. Proceeding in this way, regression models of arbitrary dimension can be built up.

We now have a flexible logistic regression model which can handle multi-dimensional, one-dimensional or ordered regression relationships. The special case of indistinguishability, mentioned in the Introduction, has a particularly simple representation in this logistic framework. For models (5) and (6), the hypothesis that $y = y_s$ is indistinguishable from $y = y_t$, with respect to $\mathbf{x}$, has the parametric form $H_0: \boldsymbol{\beta}_s = \boldsymbol{\beta}_t$. In one-dimensional models, this simplifies to $H_0: \phi_s = \phi_t$ and there are equally simple hypotheses to test in other dimensions. The choice between the models of different levels of complexity is made on empirical grounds. Within the hierarchy of models, we take the simplest that fits the data well. Since all the choices are made within the same family, there is no inferential difficulty of the kind mentioned in the previous section for the grouped continuous model.

We believe that the logistic family of regression models is particularly appropriate for assessed, ordered response variables. This is partly on pragmatic grounds, since the looser definition of such variables usually suggests that questions about multi-dimensionality and distinguishability are relevant. It is also because we conjecture that there may be a link between the way that assessed variables are generated and the logistic approach. In the category assessment procedure an indeterminate amount of information, J, is processed in an unknown way. One possibility is that the judge has loose stereotypes for each category and that a new case for categorization is fitted into the most appropriate category. This is akin to discrimination on the basis of J and suggests different distributions for J depending on the category. This also suggests different distributions for $\mathbf{x}$ in each category, especially if J and $\mathbf{x}$ are related. In this case, $\mathrm{pr}\,(y = y_s \mid \mathbf{x})$ can be modelled as in discrimination to give the equation following (6). The general (unordered) logistic model (5) follows in many contexts as seen in Section 3.1. An ordered model comes about if the distributions of J and $\mathbf{x}$, given the category, are ordered. These ideas of ordered distributions are explored further in Sections 3.3 and 3.5. The term stereotype model is applied only to the ordered model although it could be used for the unordered model (5). The pragmatic argument for the logistic family holds good irrespective of how the assessed, ordered variables are generated. In any case, there must be a good fit to the data.

The logistic approach has a strong appeal from the empirical, data-analytic point of view but it also has useful theoretical properties. We consider two special cases in Sections 3.3 and 3.4 before proceeding to the general case. In all cases, particular attention is focussed on establishing (i) the one-dimensional nature of models (8) and (9) and (ii) the ordered properties introduced by (10).

3.3. *Ordered Multivariate Normal Distributions*

Suppose that the distribution of $\mathbf{x}$ conditional on $y = y_s$ is $N(\boldsymbol{\mu}_s, \Sigma)$, $s = 1, \ldots, k$. Equation (6) is satisfied by the density functions, where $\boldsymbol{\beta}_s^{\mathrm{T}} = (\boldsymbol{\mu}_s - \boldsymbol{\mu}_k)^{\mathrm{T}} \Sigma^{-1}$ and $\beta_{0s} = (\boldsymbol{\mu}_k^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_s^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_s)/2$ for all s. Equation (5) is also satisfied, taking these densities to be mixed as in Section 3.1.

The only differences between the distributions are in their means $(\boldsymbol{\mu}_s)$ and hence it is reasonable to suppose that the relationship between y and $\mathbf{x}$ is determined by the $(\boldsymbol{\mu}_s)$. If these are collinear, then the relationship is one-dimensional and (8) is satisfied. If, in addition, the sequence $(\boldsymbol{\mu}_s)$ is ordered as (s), then the relationship is ordered and (10) is satisfied. This is illustrated in Fig. 1.

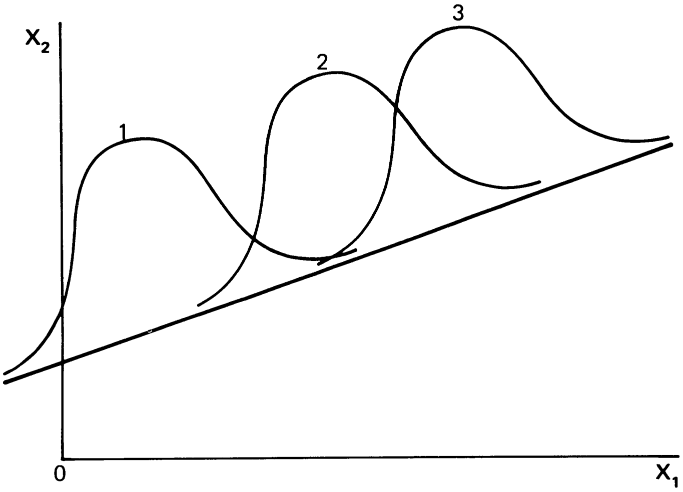

Fig. 1. Collinear bivariate normal distributions: marginal distributions on $z = \boldsymbol{\beta}^{\mathrm{T}}\mathrm{x}$.

The densities are physically displaced along a line, giving the desired one-dimensional and ordered properties.

Similarly, regression relationships between y and $\mathbf{x}$ can be defined in as many dimensions as desired by constraining the $(\boldsymbol{\mu}_s)$, and hence the $(\boldsymbol{\beta}_s)$, to lie in a space of the appropriate dimension.

3.4. *Ordered Binary Distributions*

Suppose there is one binary, regressor variable x_1, taking values x_{11} and x_{12}. The observations could be regarded as falling in a $2 \times k$ contingency table, as the response variable, y, has k possible values. Suppose the joint probabilities in this table are given by $p_{js} = \mathrm{pr}\,(x_1 = x_{1j}, y = y_s)$ for $j = 1, 2$; $s = 1, \ldots, k$, and that the one-dimensional model (8) holds. Then it follows that the model is saturated and

$$\frac{p_{js}}{p_{jk}} = \frac{\mathrm{pr}\,(y = y_s \mid x_{1j})}{\mathrm{pr}\,(y = y_k \mid x_{1j})} = \exp\,(\beta_{0s}^{*} - \phi_s\,\beta_1 x_{1j}) \qquad (j = 1, 2,; s = 1, \ldots, k),$$

or

$$\lambda_s = \log\{p_{1s}\, p_{2k}/(p_{1k}\, p_{2s})\} = \phi_s\, \beta_1 (x_{12} - x_{11}) \quad (s = 1, \ldots, k). \tag{12}$$

Thus, $\phi_s = \lambda_s/\lambda_1$ ($s = 1, \ldots, k$) recalling that $\phi_1 \equiv 1$ and $\phi_k \equiv 0$. These equations can be satisfied by any $2 \times k$ table, confirming that if there is one binary covariate, the regression relationship must be one-dimensional. The interest here is in whether the relationship is ordered.

For an ordered relationship, (10) must hold, that is, $1 > \lambda_2/\lambda_1 > \ldots . > \lambda_{k-1}/\lambda_1 > 0$. Thus, the sequences (λ_s), and hence $\{\exp(\lambda_s)\}$ must be either monotone increasing or monotone decreasing. This last sequence is just $\{(p_{1s}/p_{2s})/(p_{1k}/p_{2k})\}$, so that the sequence (p_{1s}/p_{2s}) is also monotonic increasing or decreasing. Thus, observations in a $2 \times k$ contingency table follow the stereotype regression model if, and only if, the probability ratios or cross-ratios are ordered as above. The "only if" statement follows by reversing the above argument. It follows that two categories, y_s and y_t, are indistinguishable with respect to x_1 if the probability ratios are equal: $p_{1s}/p_{2s} = p_{1t}/p_{2t}$.

An alternative interpretation of the above discussion is that order constraints have been imposed on the interaction terms of the log-linear model for the $2 \times k$ contingency table. Section 2 refers to the equivalence between special log-linear models and the stereotype model (8) for categorical **x**. There are no similar results for continuous **x**.

3.5. *Ordered Properties of* $\{P_s(\mathbf{x})\}$ *for the Stereotype Model*

It was noted in Section 3.2 that model (8) is one-dimensional by definition. We now consider its ordered properties when (10) is also satisfied, giving the stereotype model. It will be convenient to denote the values of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ by z and to write $P_s(\mathbf{x})$ as $P_s(z)$. For simplicity, z will be assumed to take continuous values in $\mathbb{R}^1$ in what follows. Discrete **x** gives discrete z but the modifications required are straightforward and are omitted.

The family of distributions (T_i), where $i \in I$, an index set, is defined to be *stochastically increasing* with respect to I if $\mathrm{pr}(T_j > t) \geqslant \mathrm{pr}(T_i > t)$ for all t, implies and is implied by $j > i$ for all $i, j \in I$. If so, the family $(-T_i)$ is *stochastically decreasing.*

The random variable Y_z is defined by $\mathrm{pr}(Y_z \leqslant y_s \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = F_z(s) = \Sigma_{t=1}^{s} P_t(z)$, where the $\{P_s(z)\}$ are given by the stereotype model, (8) with (10). The family of distributions $\{P_s(z)\}$ or $\{F_z(s)\}$ has many ordered properties including:

Property: P1. The family (Y_z) is stochastically increasing with respect to $Z = \{z: z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$ if (10) holds.

Property: P2. For each s, the functions $\{P_s(z)\}$ are unimodal with modes at (z_s), say, for $s = 2, \ldots, k-1$. Note that $P_1(z)$ and $P_k(z)$ are monotone decreasing and increasing, respectively. The modes are ordered: (i.e. $z_s < z_t$ for all $s < t$) if and only if (10) holds.

Property: P3. If (10) holds:

(i) There exist values (z_{st}^*) such that, for $s < t$,

$$P_s(z) \underset{<}{\geqslant} P_t(z) \quad \text{for } z \underset{\geqslant}{<} z_{st}^*,$$

where

$$z_{st}^* = \frac{\beta_{0s}^* - \beta_{0t}^*}{\phi_s - \phi_t} \text{ for all } s, t; s \neq t. \tag{13}$$

(ii) If the marginal probabilities $\{\mathrm{pr}(y = y_s)\}$ or (Π_s) are equal to $1/k$ for all s, then $z_{st}^* = z_{st}$, where

$$z_{st} = \frac{\beta_{0s} - \beta_{0t}}{\phi_s - \phi_t} \text{ for all } s, t; s \neq t \tag{14}$$

and $z_{st} < z_{s't'}$ if $s < s'$ and $t \leqslant t'$, or if $s \leqslant s'$ and $t < t'$.

Property: P4. One method suggested by Anderson and Philips (1981) for predicting y and $\mathbf{x}$ is to take the category, $\tilde{y}$, which maximizes the posterior probability, pr $(y \mid \mathbf{x})$ or pr $(y \mid z)$ for model (8). This gives the allocation region for $y = y_s$, $A_s = \{z: P_s(z) \geqslant P_t(z), t = 1, \ldots, k; t \neq s\}$, $s = 1, \ldots, k$. The A_s are closed intervals of $\mathbb{R}^1$ and if (10) holds, they are ordered in the sense that if $z_1 \in A_s$ and $z_2 \in A_t$, then $z_1 \leqslant z_2$ implies $s \leqslant t$ but $s < t$ implies $z_1 \leqslant z_2$, with equality possible only at boundary points.

The full proofs of these properties are omitted for conciseness but can be found in Anderson (1982). Outline proofs of Properties P1 and P3 can be found in the Appendix. These give an impression of what is involved. Other properties are also available. For example, Dr P. M. E. Altham showed that the family (Y_z) has a monotone likelihood ratio and that the functions $-\log P_s(z)$ are convex in z for all s. It is hoped that these findings will be reported elsewhere.

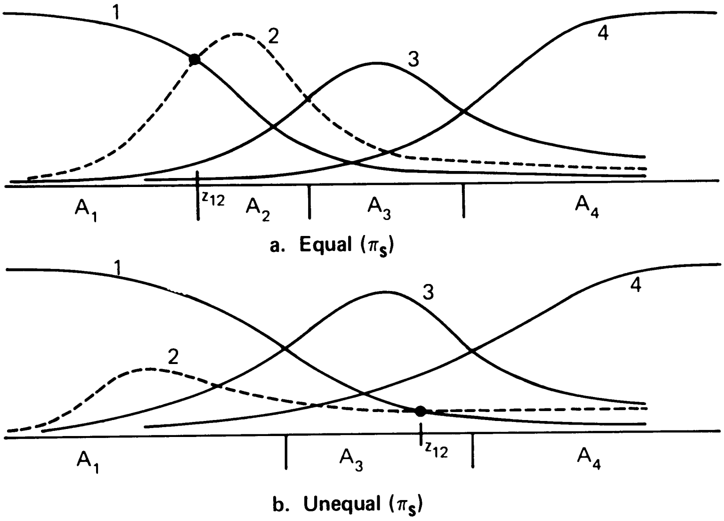

Fig. 2. Ordered properties of $\{P_s(z)\}$:

For $k = 4$, the above properties of the $\{P_s(z)\}$ are illustrated in Fig. 2a for equal $\{\Pi_s\}$ and in Fig. 2b for unequal $\{\Pi_s\}$. In both cases, the functions $P_1(z)$ and $P_4(z)$ are monotone and $P_2(z)$ and $P_3(z)$ are unimodal, with the modes ordered as s. Furthermore, the non-empty allocation regions (A_s) are ordered. The major difference between the two figures is that in (b), Π_2 is small, displacing z_{12} to the right, thus making the content of A_2 zero. It is concluded that the stereotype model possesses desirable and satisfactory ordered properties in respect of the $\{P_s(\mathbf{x})\}$.

3.6. *Ordered Properties of* $\{f_s(\mathbf{x})\}$

We now consider the family of distributions $\{f_s(\mathbf{x})\}$ given by (9) and establish its one-dimensional properties and its ordered properties when (10) is imposed. This family is very general, covering continuous and categorical x_j. It is partially parameterized in the sense used by Anderson (1983a). Although some distributional assumptions have been made, the family is not completely specified. Anderson (1983a) considered inference for the family given by (6) and for other more general families. Firstly, the equations (9) are rewritten as

$$f_s(\mathbf{x}) = \exp(\beta_{0s} - \phi_s \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) f(\mathbf{x}) \quad (s = 1, \ldots, k), \tag{15}$$

where $f(\mathbf{x}) \equiv f_k(\mathbf{x})$. Recall that $\beta_{0k} = \phi_k = 0$. In Section 3.2, it was argued that model (9) or, equivalently, model (15) gives a one-dimensional model by definition.

Ordered properties of the multivariate distributions $\{f_s(\mathbf{x})\}$ similar to those available for the multivariate normal distribution in Section 3.3 are not yet available but useful ordered properties are available for the marginal distributions along the direction $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. These are the same as the properties along x_1 if $p = 1$.

Let $z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $R = \{\mathbf{x}\colon \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = z\}$. If $y = y_s$ and $\mathbf{x}$ is continuous, the probability density function for z is

$$h_s(z) = \exp(\beta_{0s} - \phi_s z) \int_R f(\mathbf{x})\, d\mathbf{x} \tag{16}$$

or

$$h_s(z) = \exp(\beta_{0s} - \phi_s z)\, h(z), \quad s = 1, \ldots, k. \tag{17}$$

Discrete (x_j) or mixtures of continuous and discrete $\mathbf{x}$ are handled similarly. In all cases $h(z)$ is the probability content of R with respect to $f(\mathbf{x})$. Ordered properties of the family $\{h_s(z)\}$ are now considered. These properties hold if the condition (10) on the (ϕ_s) is satisfied. Points z of zero probability mass under all k distributions are excluded from consideration.

Property: F1. For $s < t$, the distributions satisfy

$$h_s(z) \underset{\leqslant}{\overset{>}{}} h_t(z) \quad \text{for } z \underset{\geqslant}{\overset{<}{}} z_{st}, \tag{18}$$

where z_{st} is given by (14).

Property: F2. The points z_{st} are ordered:

$$z_{st} < z_{s't'} \quad \text{if } s \leqslant s' \quad \text{and} \quad t < t' \quad \text{or if } s < s' \quad \text{and} \quad t \leqslant t', \tag{19}$$

Property: F3. The random variables (Z_s) with distributions $\{h_s(z)\}$ are stochastically increasing with respect to s.

Proofs of the properties are again omitted but can be found in Anderson (1982). The bivariate normal example in Fig. 1 shows the physical displacement which gives rise to the ordering. In general, the density functions need not be unimodal.

We have shown that the marginal distributions of $\{f_s(\mathbf{x})\}$, along the direction $z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, have important ordered properties, provided (9) and (10) hold. This completes our case that the model given by (8) or (9) is an ordered regression model if (10) holds.

4. INFERENCE IN THE LOGISTIC FAMILY

In principle, maximum likelihood estimation of the parameters postulated under the various models of Section 3 is reasonably straightforward. Numerical optimization methods are required to maximize the likelihood, but so far, the problems appear to be numerically well behaved. In Section 4.2, we see that there are some difficulties in testing because, in general, some of the required likelihood ratio statistics may not have asymptotic chi-square distributions. Indirect approaches are sometimes needed. However, for one regressor variable, the usual asymptotic chi-square results are available. In fact, the examples in Section 5 demonstrate that these difficulties are often more apparent than real.

4.1. *Maximum Likelihood Estimation*

The various models for $\{\mathrm{pr}(y = y_s \mid \mathbf{x})\}$, (5), (8) or (11), can be fitted using the method of maximum likelihood. Three important study designs are considered: (i) $\mathbf{x}$-*conditional*; (ii) $(y, \mathbf{x})$ *joint* or mixture; and (iii) *y-conditional* or separate. These correspond to sampling the distributions $y \mid \mathbf{x}$, $(y, \mathbf{x})$ and $\mathbf{x} \mid y$, respectively. In all three cases, the maximum likelihood estimates are given by the same iterative algorithm.

The simplest design to consider is the **x**-conditional. This is perhaps the classic regression study and its use is exemplified by a dose-response investigation in which the response, y, is noted at each of a series of dose-levels, **x**. Suppose that $n_s(\mathbf{x})$ sample points are observed at **x** with response $y = y_s$. Then the likelihood is

$$L_c = \prod_{s=1}^{k} \prod_{\mathbf{x}} \{\mathrm{pr}\,(y = y_s \mid \mathbf{x})\}^{n_s(\mathbf{x})}, \tag{20}$$

where the product $\Pi_{\mathbf{x}}$ is taken over observed **x**-values. The appropriate model for the $\{\mathrm{pr}\,(y = y_s \mid \mathbf{x})\}$, (5), (8) or (11), is substituted into (20) and then L_c is maximized with respect to the parameters of the model.

The joint sampling plan is also straightforward. Examples of its use are in prospective epidemiological studies and prognosis. In the above notation, $n_s(\mathbf{x})$ is now the number of individuals noted at $(y = y_s, \mathbf{x})$. The full likelihood is

$$L_J = \prod_{s=1}^{k} \prod_{\mathbf{x}} \{L\,(y = y_s, \mathbf{x})\}^{n_s(\mathbf{x})}. \tag{21}$$

Conditioning on **x** leads to L_c, as given in (20). None of the models (5), (8) or (11) makes any assumptions about the form of the marginal distribution for **x**, $L(\mathbf{x})$. Hence L_c contains all the information about the relevant parameters and the maximum likelihood estimates are obtained by maximizing L_c.

The y-conditional design is referred to by some authors as inverse sampling. It also occurs in epidemiology in case-control studies and in discrimination as separate sampling. In this case, sample points are taken from each $f_s(\mathbf{x})$ separately $(s = 1, \ldots, k)$, with $n_s(\mathbf{x})$ observed at **x**. It is most natural to think of estimating the parameters of the ratios $\{f_s(\mathbf{x})/f_k(\mathbf{x})\}$ as modelled in (6) and (9). Anderson (1982) showed that in a general setting, parameters of the likelihood ratios in y-conditional sampling could be estimated by maximizing the expression L_c in (20) where now $\beta^*_{0s} = \beta_{0s} + \log\,(n_s/n_k)$. For categorical **x**, Anderson (1982) showed that this was a maximum likelihood procedure. For continuous **x**, Prentice and Pyke (1979) showed that this was a maximum likelihood procedure for the model (5). Anderson and Blair (1982) showed that it was also a penalized maximum likelihood procedure. It seems reasonable to conjecture that the above estimation procedure for the specializations (9) and (11) of model (5) are also maximum likelihood procedures for continuous **x**.

The maximization for L_c proceeds in the same way, irrespective of sampling plan. An iterative optimization procedure is required; a good choice is the quasi-Newton algorithm (Gill and Murray 1972) as implemented in the NAG library. The fitting of the multiple logistic model (5) is straightforward numerically and standard programs are available. A program is also available to fit the one-dimensional model (8). Alternatively, Green's (1982) approach to iteratively reweighted least squares could be used. Imposition of order constraints (10) could cause numerical difficulty but, in practice, this has not been necessary; see the examples, Section 5. In principle, there is no difficulty in fitting models of dimensionality $2, 3, \ldots, k-2$ (for example, that given by (11)); the appropriate likelihood function is maximized iteratively. This has not yet been attempted because these models have not often been required. Instead, indirect methods have been used to check that a one-dimensional fit is adequate. See the end of Section 4.2 and the examples.

4.2. *Testing: Choice of Dimensionality, Distinguishability, Ordering*

Given a data set, collected according to one of the above sampling plans, the choice of model within those given by (5) is best dealt with empirically. The response variable is a k-category variable so there is a choice between 1-, 2-, . . . and d-dimensional models, in the terminology of Sections 3.1 and 3.2. Recall that $d = \min(p, k-1)$. This choice can be made on the basis of the corresponding maximized log-likelihoods using asymptotic methods of inference. Suppose the maximized log-likelihood, L_c, under the hypothesis, H_s, of an s-dimensional relationship, is l_s, $s = 0, 1, \ldots, d$. Note that H_0 corresponds to the hypothesis that there is no relationship between y and $\mathbf{x}$, $\boldsymbol{\beta}_s = 0$, $\forall\ s$. The statistics $2(l_s - l_{s-1})$, $s = 1, \ldots, d$, test the nested sequence of hypotheses that the dimensionality of the regression is $1, \ldots, d$, respectively. Unfortunately, the usual arguments leading to asymptotic chi-squared distributions for the above statistics do not hold, in general. For example, in testing $H_0(\boldsymbol{\beta}_s = \mathbf{0})$ against $H_1(\boldsymbol{\beta}_s = -\phi_s \boldsymbol{\beta})$, the parameters (ϕ_s) $(s = 2, \ldots, k-1)$ are not identifiable under the hypothesis that $\boldsymbol{\beta} = \mathbf{0}$. This disallows the usual asymptotic argument. Similar difficulties occur in inference about the existence of break-points in regression lines (Hinkley, 1970) and in many other contexts.

Note that the hypotheses about the collinearity of multivariate normal means $(\boldsymbol{\mu}_s)$ considered by Rao (1973, p.577) can be tested using the above approach. The reduction in power compared with Rao's normal theory tests is compensated by the gain in robustness for non-normal data.

The above difficulties do not apply in some special but important cases. Firstly, we note that $2(l_d - l_0)$ always has an asymptotic chi-squared distribution with $p(k-1)$ degrees of freedom on the usual argument. If the observed value of l_d is significant and close to the observed value of l_1, this may be sufficient to suggest that the model is one-dimensional. If $p = 1$, or $k = 2$, the relationship is at most one-dimensional and $2(l_{k-1} - l_0)$ gives a direct test of this. Other small values of p and/or k give obvious bounds on the dimensionality. Further work is required on the asymptotic distribution of $2(l_s - l_{s-1})$ in the general case but, in the interim, these statistics provide a useful yardstick. An alternative, indirect approach is suggested below.

Having decided on the dimensionality of the model, there are then questions about ordering and model simplification, perhaps using distinguishability as a criterion. One-dimensional relationships are fairly frequent in practice but a common finding is that the estimates, $(\hat{\phi}_s)$, of (ϕ_s) have large standard errors. One possibility is that the categories are not all distinguishable so we now consider testing the corresponding hypotheses.

There is some information about distinguishability in the dimensionality of the regression model selected as above. If the dimensionality is 0 or d, then none or all the categories are distinguishable, respectively. If the dimensionality has any other value some, but not all, of the categories may be indistinguishable.

Taking the one-dimensional case first, models (7), (8) or (9) apply and distinguishability is tested by hypotheses of the form $H_0\colon \phi_s = \phi_t$. Since $\boldsymbol{\beta} \neq \mathbf{0}$, the corresponding likelihood ratio statistics have χ_1^2 distributions. Because of the ordered background, it will usually be sensible to test first the hypotheses $\phi_s = \phi_{s+1}$, $s = 1, \ldots, k-1$. These are simultaneous tests of hypotheses, similar to those occurring in the analysis of variance. Examples are given in the next Section. It is not necessary to have a different numerical algorithm for maximizing L_c under the hypothesis that, say, $\phi_1 = \phi_2$. Combination of categories $y = y_1$ and $y = y_2$ and maximization of L_c give, say, L_{cmb}. Then the required maximized likelihood under $\phi_1 = \phi_2$ is

$$L_{\text{ind}} = \left(\frac{n_1}{n_1 + n_2}\right)^{n_1} \left(\frac{n_2}{n_1 + n_2}\right)^{n_2} L_{\text{cmb}}, \tag{22}$$

where n_1, n_2 are the number of observations in categories 1 and 2, respectively. This result can be extended immediately to tests of more complicated hypotheses such as $\phi_1 = \phi_2 = \phi_3$. Expression (22) follows from noting that, under the hypothesis $\phi_1 = \phi_2$,

$$P_s(\mathbf{x}) = \kappa_s\{P_1(\mathbf{x}) + P_2(\mathbf{x})\}, \quad s = 1, 2,$$

where $\kappa_s = e^{\beta_{0s}{}^*}/(e^{\beta_{01}{}^*} + e^{\beta_{02}{}^*})$, $s = 1, 2$. Maximizing the likelihood, L_c, under the indistinguishability hypothesis leads to maximizing the combined likelihood with $\hat{\kappa}_1 = n_1/(n_1 + n_2)$.

Distinguishability may also be tested in higher dimensions. In the two-dimensional model given by (11), indistinguishability between $y = y_s$ and $y = y_r$ implies that $\phi_s = \phi_t$ and $\psi_s = \psi_t$. Simultaneous inference is again required to establish which categories are distinguishable. Shortcuts, similar to (22), are available for computing maximized likelihoods under the above hypotheses. Similar results obtain in other dimensions.

The above approach to testing distinguishability is best described in the context of examples. See Section 5.1, where further discussion of the testing procedure is given.

An alternative procedure for investigating structure is now possible. Firstly, consider the hypotheses that there are two groups of $(\boldsymbol{\beta}_s)$ in model (5):

$$H_{(2;r)}: \boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_r; \quad \boldsymbol{\beta}_{r+1} = \ldots = \boldsymbol{\beta}_k = 0; \quad r = 1, \ldots, k-1, \tag{23}$$

with corresponding maximized log-likelihoods $l(2;r)$. Recall that the hypothesis H_0, of no relationship, is equivalent to $\boldsymbol{\beta}_s = \mathbf{0}\ \forall\ s$, or that there is one group of $(\boldsymbol{\beta}_s)$. The corresponding log-likelihood is l_0. The likelihood ratio test $2(l(2;r) - l_0)$ for $H_{(2;\,r)}$ against H_0 has an asymptotic χ^2_p-distribution for each $r = 1, \ldots, k-1$. To test the "two groups" hypothesis, $H_{(2)}$, against H_0 at significance level α, take $l(2) = \max_r \{l(2;r)\}$ and test $2(l(2) - l_0)$ as a χ^2_p variate at the $1-(1-\alpha)^{1/(k-2)}$ significance level. This is Tippett's procedure (Koziol and Perlman, 1978) and is only approximately valid since independence between the $(k-1)$ statistics $2(l(2;r) - l_0)$ is assumed.

The above maximizing value of r, r^*, gives the "best" division of the $(\boldsymbol{\beta}_s)$ into two groups. If two groups are accepted, the dimensionality of the regression relationship is at least one.

Secondly, consider the hypothesis that there are three groups of $(\boldsymbol{\beta}_s)$:

$$H_{(3;s,t)}: \boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_s = \boldsymbol{\beta}; \quad \boldsymbol{\beta}_{s+1} = \ldots = \boldsymbol{\beta}_t = \gamma; \quad \boldsymbol{\beta}_{t+1} = \ldots = \boldsymbol{\beta}_k = 0$$
$$(s, t = 1, \ldots, k-1; s < t), \tag{24}$$

with corresponding maximized log-likelihoods $l(3;s,t)$. To maintain a hierarchical structure of hypotheses, one of s or t must be equal to r^*. With this restriction, Tippett's procedure can be used to test at level α three groups of $(\boldsymbol{\beta}_s)$ against two groups using as statistic $2(l(3) - l(2))$, where $l(3) = \max_{s,t}\{l(3;s,t)\}$. This statistic is tested as a χ^2_p variate at the $1-(1-\alpha)^{1/(k-2)}$ level of significance. Suppose that (s^*, t^*) are the maximizing values of (s,t), then these give the best division of the $(\boldsymbol{\beta}_s)$ into three groups. If three groups are accepted, then we may wish to test for four, five or s groups of indistinguishable categories with corresponding likelihoods, $l(4), l(5)$ or $l(\mathrm{s})$ defined similarly to $l(2)$ and $l(3)$. Note that $l(k)$ corresponds to the hypothesis that all k $(\boldsymbol{\beta}_s)$ in model (5) are different and hence $2(l(k) - l_0)$ can be tested as a $\chi^2_{p(k-1)}$ variate. These procedures can be extended to test similar hypotheses about distinguishability of categories when the model is restricted to be one-dimensional (model (8)), two-dimensional (see (11)) or any specified dimensionality. Section 5.3 gives an example of this for a one-dimensional model.

With the questions about dimensionality and distinguishability settled, we have a regression model which is as economical as possible in the number of parameters required. If the dimensionality is one, there is the further question about ordering. The final model in one-dimension has $(\hat{\phi}_s)$ as estimates of (ϕ_s), where some subsets of the $(\hat{\phi}_s)$ may be equal, indicating indistinguishability. Then the estimated regression model is ordered if

$$1 = \hat{\phi}_1 \geqslant \hat{\phi}_2 \geqslant \ldots \geqslant \hat{\phi}_k = 0. \tag{25}$$

Note that this is not a strict ordering as adjoining categories may be indistinguishable. If (25) is not satisfied, then the ordering of the regression model is not that suggested by the category definitions.

The above procedure for model selection has side-stepped the problem of directly testing whether or not a given one-dimensional regression model (8) is ordered. This is equivalent to the hypothesis that the strict inequality (10) on the (ϕ_s) holds for most practical purposes. This is virtually the same as the hypothesis that:

$$1 = \phi_1 \geqslant \phi_2 \geqslant \ldots \geqslant \phi_k = 0. \tag{26}$$

A likelihood test for ordering is given by computing the maximized likelihood, l_{ORD}, under (10) or (26) and considering the statistic $2(l_1 - l_{\text{ORD}})$. If $l_{\text{ORD}} = l_1$, then the constrained and unconstrained maxima are equal and we accept the ordered hypothesis. If not, there is the difficulty that the test statistic may not be asymptotically distributed as chi-squared so that inference is difficult. An alternative approach would be to consider the asymptotic distribution of $\hat{\phi}_2, \ldots, \hat{\phi}_{k-1}$ as derived from the one-dimensional fit of model (8) and to test from that

$$H_{\text{ORD}}: \phi_1 - \phi_2 \geqslant 0, \phi_2 - \phi_3 \geqslant 0, \ldots, \phi_{k-2} - \phi_{k-1} \geqslant 0. \tag{27}$$

Further work is required on testing ordered hypotheses but the first approach introduced above will be adequate for many purposes. In practice, there will often be only two groups of $(\boldsymbol{\beta}_s)$ which ensures ordering.

4.3. *Stepwise Choice of Regressor Variables*

The above material refers to the choice of type of model but, in some practical situations, a large number of regressor variables have been observed and we need a method of selecting subsets which adequately describe the relationship. Stepwise methods are often used for this purpose and, *given* the type of model selected as in Section 4.2 above, standard selection procedures can be adapted. However, the question then arises of whether the forms of models selected in Section 4.2 are invariant under various choices for the subsets of the regressor variables. Some iteration between choice of model and choice of regressor variables will help to determine how stable these choices are.

For moderate numbers of regressor variables, these difficulties do not arise and there may be no need for regressor variable selection.

5. EXAMPLES

5.1. *Nausea Data*

Farewell (1982) reports a study of patients undergoing chemotherapy, who were categorized by (i), y, the amount of nausea experienced on 6-point scale ranging from none to severe and (ii), whether or not their therapy included Cisplatinum. Table 1 gives the data. Nausea, as judged by this 6-point scale, is an assessed, ordered categorical variable as defined in the Introduction. Hence it is natural to try fitting the logistic family of regression models (5) to the data. Note that there is one binary, regressor variable, x_1, which takes values 0 and 1 for no Cisplatinum and Cisplatinum, respectively. Hence, the dimensionality of the regression relationship is at most 1. Also, note that the data fall in a 2 x 6 contingency table of the type discussed in Section 3.4. Estimates of the odds ratios mentioned there are given in Table 1, and it is clear that they are not

TABLE 1
The severity of nausea in patients receiving chemotherapy

	Severity of nausea						
	None	*Mild*	*Moderate*			*Severe*	
	0	1	2	3	4	5	
No Cisplatinum	43 (42.4)	39 (39.0)	13 (13.6)	22 (21.0)	15 (18.5)	29 (26.5)	161
Cisplatinum	7 (7.6)	7 (7.0)	3 (2.4)	12 (13.0)	15 (11.5)	14 (16.5)	58
Odds ratio $\hat{p}_{1s}/\hat{p}_{2s}$	2.21	2.01	1.56	0.66	0.36	0.75	
$\hat{\phi}_s$	1	0.91	0.68	−0.11	−0.67	0	
S.E.	–	0.47	0.63	0.45	0.58	–	

The expected numbers under the stereotype model (8), with grades 0, 1, 2, and 3, 4, 5 indistinguishable, are given in parentheses. The goodness-of-fit is 2.70 on 4 d.f.

monotone. Hence, using the criterion of Section 3.4, the *observed* relationship between grade of nausea and Cisplatinum history is not ordered in the sense of (10). Further analysis is required to see whether this could be caused by sampling fluctuations, giving ordering as in (26).

Logistic regression models of varying complexity were fitted to the Cisplatinum data and the corresponding likelihoods are reported in Table 2. Recall that, in Section 4.2, we demonstrated that model (5) is one-dimensional if $p = 1$, as here, and that in this case $2(l(6) - l_0)$ has an asymptotic χ_5^2 distribution. From Table 2, the observed value of this statistic is 18.00 which is significant at the 0.005 level on a χ_5^2 test. Hence, the hypothesis of no relationship is rejected in favour of a one-dimensional relationship.

The maximum likelihood estimates of the (ϕ_s) for the saturated 6-category, one-dimensional model are given in Table 1 and it can be seen that:

(i) the $(\hat{\phi}_s)$ are not monotone, as expected since the odds ratios are not monotone;

(ii) the asymptotic standard errors of the $(\hat{\phi}_s)$ are high.

This suggests that questions about distinguishability should be settled before asking about the ordering of the relationship.

We approach the question of indistinguishability as in the latter part of Section 4.2. Firstly, we test $H_{(2)}$, for two groups of categories, against one group (H_0). The test statistic is $2(l(2) - l_0) = 15.36$, from Table 2 and $r^* = 3$. Using Tippett's procedure based on a χ_1^2 variate, this is statistically significant at the 1 per cent level, to the accuracy available in standard tables. Significant evidence in favour of two groups was expected since the case for a one-dimensional model has been established already.

TABLE 2

Likelihood analysis of the Cisplatinum data

Model	*Symbols*	*Number of parameters*	*Maximized* log L
Logistic/stereotype			
No relationship, $\boldsymbol{\beta} = 0$	l_0	5	−380.46
1-dim model, six categories (saturated model)	$l(6)$	10	−371.46
Distinguishability			
Two groups, $r^* = 3$	$l(2)$	6	−372.78
Three groups, $s^* = 2$, $t^* = 3$	$l(3)$	7	−372.69
Grouped continuous: proportional hazards			
Six categories: basic model		6	−383.18
Six categories: model modified for observer error		8	−373.86

Now we test $H_{(3)}$, three distinguishable groups of categories with respect to x_1, against two groups of categories of significance level α. Since $p = 1$, these are all one-dimensional hypotheses. From Section 4.2, the test statistic is $2\{l(3) - l(2)\} = 0.18$, from Table 2. Tippett's procedure tests this statistic as a χ_1^2 variate at the $1 - (1 - \alpha)^{1/4}$ level. Clearly, it is not significant.

In general, consideration would now be given to greater numbers of distinguishable categories. This is not necessary here because $2\{l(6) - l(2)\}$ is equal to the sum of the test statistics for three, four and five distinguishable groups of categories. These statistics are tested as χ_1^2 variates using Tippett's procedure and none of them is significant since, from Table 2, $2\{l(6) - l(2)\} = 2.64$, which is not significant at the 0.05 level, even as an ordinary χ_1^2 statistic. This argument covers the three group case above which was included for illustrative purposes. Hence it is concluded that there are only two groups of distinguishable grades (0, 1, 2; 3, 4, 5). The observed and expected numbers under this hypothesis are given in Table 1 and give the chi-squared goodness-of-fit statistic as 2.70 on 4 degrees of freedom. For comparison, fitting the next most complicated hypothesis with three groups of distinguishable grades (0, 1; 2; 3, 4, 5) gives a chi-squared statistic of 2.50 on 3 degrees of freedom. This is virtually the same value as for the simpler hypothesis. Both fits are acceptable, but the simpler hypothesis is preferred using the principle of parsimony. It is concluded that there is a significant relationship between y (grade of nausea) and x_1 (Cisplatinum history) and that nausea is worse on Cisplatinum. Furthermore, the relationship is as well described in terms of two distinguishable groups of grades as six.

It is interesting to compare the above results on the Cisplatinum data with those obtained by Farewell (1982), fitting a modification of the grouped continuous model (1) to allow for observer variation. Taking the underlying distribution, $F(\cdot)$, to have the proportional hazards form, the likelihoods for the basic and modified models are given in Table 2. It can be seen there that the results for the basic model compare unfavourably with the other models. The log-likelihood for the modified model is close to $l(3)$ for the two-group logistic model; it is actually greater than $l(3)$ but requires two more parameters. Statistical comparison of the two types of fit is not easy and would not be particularly rewarding. The information in the data available is insufficient to make a definitive choice between the models but in the absence of *a priori* information, the two-group logistic model has the advantage in the terms of simplicity.

5.2. *Severity of Disturbed Dreams*

Maxwell (1961, p.70) reports a study in which boys aged 5-15 are categorized by y, the severity of disturbed dreams on a four-point scale of increasing severity: 1 (not severe), 2, 3, 4. The data are reported in Table 3, cross-classified by age, x_1, and grouped in 2- or 3-year strata.

TABLE 3
Severity of disturbance of dreams in boys by age

	Degree of suffering from disturbed dreams				
	Not severe			*Very severe*	
Age	1	2	3	4	*Total*
5-7	7 (3.9)	4 (5.8)	3 (5.7)	7 (5.6)	21
8-9	10 (14.7)	15 (11.7)	11 (11.4)	13 (11.2)	49
10-11	23 (20.7)	9 (10.0)	11 (9.8)	7 (9.5)	50
12-13	28 (31.7)	9 (9.3)	12 (9.1)	10 (8.9)	59
14-15	32 (29.0)	5 (5.1)	4 (5.0)	3 (4.9)	44
ϕ_s	1	0.19	0.36	0	
S.E.	–	0.25	0.24	–	

The expected numbers under the stereotype model (8), with grades 2, 3, 4 indistinguishable, are given in parentheses. The goodness-of-fit statistic is 11.4 on 11 d.f.

The variable x_1 was assigned mid-point values for each stratum and was analysed as a continuous regressor variable. Hence there was one regressor variable and $p = 1$. Severity of disturbance is an assessed, ordered categorical variable, so there are good reasons to investigate the relationship between y and x_1 using the logistic family (5), with y as the response variable.

Since $p = 1$, all the models of type (5) are at most linear and are equivalent to type (8). The likelihoods for fitting the basic "four category" and the "no relationship" models are given in Table 4, together with the likelihoods for testing distinguishability. The notation is as in Section 4.2. Here $2(l(4) - l_0) = 22.72$ which is tested as an asymptotic χ_3^2 and found to be significant at the 0.001 level. The hypothesis of no relationship is rejected in favour of a one-dimensional relationship. The maximum likelihood estimates of the (ϕ_s) of the four-category model and their standard errors are given in Table 3 and again it can be seen that

(i) the $(\hat{\phi}_s)$ are not monotone and
(ii) the $\hat{\phi}_s$/S.E. ratios suggest questions about distinguishability.

TABLE 4

Likelihood analysis of the severity of disturbed dreams study

Model	*Symbols*	*Number of parameters*	*Maximized* log *L*
Logistic/stereotype			
No relationship, $\boldsymbol{\beta} = 0$	l_0	3	−288.49
1-dim model, four categories	$l(4)$	6	−277.13
Distinguishability			
Two groups, $r^* = 1$	$l(2)$	4	−277.98
Three groups, $s^* = 1$, $t^* = 3$	$l(3)$	5	−277.32
Grouped continuous: logistic			
Four categories: ordinary model		4	−278.47
Four categories: 2, 3, 4 indistinguishable		4	−277.98
Four categories: 2, 3 indistinguishable		4	−277.31

Using the approach of Section 4.2, distinguishability is examined stepwise up. Firstly we test for two groups of $(\boldsymbol{\beta}_s)$ against one group, H_0, and expect to reject H_0, as above. From Table 4, the test statistic $2(l(2) - l_0) = 21.02$. Using Tippett's procedure, this is tested as a χ_1^2 variate at the $1 - (1 - \alpha)^{1/3}$ level to give overall significance at the α-level. To the accuracy of standard Tables, this gives significance at the 0.01 level. Instead or proceeding along the lines of Section 4.2, we note that $2(l(3) - l(2)) = 1.70$. Hence, neither of the statistics for three and four groups of $(\boldsymbol{\beta}_s)$ can be significant using Tippett's procedure based on χ_1^2 variates. We conclude that there are only two distinguishable groups of categories, (1) and (2, 3, 4) with respect to x_1. The expected values for the stereotype model under this indistinguishability hypothesis are given in parentheses in Table 3 and give the chi-squared goodness-of-fit statistic $\chi^2 = 11.4$ on 11 degrees of freedom. This suggests an acceptable fit. For comparison, fitting the next most complicated model, with three indistinguishable groups of categories (1; 2, 3; 4) gives a chi-squared value of 11.2 on 10 degrees of freedom. This is virtually the same value as that given by the simpler hypothesis. We conclude that there is a significant relationship between age and severity of dreams which loses little by being described in terms only of "severe" and "not severe" disturbances.

It is instructive to compare the above results with those obtained from the grouped continuous model (1) with the logistic underlying distribution, $F(\cdot)$. It can be shown that the likelihood with this model under an indistinguishability hypothesis is also as given in (22). McCullagh (1980) discusses the basic likelihood fit and this is given in Table 4, together with the likelihoods allowing for two indistinguishability hypotheses. There is a small *gain* in likelihood terms by allowing indistinguishability. However, as seen in Section 2.1, formal tests for indistinguishability are not available here. It is concluded that the relationship between severity of dream disturbance and age is as well described in terms of two categories for y as four categories for both the grouped continuous and stereotype models. The latter model is preferred because of the unnaturalness of the indistinguishability concept in the grouped continuous model, as seen in Section 2.1.

5.3. *Back Pain Prognosis*

Now we discuss a multiple regression problem with 3 regressors. Doran and Newell (1975) describe a study in which patients suffering from back pain had several prognostic variables $(\mathbf{x})$ recorded at presentation. Three weeks after treatment, their progress, y, was assessed on a 6-point scale: *worse*, *same*, *slight improvement*, *moderate improvement*, *marked improvement* or *complete relief*. Anderson and Philips (1981) analysed the 101 patients treated by manipulation using the grouped continuous model (1) with the logistic underlying density. Using a stepwise approach, they found that it was necessary to include only three regressor variables: length of previous attack ($x_1 = 1, 2$), pain change ($x_2 = 1, 2, 3$) and lordosis ($x_3 = 1, 2$).

TABLE 5

Likelihood analysis of the back pain study

Model	*Symbol*	*Number of parameters*	*Maximized log-likelihood*
Logistic family			
Six groups: three-dimensional	$l(6)$	20	−149.51
Six groups: one-dimensional	l_1	12	−151.55
Six groups: no relationship	l_0	5	−171.53
Distinguishability: one-dimensional			
Five groups: $\phi_2 = \phi_3$	$l^*(5)$	11	−151.60
Four groups: $\phi_2 = \phi_3 ; \phi_5 = \phi_6$	$l^*(4)$	10	−152.79
Three groups: $\phi_2 = \phi_3 = \phi_4 ; \phi_5 = \phi_6$	$l^*(3)$	9	−154.39
Grouped continuous: logistic			
Six categories		8	−159.05

Taking $\mathbf{x}^{\mathrm{T}} = (x_1, x_2, x_3)$, the above data on 101 patients were fitted to the following logistic regression models: (i) no relationship; (ii) model (8) – one-dimensional with all (ϕ_s) different; (iii) model (5)–all six $(\boldsymbol{\beta}_s)$ different. The corresponding maximized likelihoods, l_0, l_1 and $l(6)$, are given in Table 5. Note that, since $p = 3$, model (iii) has the maximum dimensionality of three. Proceeding as in Section 4.2, we may test the overall relationship between y and $\mathbf{x}$ by $2(l(6) - l_0) = 44.0$ as a χ^2_{15} variate. The statistic for testing model (ii) against model (iii) is $2(l(6) - l_1) = 4.08$. On the standard argument, this would be tested as a χ^2_8 variate whose 5 per cent point is 15.51. Hence it seems safe to accept a one-dimensional model, despite the warning in Section 4.2 that the distribution of the statistic may not be χ^2_8. The statistic for testing a one-dimensional relationship against no relationship is $2(l_1 - l_0) = 39.96$. This would usually be tested as a χ^2_7 variate, whose 0.1 per cent is 24.32. Again, the warning in Section 4.2 is valid but it still seems safe to accept the one-dimensional model.

TABLE 6

Maximum likelihood estimates of the (ϕ_s) *and* $\boldsymbol{\beta}$ *for the one-dimensional logistic model* (8): *(i) six categories and (ii) three groups of categories*

	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$	$\hat{\phi}_5$	$\hat{\phi}_6$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Six categories	1	0.31	0.35	0.51	0.14	0	−2.63	−2.15	−1.31
S.E.	–	0.13	0.14	0.17	0.10	–	0.93	0.75	0.51
Three groups of categories $(\phi_2 = \phi_3 = \phi_4 ; \phi_5 = \phi_6)$									
	1		0.30		0		−2.79	−1.80	−1.05
S.E.	–		0.13		–		1.31	0.74	0.47

Maximum likelihood estimates of the (ϕ_s) and $\boldsymbol{\beta}$ for the six-category one-dimensional logistic regression model (9) are given in Table 6, together with their estimated standard errors. It can be seen that the $(\hat{\phi}_s)$ are not monotone and the standard errors are high compared with the differences $(\hat{\phi}_s - \hat{\phi}_{s+1})$. However, a significant relationship has been demonstrated and this is reinforced by comparison of $\hat{\boldsymbol{\beta}}$ with its standard errors. If we proceed along the lines discussed in earlier Sections, the question of distinguishability arises, particularly when we recall the difficulty of deciding the actual category on the six-point scale. The likelihoods required to test hypotheses of distinguishability are given in Table 5, this time proceeding stepwise down. If we extend the notation of Section 4.2, $l^*(s)$ is defined as the maximum of the likelihoods for s groups of indistinguishable categories in the one-dimensional model (8). As s decreases from 5 to 2, restrictions are imposed to ensure that categories judged indistinguishable for one value of s remain indistinguishable for smaller values of s. It can be seen that little is lost in taking five, then four groups of distinguishable categories, instead of the original six categories. The statistic $2\{l^*(4) - l^*(3)\} = 3.58$ tests whether four groups of categories are required as opposed to three groups. This statistic is not quite significant at the 0.05 level when judged as a χ_1^2 variate. It is even less significant when Tippett's procedure is used to allow for the implicit consideration of several statistics. Hence, three groups of categories are preferred to four groups. The question of whether two groups are preferred to three, is approached indirectly. We fit the model (8) under the hypothesis: $\phi_1 = 1$; $\phi_2 = \phi_3 = \phi_4 = \phi_2'$, say; $\phi_5 = \phi_6 = 0$. Table 6 gives 0.30 as the estimate of ϕ_2' with standard error 0.13. It is concluded that there are three distinguishable groups of categories since the 95 per cent confidence interval for ϕ_2', (0.04, 0.56), contains neither 1 ($= \phi_1$) nor 0 ($= \phi_5 = \phi_6$). Furthermore, the $(\hat{\phi}_s)$ are monotone in the sense of (26), indicating an ordered relationship. Table 6 also gives two estimates of $\boldsymbol{\beta}$, for six and three groups of distinguishable categories, respectively. There is no marked difference between the two estimates.

The question of model fit is never easy to resolve when the data are sparse, as here. To reduce this effect, the response categories were grouped, as above, into three distinguishable sets (1; 2, 3, 4; 5, 6). Expected and observed frequencies are given for these sets in Table 7, classified by the $12(= 2 \times 3 \times 2)$ possible $\mathbf{x}$-values and labelled by their $\hat{z} = \hat{\boldsymbol{\beta}}^T\mathbf{x}$ values. The goodness-of-fit statistic $\chi^2 = 12.1$ was calculated from the 36 cells without grouping for small expectations to avoid any subjective choice. To ignore any such difficulties would give 18(= 24–6) degrees of freedom and any grouping would be unlikely to give fewer than 6 degrees of freedom or to increase χ^2. Hence, judging $\chi^2 = 12.1$ on 6 degrees of freedom, the hypothesis of an adequate fit is accepted.

For comparison with the grouped continuous model, note that Anderson and Philips (1981) found the maximized log-likelihood to be -159.05 with 6 categories; see Table 6. When compared with $l^*(3) = -154.39$, this is perhaps slightly unfavourable to the grouped continuous model, particularly when the greater simplicity of the "three groups of categories" model is taken into account. If we proceed as in Section 5.2, tests of distinguishability of the categories within the grouped continuous model suggest that three groups of categories represent the data almost as well as six categories. Certainly the fall in the log-likelihood is small but the assessment of this is difficult; see Section 2. Since distinguishability is important here, again we prefer the logistic stereotype approach.

6. DISCUSSION

We have shown that the logistic family of regression models (5) contains models of varying dimensionality and an ordered regression model, the stereotype model. We have seen that the stereotype model can fit ordered regression data at least as well as the grouped continuous family of models and hence the logistic family is preferable on the grounds of its greater flexibility.

In the context of the logistic family, it is natural to question whether there is evidence of differences among all the categories. In one dimension this is equivalent to asking about differences among the (ϕ_s). We have chosen to look for subsets of the (ϕ_s) which are not significantly different, largely on the grounds that this can lead to model simplification. When all the $(\hat{\phi}_s)$ are not significantly different, an alternative is to fit a linear model for the (ϕ_s) as a function of s, $\phi_s = (k-s)/(k-1)$. The full specification of the linear model follows from the constraints $\phi_1 = 1$ and $\phi_k = 0$. Andrich (1979) suggests a similar approach for a related model; see below. An advantage of the linear model for (ϕ_s) is that the asymptotic likelihood ratio for a one-dimensional model against no relationship now has an asymptotic χ^2-distribution since the (ϕ_s) are now fully specified; see Section 4.3. However, there is a considerable price to pay in terms of model fitting when the (ϕ_s) are *not* linear as specified. Furthermore, the linear form does not permit any model simplification. We prefer the subset's approach for its simplicity and use the linear form for the (ϕ_s) only when the occasion demands it. In practice, there may be little demand for this linear model as there is accumulating empirical evidence that often little

TABLE 7

Observed and expected frequencies for the back pain study

y \ $\hat{z}$	−5.99	−3.82	−3.40	−1.23	−0.81	−0.28	1.36	1.89	2.31	4.48	4.90	7.07	*Total*
1 *Worse*	2 (2.49)	2 (1.18)	1 (0.95)	0 (0.28)	0 (0.04)	0 (0.04)	0 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	5
2, 3, 4 *Same, some improvement*	5 (3.99)	8 (8.68)	11 (9.40)	12 (12.84)	1 (2.46)	2 (2.90)	3 (2.77)	0 (1.70)	3 (1.58)	5 (3.81)	1 (0.92)	1 (0.95)	52
5, 6 *Marked improvement, better*	0 (0.52)	2 (2.15)	1 (2.84)	8 (6.88)	3 (1.50)	3 (2.06)	3 (3.22)	4 (2.30)	1 (2.42)	10 (11.19)	3 (3.08)	6 (6.05)	44
Total	7	12	13	20	4	5	6	4	4	15	4	7	101

The expected frequencies under the stereotype model (8), with three groups of indistinguishable categories (1; 2, 3, 4; 5, 6), are given in parentheses.

information loss results from approximating the ordered regression models by binary regression models. The examples in Sections 5.1 and 5.2 demonstrate this.

Andrich (1979) introduced a model very similar to the stereotype model. With suitable changes of notation and interpretation, his model (6) is identical to model (8) here. However, he chose to restrict his (ϕ_s) to be linear in s, leading to the above difficulties and did not advocate the use of his equivalent of the stereotype model. Andrich's motivation for his models was quite different from the approach taken here and was rather difficult to accept (McCullagh, 1980). He did not relate his model to the logistic family (5).

The approach of this paper can be modified for use with ordered categorical regressor variables. The objective is to find the regression model which is most economical in the number of parameters postulated, subject to adequate model fit. Possible forms for an ordered variable range from unordered categorical to binary. Anderson (1983b) discusses these ideas in more detail. Barlow *et al.* (1972) introduced the isotonic regression method which gives an alternative approach for an ordered regressor variable when the response variable is univariate normal.

A major advantage of the stereotype model is that it belongs to the exponential family, whereas the grouped continuous model does not. Hence, the stereotype model can be used in more complicated situations, such as block designs and paired comparisons, where conditional inference is required.

ACKNOWLEDGEMENTS

The programs referred to in the text were written by Dr P. R. Philips; his help is gratefully acknowledged. Comments by P. M. E. Altham, V. T. Farewell, P. McCullagh, R. L. Plackett and the referees on earlier drafts were very helpful. Partial support for this work was provided by NIH grants GM-28314 and GM-24472.

REFERENCES

Anderson, J. A. (1972) Separate sample logistic discrimination. *Biometrika*, **59**, 19–35.

——— (1982) Regression modelling for ordered categorical variables. Technical Report No. 52, Department of Biostatistics, Seattle, University of Washington.

——— (1983a) Robust inference using logistic models. *Bull. Int. Statist. Inst.* **48** (2), 35–53.

——— (1983b) Logistic regression methods in risk assessment. In *Environmental Epidemiology: Risk Assessment* (R. L. Prentice and Alice S. Whittemore, eds). Philadelphia: SIAM.

Anderson, J. A. and Blair, V. (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*,,**69**, 123–136.

Anderson, J. A. and Philips, P. R. (1981) Regression, discrimination and measurement models for ordered categorical variables. *Appl. Statist.*, **30**, 22–31.

Andrich, D. (1979) A model for contingency tables having an ordered response classification. *Biometrics*, **35**, 403–415.

Ashford, J. R. (1959) An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics*, **15**, 573–581.

Atkinson, A. C. (1970) A method for discriminating between models (with discussion). *J. R. Statist. Soc.* B, **32**, 323–353.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference Under Order Restrictions*. New York: Wiley.

Chung, K. L. (1968) *A Course in Probability Theory*. New York: Harcourt, Brace.

Cox, D. R. (1970) *The Analysis of Binary Data*. London: Methuen.

Doran, D. M. L. and Newell, D. J. (1975) Manipulation in treatment of low back pain: a multicentre trial. *Brit. Med. J.*, **1**, 161–164.

Farewell, V. T. (1982) A note on regression analysis of ordinal data with variability of classification. *Biometrika*, **69**, 533–538.

Fienberg, S. E. (1980) *The Analysis of Cross-Classified Data*, 2nd ed. Cambridge: MIT Press.

——— (1982) Using information on ordering for log-linear model analysis of multi-dimensional contingency tables. Presented at the 1982 International Biometrics Conference, Toulouse, France.

Gill, P. E. and Murray, W. (1972) Quasi-Newton methods for unconstrained optimization. *J. Inst. Math. & Applic.*, **9**, 91–108.

Goodman, L. A. (1981) Association models and canonical correlation in the analysis of cross-classified data having ordered categories. *J. Amer. Statist. Assoc.*, **76**, 320–334.

Green, P. J. (1983) Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. Unpublished manuscript.

Haberman, S. J. (1981) Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Ann. Statist.*, **9**, 1178–1186.

Hinkley, D. V. (1970) Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17.

Koziol, J. A. and Perlman, M. D. (1978) Combining independent chi-squared tests. *J. Amer. Statist. Assoc.*, **73**, 753–763.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

Maxwell, A. E. (1961) *Analysing Qualitative Data.* London: Methuen.

Plackett, R. L. (1981) *The Analysis of Categorical Data*, 2nd ed. London: Griffin.

Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

Thompson, R. and Baker, R. J. (1981) Composite link functions in generalized linear models. *Appl. Statist.*, **30**, 125–131.

APPENDIX

Outline proofs for Properties 1 and 3

Property 1

Substituting $z = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ in (8), we obtain

$$d\{P_t(z)\}/dz = -P_t(z) \sum_{u=1}^{k} (\phi_t - \phi_u)\, P_u(z).$$

On substitution of these into $d\{F_z(s)\}/dz = \Sigma_{t=1}^{s}\, d\{P_t(z)\}/dz$, it follows that this derivative is negative for all z, if (10) is true. Hence, $F_z(s)$ is monotone decreasing in z and (Y_z) is stochastically increasing with respect to z.

Property 3

(i) This follows directly from (8), using (10).

(ii) If the (π_z) are equal, from Section 3.1, $\beta_{0s}^* = \beta_{0s}$ and $z_{st}^* = z_{st}$. Denote by Z_s the random variable with density function $h_s(z)$ in (16). Since the probability content of Z_s is unity for $s = 1, \ldots, k$, z_{st} in (14) can be rewritten as

$$z_{st} = -\{1/(\phi_s - \phi_t)\} \log E_{Z_t} [\exp\{-(\phi_s - \phi_t)\, z\}] \text{ for all } s, t; s \neq t.$$

Define $H(\phi) = -(1/\phi) \log E_Z [\exp(-\phi_z)]$. With the help of the Liapounov inequality (Chung, 1968, p. 45), $H(\phi)$ can be shown to be monotone decreasing in ϕ, for a range of values of ϕ. This includes the (ϕ_s) for $Z = Z_s$, $s = 1, \ldots, k$. If (10) holds, the sequences (ϕ_s) and $(\phi_s - \phi_t)$ are monotone decreasing for all s and t and the required ordering for z_{st} follows.

DISCUSSION OF PROFESSOR ANDERSON'S PAPER

Dr P. McCullagh (Imperial College of Science and Technology, London): First let me record my deepest sympathy for the Anderson family.

John wrote part of his paper during the summer of 1982 when he and I were both visiting the Fred Hutchinson Cancer Research Center in Seattle. Although our views were seldom in close agreement, we had numerous frank but always friendly discussions on the merits of various models for ordinal responses. My contribution this evening is based partly on my recollection of those discussions.

First, we were in agreement that models, particularly of the log-linear variety, developed primarily for nominal or unordered responses, are seldom appropriate for ordinal variables. This is not because of failure to fit the data but because the parameters are seldom interpretable or relevant in the context of the ordinal response observed. There is therefore a need for specific models for ordinal response variables, e.g. the grouped continuous models discussed in Section 2.1.

Grouped continuous models are appropriate only if the order of the categories of the measurement scale is well defined as is the case in Examples 5.1, 5.2 and 5.3. However, it may occasionally be the case that different orders are appropriate for different purposes or, for some other reason, the appropriate order is unclear. For example, newspapers could be ordered according to either "quality" or "political bias". The so-called stereotype models of Section 3, which I prefer to call canonical regression models, have the property that the order need not be specified in advance and would therefore seem appropriate in the latter case where order is ambiguous. To explore the connection with canonical regression, let π_{is} be the probability, for the ith multinomial sample with covariate vector $\mathbf{x}_i$, of observing response category s. Professor Anderson's model (8) may be written

$$\log \pi_{is} = \alpha_i + \beta^*_{os} - \phi_s \quad \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \qquad s = 1, \ldots, k; \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ and $\mathbf{x}_i$ are of length p. Since α_i is an irrelevant nuisance parameter, the analogous multivariate linear model is

$$E(Y_{is}) = \mu_{is} = \beta^*_{os} - \phi_s \quad \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i. \tag{2}$$

In other words, the k regression planes have arbitrary intercepts and proportional slopes. (Use of the word "parallel" in Section 3.2 may therefore be misleading.) The linear combination of the Y's having maximal regression on $\mathbf{X}$ is $\Sigma\phi_s \mathbf{Y}_s$, also called the linear discriminant function. All orthogonal combinations have zero regression on $\mathbf{X}$. Thus (2) implies (and is in fact equivalent to) the statement that there is a single non-zero canonical root with vectors $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$. In a similar way, models corresponding to several canonical roots can be constructed as in equation (11) in the paper, so that Professor Anderson's "dimensionality" is the number of canonical roots.

It is not difficult to think of examples where model (2) would be appropriate. For example, if the components of $\mathbf{Y}$ were k lineal measurements on n crabs of p different species ($\mathbf{X}$) having common shape but different size, model (2) would apply with $\beta^*_{os} = 0$. In other words,

$$E(Y_{is}) = \phi_s \beta_i$$

so that $\{\phi_s\}$ determines the crab shape. Similarly, if $\beta^*_{os} \neq 0$, then taking $\beta_1 = 0$, $\phi_s \beta_i$ determines the difference in mean size between the ith species and the first so that, in this case, $\{\phi_s\}$ determines the direction of change in size. In Professor Anderson's model $\{\phi_s\}$ determines the direction of change of the probability vector per unit change in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. However, I find it difficult to construct equally compelling examples for the discrete model.

Nevertheless, as an exploratory tool, the canonical regression model with one or two roots ((8) or (11) in the paper) may prove useful for seriation or discrimination when the order of the categories is in doubt or is thought to be irrelevant. In the more usual case where the order is important and not in doubt, the grouped continuous models would appear to be preferable because of the ease of parameter interpretation and invariance under combination of adjacent categories.

Significance tests for the hypothesis of no relationship (no non-zero canonical roots) present problems because of the lack of identifiability under H_0: $\boldsymbol{\beta} = \mathbf{0}$. In view of the discussion above, it is not surprising that the approximate null distribution of the likelihood-ratio statistic is that of the largest root of a Wishart matrix, a result proved by Haberman (1981). The adequacy of normal approximations for the parameter estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\beta}}$ does not seem to have been investigated.

This paper is a substantial contribution to an important and lively area of Statistics. In its emphasis on applications, it is a fitting testimonial to John Anderson the Statistician.

Dr G. D. Murray (University of Glasgow): This paper makes a significant contribution to what is becoming an important growth area in statistics, and it is fitting that it should show so clearly John's unique blend of mathematical sophistication and a commonsense approach to the analysis of real data.

The paper rightly makes the case that an ordered categorical response variable requires careful modelling, but I should like to extend the argument and make an appeal for special treatment for ordered categorical regressor variables. It is mentioned briefly in Section 6 that isotonic regression could be used in this situation if the response variable is univariate normal, but the approach is equally applicable with an ordered categorical response variable. This is seen most easily in the situation where the response can take only two values, good and bad say. The problem is to estimate the probability of a good response subject to the constraints that this response probability should increase monotonically as each regressor ranges from its worst to its best value. This imposes a partial order on the binomial response probabilities, and is a standard isotonic regression problem (Barlow *et al.*, 1972, p. 38). When the response y can take several values, the problem is then to estimate a set of multinomial response probabilities for each vector $\mathbf{x}$, and these probabilities can be constrained by saying that for every possible cutpoint which could group the responses into a "good" and a "bad" set, the probability of lying in the good set should be a monotonic function of each regressor.

One advantage of this approach over the logistic approach is that it is simple to handle missing data on a regressor variable. If the value of a particular regressor variable is not observed then one constrains the response probability to lie between the values it would take when the regressor takes its best possible and worst possible values. It is of course possible to think of situations where "missing" would be associated with a higher (lower) probability of a good response than the best (worst) possible value which the regressor might take, but the assumption adopted with this approach is far weaker than the usual assumption adopted to handle missing data, namely that the data are missing at random.

Dr V. T. Farewell (Fred Hutchinson Cancer Center, USA): First, on behalf of many people in the Pacific Northwest, I should like to say that it was most enjoyable to have Professor Anderson and his family in Seattle during the preparation of this paper. At one point during John's visit we were playing tennis and I was having some success, a rare event. On a critical point I hit, surprisingly, a one-handed top-spin backhand across court and felt sure the point was one. As I admired the shot, John scrambled across the court to hit an extremely unorthodox winner down the line. He went on to win the match.

In some sense that incident parallels this paper. Grouped continuous ordinal regression models are a current favourite. This paper, which is perhaps a little *ad hoc* in methodology, asks significant questions about those models and deserves careful reading. My feeling, however, is that the paper, although important, is not quite so devastating as Professor Anderson's tennis.

The quality of available data must be a concern with an assessed ordinal variable. In the examples, indistinguishability is a recurring concept. However, indistinguishability appears often to be less well characterized by the relationship of categories to available covariates than by an arbitrariness in the assessment. If the process of category choice varies by category level then a truly ordinal variable does not exist. Any failure to fit resulting data may be attributed to poor quality data as easily as to a chosen regression model.

The methodology in this paper does characterize the response variable and leads to simple models when appropriate. An objection is that there may be simpler ways to proceed.

If doubts about distinguishability exist, a reasonable approach is to examine conditional regression models. For example, a four point scale might have separate models $\mathrm{pr}\,(Y=1 \text{ or } 2 \mid \mathbf{x})$, $\mathrm{pr}\,(Y=1 \mid \mathbf{x}, Y\in\{1,2\})$ and $\mathrm{pr}\,(Y=3 \mid \mathbf{x}, Y\in\{3,4\})$. If the breakdown of categories is natural, this is much simpler, and easier to explain, than is the fitting of multinomial logistic regression models. On the other hand, except for indistinguishability attributable to assessment, some consistency in a regression relationship across category boundaries is desirable and reasonable. A model which allows different covariates to discriminate between different categorizations would need considerable support from the field of application before being accepted.

A technical point concerns variability in the regression coefficient in the grouped continuous model. A regression coefficient which varies across categories does not, in principle, appear to be a major problem. The use of time-dependent covariates in survival analysis is a starting point to study this possibility. The conditional and other likelihoods discussed by McCullagh and Nelder (1983, Ch. 5, p. 116) may be useful in this regard. It will not always be easy to interpret a varying regression coefficient. If indicated, however, the appropriateness of the ordinal regression might be re-examined, as a first step.

Finally, I greatly regret that Professor Anderson is not able to defend this impressive paper against my trifling criticisms and to continue to beat me on the tennis court.

Mr R. Mead (University of Reading): Tonight's paper provides an informative and thought-provoking view of an approach to analysing subjectively assessed variables, and it bears eloquent testimony to those qualities which we shall miss with the passing of John Anderson. Discussion of the method of analysis of data should always lead to considerations of the design of the investigation producing data in the form most suitable for that method of analysis. I therefore propose to discuss design and, in particular, the design of levels for subjectively assessed variables in the context of the methods of analysis described in the paper.

First, I think there is a distinction to be drawn between the use of subjectively assessed variables in designed experiments and in observational studies. In designed experiments, such as agricultural crop trials, each experimental unit is classified by various factors, and there will usually be at most a single unit for each combination of levels. The analysis of subjective scores usually assume that the scores can be assumed to be approximately normally distributed, and this has usually been interpreted as implying that the scores should use as wide ranging a scale as is practically viable to make the approximation good.

In contrast, for observational studies, there will usually be a sizeable sample for each combination of classificatory factor levels. The data consist of frequencies of the various possible values of the subjectively assessed variables, and the analysis seeks to discriminate between the levels of treatment factors in terms of the changing frequencies of the values of the observed variable. One formulation of the design problem would then seem to be as follows: choose the values of the assessed variables so as to optimize discrimination between the distribution functions, of the incompletely defined variable, for the two or more treatment factor levels. For general distributions, as are discussed in the paper, the difference between distribution functions for two treatment levels would be expected to rise monotonically to a maximum and then to decline monotonically.

The general implications of optimal design for the estimation of such a difference function would suggest using just three values for observations. Therefore the subjectively assessed variables should have three values chosen so that, over the two treatment levels to be compared, the total occurrence frequencies of the three values should be similar.

If we now examine the three examples in the paper, the proposed analysis reduces the number of distinct values of the observed variable to two or three. This provides at least some support for the general implication of design theory that only a small number of values should be necessary.

It may be proposed that, since an analytic method which allows amalgamation of non-distinguishable values is available, then it would be sensible initially to allow more than the minimum three values. However, the disadvantage of inconsistent interpretation of the differences between many values must reduce the information available. If the credibility of the intuitive design argument is stretched to its limit, then it could be argued that the presence of surplus values in the examples has led to the confusion from which only two values can be distinguished. Alternatively, the curvature of the difference function over the observed range may be too slight to require three values, the two distinguishable values being sufficient for an approximately linear difference function.

Dr Susan R. Wilson (Australian National University): The literature on statistical models for ordered categorical response variables is relatively sparse, so John Anderson's paper is a welcome addition to the subject. Many of the currently available techniques for dealing with (ordered) categorical data are computational (Nishisato, 1980, Ch. 1) and need justification by statistical models. The particular model developed in this paper, and its relationship with canonical regression analysis, would be useful for exploring such aspects.

John was well known for his interest in statistical problems arising from practical experience, and I wish to make my comments on this paper from this viewpoint. One commonly encountered form of response which is usually considered as being "ordered" is: Yes, Don't Know, No. It will often not be particularly realistic for this middle category to have the unimodality property, *P*2. However, realistically restricting the range of z (i.e. the values of $\boldsymbol{\beta}^T\mathbf{x}$) would generally remove this difficulty. Also, one can readily envisage circumstances for which the response is an ordered categorical variable, but the properties of Section 3.5 are not appropriate. For example, for the above ordered response, one might encounter a polarization of the population into the two extreme categories, Yes and No, with increasing length of, say, media exposure, while the proportion responding in the middle category, Don't Know, decreases with length of exposure.

Fig. 2 demonstrates the use of the model given by equations (8) and (10) for predictive purposes. However, if description of the underlying phenomenon, rather than prediction, is the primary purpose of the analysis, then the corresponding diagrams may be quite different, yet useful. For example, in psychiatric (and other medical) applications, the probability of the (extreme) response, "normal", will usually dominate, over the range of z, the probabilities of the other discrete responses corresponding to (increasing) degrees of abnormality. It is the relative change of the probabilities with z that one wishes to investigate. To achieve a desirable ordering, it may be necessary to impose constraints on (β_{0s}).

It is implicit in the exposition of the proposed model that the (ordered) categorical response is univariate. All the discussed examples are of this nature. Often one encounters multivariate categorical response variables with ordering properties. Generalization of the stereotype model to such situations is not obvious.

Finally, I note that it is possible, although somewhat cumbersome, to fit the model given by equation (8) using GLIM 3, by incorporating the technique developed by Adena and Wilson (1982, Ch. 11) combined with a composite link-function approach.

Dr P. J. Green (University of Durham): In a comparison between the grouped continuous regression model and Professor Anderson's logistic models, I shall restrict attention to the narrow point of view that they provide alternative parameterizations for several multinomial distributions on the same k response categories. The logistic models then seem superficially unattractive for $k > 2$; equation (5) is only easily interpreted in comparing *pairs* of categories.

However, I have found the results of fitting such models very appealing qualitatively, even in cases where the grouped continuous model appears more natural. Consider the data of Table D1, relating final classification to A-level performance, assembled from several years' records for a certain degree course. Various sources of variation have been ignored here, so we should not overemphasize the quantitative results of any analysis.

TABLE D1
Final degree classification and A-level performance

A-level	*Degree*				
score	I	II_1	II_2	III	*Pass*
15	22 (13.6, 18.1)	13 (18.4, 13.3)	10 (10.8, 12.3)	3 (4.1, 3.3)	0 (1.1, 1.0)
14	20 (17.5, 20.7)	21 (30.3, 25.2)	31 (22.6, 25.2)	9 (9.9, 8.9)	2 (2.7, 3.0)
13	13 (17.3, 17.2)	43 (36.8, 34.5)	31 (35.3, 37.5)	16 (18.3, 17.6)	10 (5.4, 6.3)
12	7 (9.3, 7.4)	21 (23.5, 24.6)	35 (28.9, 28.9)	18 (18.3, 18.2)	5 (5.9, 6.9)
11	3 (6.9, 4.1)	21 (19.7, 22.7)	26 (30.4, 29.0)	32 (24.2, 24.3)	8 (8.9, 9.9)
10	3 (4.1, 1.8)	17 (12.9, 16.3)	25 (24.4, 22.6)	20 (24.8, 25.2)	12 (10.7, 11.0)
9	1 (1.7, 0.5)	10 (5.7, 7.9)	9 (12.9, 11.8)	15 (16.8, 17.6)	11 (8.9, 8.2)
8	1 (0.6, 0.1)	2 (2.2, 3.3)	4 (5.8, 5.4)	12 (9.8, 10.8)	6 (6.5, 5.3)
7	0 (0.2, 0.0)	1 (0.6, 1.0)	2 (1.8, 1.8)	6 (3.9, 4.7)	1 (3.4, 2.5)
6	0 (0.0, 0.0)	0 (0.1, 0.2)	2 (0.4, 0.4)	1 (1.1, 1.5)	0 (1.3, 0.8)

The figures in parentheses give the expected numbers under the grouped continuous (logistic) model (1) and the stereotype model (8, 10), respectively. Deviances: 48.6 and 33.2. Stereotype estimates, with standard errors:

ϕ_2 : 0.463 (0.076); ϕ_3 : 0.377 (0.072); ϕ_4 : 0.070 (0.082).

Grouped continuous models based on the logistic or normal distributions, fitted by maximum likelihood, seem to provide an adequate fit, and the allegedly latent response variable has an immediate interpretation as an examination mark underlying the degree classification. However, Professor Anderson's stereotype model, equation (8), has something to add, for it provides a closer fit (see Table D1). What is more, the fitted ϕ parameters suggest that some differences are not significant, and indeed on following the procedure of Section 4.2, the two pairs of categories (II_1, II_2) and (III, pass) are indistinguishable with respect to A-level score.

It is instructive briefly to examine possible stochastic mechanisms generating the models; Consider the two-dimensional versions. In the logistic model given by equations (5) and (11) we consider a latent space in which the axes represent two combinations of covariates. Values of the regressors at each observation are represented by points, and certain directions favour allocation to the corresponding categories. These categories then play dice with each other to gain the observations. The dice can be loaded so that equation (5) holds.

In an obvious generalization of the grouped continuous model, the covariate information is displayed as before, but the categories are now also represented by points. Each observation is randomly displaced, and then allocated to the nearest category. This model *can* demonstrate indistinguishability with respect to some regressors. Presumably it is such a model that Professor Anderson rejects as "computationally impractical" but it does seem intrinsically more appealing to assign random noise to the observations rather than the competing categories.

This seems to me a thought-provoking paper presenting a valuable class of models, and I hope that others will follow up the work that John was unable to continue.

Professor D. M. Titterington (University of Glasgow): I have a few very brief remarks.

(i) It is a little disappointing that in none of the examples shown do the original $\hat{\phi}$'s turn out to be monotonic although this is subsequently induced after indistinguishability considerations.

(ii) The crux of the paper is a set of linear forms $\{\beta_{0s} + \boldsymbol{\beta}_s^T\mathbf{x}\}$, on the parameters of which certain constraints are imposed. There is surely scope for considering similar constraints within the models proposed by McCullagh (1980) or indeed within comparable Normal-theory regression models. Professor Anderson refers, in Section 4.2, to one such application, by Rao (1973); see also Dr McCullagh's comments.

My final remark is, inevitably, a personal one. John Anderson visited Glasgow frequently and I think that the best way of describing our Department's regard for him is to say that, when he did come, he was hardly regarded as a visitor at all.

Dr J. T. Kent (University of Leeds): I have three comments to make about this very interesting paper which has clarified many aspects of the use of logistic models. First, the analysis of stereotype models reduces to the use of canonical variates when underlying multivariate normality is assumed. Thus it would be enlightening to see a comparison of the stereotype analysis with the more usual analysis based on normal theory. Such an analysis would provide both an alternative interpretation of the data and a starting point for the logistic analysis.

My second comment concerns the statement in Section 4.2 that under normality it is expected that the logistic analysis will be less powerful than the standard normal theory tests. I am a bit surprised at this and would have expected the logistic tests to be asymptotically efficient in many cases. However, so far I have only been able to check this assertion in the simple case of two equally likely groups.

Thirdly, it is pointed out in Section 4.2 that the dimensionality test of H_{s-1} against H_s, $1 \leqslant s \leqslant d = \min(p, k-1)$, using the likelihood ratio statistic $2(l_s - l_{s-1})$ will not generally have an asymptotic χ^2 distribution. There are two points to raise here. First, this distributional problem does not arise in the usual approach in multivariate analysis to such problems. Under the usual method, one tests H_s against H_d sequentially, $s = 0, \ldots, d-1$, until a non-significant result is obtained. If H_s is true (but H_{s-1} is not true), then $2(l_d - l_s)$ does have an asymptotic χ^2 distribution. However, this method will be less powerful than the first approach and it ignores the problem that the successive tests will not be independent of one another. The second point, already mentioned by Dr McCullagh, is that the null asymptotic distribution for the particular test statistic $2(l_1 - l_0)$ is not intractable, but is distributed as the largest eigenvalue from a Wishart distribution. Further, this distribution is tabulated, for example, in Pearson and Hartley (1972, p. 352).

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

Dr A. Albert (University of Liège, Belgium): This extremely stimulating paper by Professor Anderson constitutes a major methodological contribution to theoretical statistics and at the same time it provides us with a new and powerful tool for practical applications. John Anderson's life was a continuous plea in favour of logistic methods, which he felt to be particularly suitable for problems involving a mixture of continuous and discrete variables. This situation arises frequently in Medicine and it is therefore not surprising that the examples used in the paper all deal with medical applications. When I first met Professor Anderson, we discussed the problem of discriminating between two "quantitatively" (or ordered) distinct groups, as opposed to the classical nominal situation. I believe this might have been the starting point of his interest in ordered categorical variables (Albert and Anderson, 1981). The method was extended to the k-group situation (Anderson and Philips, 1981), but he felt uncomfortable about the grouped continuous logistic model which did not belong to the exponential family. By introducing the "stereotype" model for assessed ordered categorical variables, Professor Anderson extended considerably the use and applicability of the logistic method. I would like to comment briefly on the maximum likelihood estimation section. In a forthcoming paper (Albert and Anderson, 1984), it is shown that the existence of MLE in logistic regression models depends on the configuration of the data points in the observation-space. These fall essentially into three mutually exclusive categories, respectively referred to as complete separation, quasi-complete separation and overlap. The first two patterns give non-unique infinite estimates while the third always yields a finite and unique solution. For assessed ordered categorical variables, complete or quasi-complete separation is unlikely to occur in practice because of the frequent indistinguishability between adjacent categories, but it might occur in small samples. I believe that the theorems proved in our paper, which could not be extended to the grouped continuous regression model (1), hold true for the stereotype model, since it belongs to the exponential family. For example, complete separation occurs if there is a set of vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ such that

$$(\boldsymbol{\beta}_s - \boldsymbol{\beta}_t)^{\mathrm{T}} x_i > 0 \quad \text{for all } i \in E_s, \quad s, t = 1, \ldots, k \quad (s \neq t),$$

where E_s is the set of points from $y = y_s$ and $\boldsymbol{\beta}_s^{\mathrm{T}} = (\beta_{0s}, \phi_s \ \boldsymbol{\beta}^{\mathrm{T}})$. Then the likelihood function attains its absolute maximum at infinity, whence the MLE do not exist. Professor Anderson has again initiated an important branch of regression analysis with a logistic model that is both general and simple. It is truly regrettable that his untimely death interrupted an already impressive yet so promising career.

Mr D. Bancroft (Consumers Union of U.S. and Yale University): One possible explanation as to why the one-dimensional model (8) seems adequate most of the time: the non-linearity of the $\phi_s \ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ components fits well a large number of random departures from null models with $\boldsymbol{\beta} = \mathbf{0}$ as well as from multi-dimensional models. This versatility is reflected in the imprecision of the $\hat{\phi}$ estimators, which tells us that the model fits well but the estimates are not reliable. In a linear model context an equivalent problem is to decide the rank of the coefficient matrix $\mathbf{B}$ in the model $E\{\mathbf{Y}\} = \mathbf{XB}$, where $\mathbf{X}$, $\mathbf{Y}$ are matrices. For normal errors independent between rows and with the same covariance matrix in each row, this reduces to an examination of the roots of $\det(\hat{\mathbf{B}}\hat{\mathbf{B}}^{\mathrm{T}} - \lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$, where $\hat{\mathbf{B}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. In this eigenvalue problem the relevant changes in the log-likelihood do not follow chi-squared distributions (Anderson, 1958, Ch. 14). For low sample sizes chi-squared approximations will suggest that a low rank $\mathbf{B}$ is adequate when it is not. This analogy suggests that the difference $2(l_d - l_1)$ in Section 4.2 will be stochastically smaller than a χ^2 random variable with the appropriate degrees of freedom, leading to optimistic acceptance of one-dimensional models.

Some care is necessary in selecting a numerical procedure for maximum likelihood parameter estimation. For the one-dimensional model (8) Newton–Raphson applied to the log-likelihood is not equivalent to the method of "efficient scores" (Rao, 1973, p. 370). The scoring procedure converges linearly with the biggest changes in estimates alternating from step to step between ϕ and β parameters. My test data set consisted of food preferences from a taste test (Scheffé, 1952), and the regression model was the Bradley–Terry model for paired comparisons together with a preference for the first in a pair. This model was fitted subject to symmetry constraints on the strength of preference categories. The scoring procedure was extremely sensitive to the initial parameter estimates, failing (catastrophically) to converge on several occasions. The Newton–Raphson procedure was more stable and converged quadratically. Compared to fitting the same regression model in the grouped continuous logistic model, the stereotype model required about two more iterations to achieve comparable accuracy in the $\boldsymbol{\beta}$ estimates. It also gave better fits than the grouped continuous logistic model, but at some cost. The coefficients of variation of the estimates of $\boldsymbol{\beta}$ were usually larger for the stereotype model, even when the variances in each model were inflated to reflect lack of model fit (Cox, 1970, p. 77).

An observation on maximizing a log-likelihood $L(\boldsymbol{\beta})$ subject to the constraint $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \mathbf{C}^{\mathrm{T}}\boldsymbol{\delta}$, where $\mathbf{C}$ and $\boldsymbol{\beta}_0$ are given, while $\boldsymbol{\delta}$ is free:

$$\frac{\partial L}{\partial \boldsymbol{\delta}} = \mathbf{C}\,\frac{\partial L}{\partial \boldsymbol{\beta}}, \quad \frac{-\partial^2 L}{\partial \boldsymbol{\delta}\,\partial \boldsymbol{\delta}} = \mathbf{C}\left(\frac{-\partial^2 L}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}}\right)\mathbf{C}^{\mathrm{T}},$$

where the $\boldsymbol{\beta}$ partial derivatives treat $\boldsymbol{\beta}$ as unconstrained. This use of the chain rule allows constraints more complicated than those leading to (22) to be fitted very easily.

Professor D. J. Bartholomew (London School of Economics): I am sorry that I am unable to be present at the meeting to add my tribute to John Anderson whose contributions to statistics are so well exemplified in this paper. Apart from the intended application to regression analysis, some of the ideas promise to be useful in constructing models for the factor analysis of ordered categorical data. The connection arises because, in the factor analysis context, the observed categorical variables are assumed to be regressed on latent, rather than observable, random variables. Models developed in one area are thus potentially useful in the other.

In my paper read to the Society on the factor analysis of categorical data (Bartholomew, 1980) I proposed a model of the same form as that of McCullagh's given in equation (1) with a logistic function for F. This seemed natural at the time but some disadvantages have since become apparent. In part these arise from considerations like those the author raises in connection with the nature of assessed variables. But more important is the fact that the model does not, as in the binary case, lead to a single sufficient statistics (in the Bayesian sense) for the distribution of each latent variable. In order to preserve this property I formulated an alternative generalization of the binary model which was reported in Bartholomew (1983). This proposes a model for the distribution of each manifest variable, y, of the form given in equation (5) with appropriate order restrictions on the β's. Bearing in mind that there will be a different set of β's for each y, the number of parameters to be estimated and interpreted is typically rather large. The stereotype model offers one way of reducing the number to manageable proportions and I shall certainly explore its possibilities. In factor analysis applications, normal rather than logistic regression models have been used. Although the practical difference between the two is small, this paper adds further weight to the case for the logistic.

I am sure that in years to come we shall have many occasions to refer back to this paper and thus be reminded not only of John's achievements in our subject but also of what might have been.

Professor S. E. Fienberg (Carnegie–Mellon University, USA): It was with mixed emotions that I read this paper. John Anderson and I became personal friends during my extended visit to the University of Newcastle in 1978, and I was shocked by the news of his untimely death. My distress that this may be the last paper of his that we will see was, however, in part counteracted by the enjoyment and stimulation I felt as I read Professor Anderson's innovative ideas on the analysis of ordered categorical response structures using the stereotype regression model. Although we came to this problem from quite different directions, our views turned out to be remarkably similar, and I wish he were able to respond to my comments and queries.

At various points in the paper, Professor Anderson discussed the superposition of the ordering restriction on the scaling parameters, i.e. the $\{\phi_i\}$, and noted "potential" difficulties. These difficulties should not be underestimated. When ordering is not imposed *a priori* and the estimated regression model is in fact ordered (i.e. (25) holds) then all is well. Otherwise a markedly different estimation algorithm may need to be used. For the one-dimensional stereotype model a variant of Anderson's estimation approach using something like the pool-adjacent-violators algorithm (see, for example, Barlow *et al.*, 1972) will likely converge, but I would speculate that for the two- and higher-dimensional models a more complex procedure will be required. Also the corresponding likelihood ratio test then *will not* be asymptotically distributed as chi-squared under the null hypothesis. Even in the unrestricted case, I would be somewhat wary of the blind use of the recommended quasi-Newton algorithm, especially for large problems. My own experience in such settings is that the likelihood is typically maximized on or near the boundary of the parameter space and modifications in standard algorithms often are required. It would not surprise me, however, to learn that GLIM could be adapted to handle the general stereotype model. Has anyone attempted to do this?

Professor Anderson observed that in testing H_0 ($\boldsymbol{\beta}_s = \mathbf{0}$) against H_1 ($\boldsymbol{\beta}_s = -\phi_s \boldsymbol{\beta}$) the ϕ_s's are not identifiable under $\boldsymbol{\beta} = \mathbf{0}$, and the usual chi-squared arguments fail. For a single categorical explanatory variable, Haberman (1981) actually gives the asymptotic distribution as following an F-distribution. I am unable to understand why this difficulty does not apply when l_1 is replaced by l_d, i.e. why does the usual argument hold for $2(l_d - l_0)$ but not for $2(l_1 - l_0)$?

Finally, despite the attractiveness of the general stereotype model and the approach suggested by Professor Anderson, I note that the examples are just too simple to be compelling. Others will undoubtably use the model to analyse more complex examples and thus put my reservations to rest, but I wish it were John Anderson who was able to do so.

REFERENCES IN THE DISCUSSION

Adena, M. A. and Wilson, S. R. (1982) *Generalised Linear Models in Epidemiological Research: Case Control Studies*. Sydney: The Intstat Foundation.

Albert, A. and Anderson, J. A. (1981) Probit and logistic discriminant functions. *Commun. Statist.*, **A10**, 641–657.

——— (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, to appear.

Anderson, T. W. (1958) *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bartholomew, D. J. (1980) Factor analysis for categorical data. *J. R. Statist. Soc.* B, **42**, 293–321.

——— (1983) Latent variable models for ordered categorical data. *J. Econometrics*, **22**, 229–243.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models*. London: Chapman and Hall.

Nishisato, S. (1980) *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

Pearson, E. S. and Hartley, H. O. (1972) *Biometrika Tables for Statisticians*, Vol. II. Cambridge: Cambridge University Press.

Scheffé, H. (1952) An analysis of variance for paired comparisons. *J. Amer. Statist. Ass.*, **47**, 381–400.