

Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels

Mark Lunt^{*,†}

*ARC Epidemiology Unit, University of Manchester, Stopford Building, Oxford Road,
Manchester M13 9PT, U.K.*

SUMMARY

There are a number of regression models which are widely used to predict ordinal outcomes. The commonly used models assume that all predictor variables have a similar effect at all levels of the outcome variable. If this is not the case, for example if some variables predict susceptibility to a disease and others predict the severity of the disease, then a more complex model is required. One possibility is the multinomial logistic regression model, which assumes that the predictor variables have different effects at all levels of the outcome variable.

An alternative is to use the stereotype family of regression models. A one-dimensional stereotype model makes the assumption that the effect of each predictor is the same at all outcome levels. However, it is possible to fit stereotype models with more than one dimension, up to a maximum of $\min(k - 1, p)$ where k is the number of outcome categories and p is the number of predictor variables. A stereotype model of this maximum dimension is equivalent to a multinomial logistic regression model, in that it will produce the same predicted values and log-likelihood. If there are sufficient outcome levels and/or predictor variables, there may be a number of stereotype models of differing dimension.

The method is illustrated with an example of prediction of damage to joints in rheumatoid arthritis. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: statistical model; ordinal regression; multi-dimensional

1. INTRODUCTION

There are a number of regression models that can be used to analyse ordinal outcomes [1, 2]. The most common models, the proportional odds (PO) model and the continuation ratio (CR) model, assume that each predictor variable has the same effect, on some scale, at every level of the outcome variable. If this assumption is inappropriate, these models may provide a poor fit to the data, or be misleading as to how individual predictors affect the outcome.

*Correspondence to: Dr. Mark Lunt, ARC Epidemiology Unit, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, U.K.

†E-mail: mark.lunt@man.ac.uk

If a predictor variable is only important across part of the range of outcomes, the PO and CR models cannot hold. For example, a disease outcome may be categorized as none, mild, or severe. It may be that one factor is related to susceptibility to the disease (i.e. the change from 'none' to 'mild' or worse), but other factors determine the severity of the disease (i.e. the change from 'mild' to 'severe'). In this situation, the assumption that the effect of each predictor is the same at all levels of the outcome is clearly untenable.

Such situations require a model in which different combinations of predictor variables are used for distinguishing between different levels of the outcome. Such models were described as multi-dimensional models by Anderson [3]. In such a situation a multinomial logistic (ML) model can be used, which fits a separate linear combination of predictors for each outcome level. However, this model makes no use of the ordinal nature of the outcome: it treats the outcome simply as an unordered categorical variable.

For this reason, Anderson proposed an alternative model, the stereotype (ST) model [3]. This can be thought of as a multinomial logistic model constrained to respect the ordinal nature of the outcome. Anderson gave several examples of the use of the stereotype model, all of which were one dimensional (i.e. the same combination of predictors could be used for all levels of the outcome). Others have also used one-dimensional stereotype models [1, 4–6]), although it has not been used as widely as it might have been due partly to a lack of suitable software for fitting it [2]. There are no examples of the use of multi-dimensional stereotype models in the literature. Such models had been considered, but Anderson [3], reported that he had never found a data set which required their use.

This paper presents an example in which a multi-dimensional stereotype model is the simplest model to fit the data adequately. The software used to fit these multi-dimensional stereotype model using Stata is available from the author.

In what follows, we are concerned with modelling an outcome variable Y , which takes one of the ordered values $1, 2, \dots, k$. There are p variables, x_1, \dots, x_p , which can be used to predict Y .

2. GENERALIZED LINEAR MODELS

Generalized linear models for ordinal outcomes generally involve modelling the cumulative response probability $\gamma_j(x_1, \dots, x_p) = \text{pr}(Y \leq j | x_1, \dots, x_p)$. In this model, a link function $g(\gamma_j)$ is modelled as a linear function of the predictor variables x_1, \dots, x_p , i.e.

$$g(\gamma_j(x_1, \dots, x_p)) = \theta_j + \sum_{i=1}^p \beta_i x_i \quad (1)$$

Different choices for the link function $g()$ give different models. The most common choices for $g()$ are the logit function $g(\gamma) = \log(\gamma/(1 - \gamma))$ (which gives the PO model), or the complementary log log link $g(\gamma) = \log(-\log(1 - \gamma))$ (which gives the CR model).

2.1. Interpretation of coefficients

In both the PO and CR models, e^{β_i} can be interpreted as an odds ratio. In the PO model, e^{β_i} is the proportional change in the odds of being in outcome category j or above when x_i changes by 1. This is assumed to be the same for all values of j . In the CR model, e^{β_i} is the

proportional change in the odds of being in any outcome category greater than j given that the subject is in the category j or above, when x_i changes by 1. Again, this odds ratio is assumed to be the same for all j .

3. THE MULTINOMIAL LOGISTIC REGRESSION MODEL

If we ignored the ordering in Y , we could fit a multinomial logistic (ML) model of the form

$$\text{pr}(Y = s | x_1 \dots x_p) = \frac{\exp(\theta_s + \sum_{i=1}^p x_i \beta_{is})}{\sum_{j=1}^k \exp(\theta_j + \sum_{i=1}^p x_i \beta_{ij})} \quad (2)$$

As it stands, this model is not identified: adding a fixed constant to every β will give exactly the same predicted probabilities. To identify the model, constraints need to be placed on the parameters. Commonly β_{i1} , $0 \leq i \leq p$ are constrained to be 0. In this case, outcome 1 is treated as a baseline group and the other outcome levels compared to it. This is similar to performing $k - 1$ separate logistic regressions, one to distinguish between outcomes 1 and 2, one to distinguish between outcomes 1 and 3, etc. However, whilst the parameter estimates from the two methods are generally very similar, the multinomial logistic regression method is preferable as it is more efficient [7].

4. THE STEREOTYPE MODEL

The multinomial model assumes that different linear combinations of the predictor variables are required to discriminate between the baseline level and each of the other levels of the outcome variable. However, if the outcome is ordinal, rather than categorical, it may be that a single linear combination can discriminate between all levels. If this is the case, the above model can be simplified to

$$\text{pr}(y = s | x_1 \dots x_p) = \frac{\exp(\theta_s + \phi_s \sum_{i=1}^p x_i \beta_i)}{\sum_{j=1}^k \exp(\theta_j + \phi_j \sum_{i=1}^p x_i \beta_i)} \quad (3)$$

This is a one-dimensional stereotype model. In this model, the β parameters no longer differ between the different levels of the outcome. The combination that best discriminates between the levels of the outcome variable is given by $\sum_{i=1}^p x_i \beta_i$ and the distance between the outcome levels in terms of this linear predictor is given by the ϕ_j parameters. If the relationship between the predictors and outcome is ordinal, (i.e. an increase in $\sum_{i=1}^p x_i \beta_i$ increases the probability of being in a higher outcome category rather than a lower one), then we can expect that $\phi_1 \geq \phi_2 \geq \dots \geq \phi_k$, although this is not necessary in order to fit the stereotype model.

Again, constraints are needed to make the model identifiable. Anderson recommended setting $\phi_k = 0$ and $\phi_1 = 1$ [3], but other constraints are possible. Indeed, it is not strictly necessary to estimate the ϕ parameters at all: they may be constrained to take pre-specified values based on expected spacings or other background information. If the predictor variables are all discrete and the scores pre-specified, the stereotype model is equivalent to an ordered loglinear model, commonly used in categorical data analysis [8]. The stereotype model

can be considered as a constrained multinomial model, in which the ratios β_{ij}/β_{ij} , are constant for all variables x_i .

4.1. Dimensionality

Equation (3) assumes that a single linear combination is adequate to distinguish between all k outcome levels, whilst equation (2) assumes that $k - 1$ linear combinations are necessary. It is possible to fit models that use between 2 and $k - 2$ linear combinations: such models require more parameters than equation (3) but fewer than equation (2).

For example, consider an outcome variable with four levels: none, mild, moderate and severe. It may be that one set of variables determine susceptibility to a disease, but different variables are important in determining the severity of the disease. In this case, two linear combinations would be used: one combination to determine the probability of being classed as mild, moderate or severe rather than none, and the other to determine how far along the progression from mild through moderate to severe the subject is likely to have passed. The PO model, by contrast, uses a single combination to determine the progression from none to severe, and hence cannot capture the fact that different variables are important at different stages of the disease. The ML model assumes that the effect of each variable is different at each stage of the disease.

If two linear combinations are used, the model is referred to as a two-dimensional stereotype (ST(2)) model. A two-dimensional stereotype model would have the form

$$\text{pr}(y = s | x_1 \dots x_p) = \frac{\exp(\theta_s + \phi_{s1} \sum_{i=1}^p x_i \beta_{i1} + \phi_{s2} \sum_{i=1}^p x_i \beta_{i2})}{\sum_{j=1}^k \exp(\theta_j + \phi_{j1} \sum_{i=1}^p x_i \beta_{i1} + \phi_{j2} \sum_{i=1}^p x_i \beta_{i2})} \quad (4)$$

The concept of dimensionality is illustrated in Figures 1 and 2. Figure 1 shows three groups of subjects, in which an increase in x is associated with an increased probability of being in a higher outcome category Y , as is an increase in z . The data were simulated, with both x and z being generated as $1.5Y + 2u$, where u followed a uniform distribution on the interval $[0, 1)$. The lines divide the plane up into three regions: a region in which the most likely outcome is $Y = 1$, one in which the most likely outcome is $Y = 2$ and one in which the most likely outcome is $Y = 3$.

In Figure 1(a), the lines separating the groups are based on the proportional odds model. These lines are parallel, having the equations $\beta_1 x + \beta_2 z = c$, for appropriate choices of constant c , where β_1 and β_2 are the parameters from equation (1). The multinomial model, shown in Figure 1(b) does not use parallel lines to distinguish between groups. Outcome $Y = 1$ is the most likely outcome in the region where $((\beta_{01} + \sum_{i=1}^p x_i \beta_{i1}) > (\beta_{02} + \sum_{i=1}^p x_i \beta_{i2}))$ and $((\beta_{01} + \sum_{i=1}^p x_i \beta_{i1}) > (\beta_{03} + \sum_{i=1}^p x_i \beta_{i3}))$, where the β parameters are those from equation (2). Similar pairs of equations define the regions in which $Y = 2$ and $Y = 3$ are the most likely outcomes. In Figure 1(a), both x and z increase with each increase in Y , so parallel lines are a reasonable way to distinguish the groups. The added complexity of the multinomial logistic model does not lead to a great improvement in discrimination.

The data shown in Figure 2, were generated as

$$x = \begin{cases} 1.5 \times u & \text{if } Y = 1 \text{ or } Y = 2 \\ 1.5 \times u + 1 & \text{if } Y = 3 \end{cases}$$

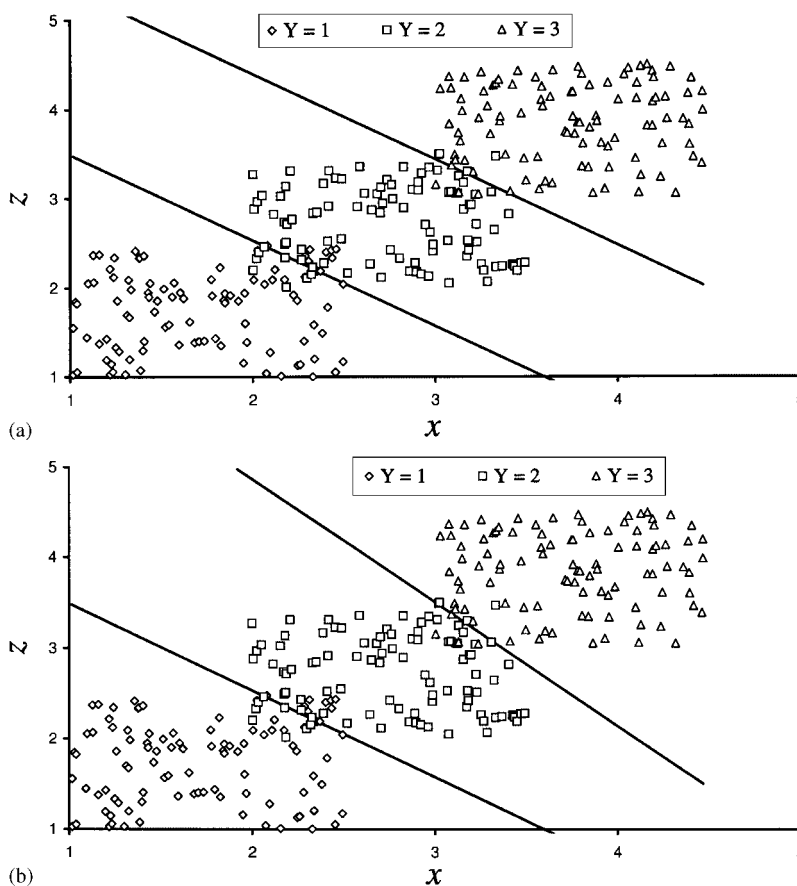


Figure 1. Predicting group membership from ordinal regression models when predictor variables affect all outcome levels: (a) proportional odds model; (b) multinomial logistic regression model.

and

$$z = \begin{cases} 1.5 \times u & \text{if } Y = 1 \\ 1.5 \times u + 1 & \text{if } Y = 2 \text{ or } Y = 3 \end{cases}$$

where u again followed a uniform distribution on the interval $[0, 1)$. In this case, z predicts the progression from $Y = 1$ to 2, while x predicts the progression from $Y = 2$ to 3. Thus x does not differ between groups 1 and 2, whilst z does not differ between groups 2 and 3. Therefore fitting parallel lines to distinguish between the groups is not very successful, as Figure 2(a) shows. Groups 1 and 3 are distinguished reasonably well, but the subjects predicted to be in group 2 include considerable numbers who are in groups 1 and 3.

However, the non-parallel lines fitted by the ML model in Figure 2(b) can distinguish between the groups far better, using a horizontal line to distinguish group 1 from group 2

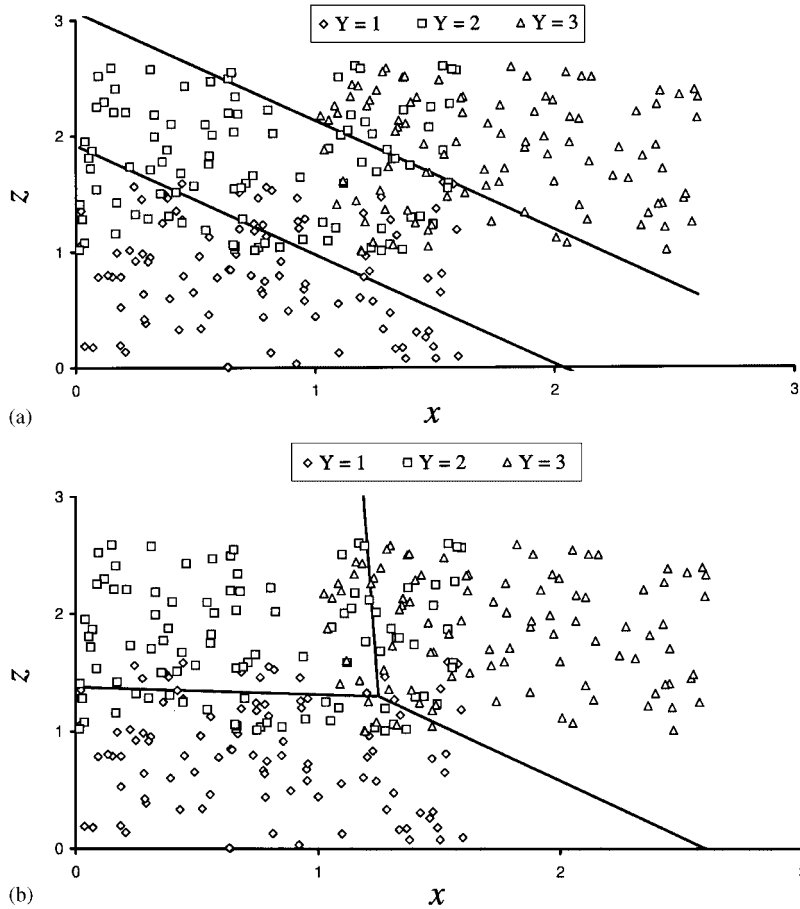


Figure 2. Predicting group membership from ordinal regression models when predictor variables do not affect all outcome levels: (a) proportional odds model; (b) multinomial logistic regression model.

(i.e. z is not used to discriminate between groups 1 and 2) and a vertical line to discriminate between groups 2 and 3 (i.e. x is not used to discriminate between groups 2 and 3). The line distinguishing between groups 1 and 3 is diagonal, since both x and z differ between these groups. This is illustrated in Figure 2(b).

If parallel lines are not an appropriate way to distinguish between the groups, a multi-dimensional model is required. This may be a multinomial logistic model or a (possibly simpler) multi-dimensional stereotype model. In Figure 2, a two-dimensional stereotype model would be exactly the same as the ML model, but in more complex situations the stereotype model could be simpler.

4.1.1. Distinguishability. In equation (3), the β parameters define a linear combination of the x_i which discriminates between the outcome levels, and the ϕ parameters measure the

Table I. Number of parameters in several ordinal regression models with k outcome levels and p predictor variables.

Model	θ	Number of parameters	
		β	ϕ
Null model (intercept only)	$k - 1$		
Proportional odds	$k - 1$	$+p$	
Continuation ratio	$k - 1$	$+p$	
One-dimensional stereotype	$k - 1$	$+p$	$+(k - 2)$
d -dimensional stereotype	$k - 1$	$+dp$	$+d(k - d - 1)$
Multinomial logistic	$k - 1$	$+(k - 1)p$	

distance apart of these groups. It is expected that $\phi_1 \leq \phi_2 \leq \dots \leq \phi_k$. However, it may be that $\phi_j = \phi_{j+1}$, in which case it is not possible to distinguish between outcome levels j and $j + 1$ using the x_i . For example, if we were fitting a stereotype model to the data in Figure 2, using x as the only predictor, we would expect $\phi_1 = \phi_2$.

As part of the process of simplifying the stereotype model, it is possible to constrain ϕ parameters to be equal to each other. The constrained stereotype (CST) model is a more parsimonious model than the full ST model, and may be preferable if the reduction in fit caused by the constraints is not too great. Anderson [3] proposed a stepwise method for determining which outcome levels were distinguishable, i.e. which constraints on the ϕ did not cause an unacceptable reduction in the fit of the model. Such constraints can provide useful etiologic information, since $\phi_j = \phi_{j+1}$ suggests that the predictors do not influence the probability of having outcome $j + 1$ rather than outcome j .

The unconstrained stereotype models always require more parameters than the PO and CR models, although if the scores are constrained to pre-specified values, the number of parameters is the same. The maximum dimension of a stereotype model that can be fitted is either $k - 1$ or p , whichever is the smaller. A stereotype model of this maximum dimension is exactly equivalent to a multinomial logistic regression in terms of predicted probabilities and likelihood, although the actual parameters fitted and their interpretations differ. The number of parameters in each of the models when there are p predictor variables and k outcome levels are given in Table I.

4.2. Fitting the stereotype model

The stereotype model is non-linear in the parameters (since the ϕ and β parameters are multiplied together), and conventional generalized linear model methods cannot be used to estimate it. One suggestion [6] for solving this problem of non-linearity is to use an iterative process: first fix the ϕ parameters and estimate the β s, then fix the β s at these estimated values and re-estimate the ϕ s. This procedure will give unbiased estimates for the parameters, but will underestimate the standard errors, since the standard errors for the ϕ parameters are calculated conditional on the estimated β parameters, and vice versa. Greenland recommended using bootstrapped standard errors to avoid this problem.

However, the approach used here was to linearize the using the method of Box and Tidwell [9] for fitting terms of the form $g(x; \theta)$ where θ is an unknown parameter. An initial value,

θ_0 , is chosen for θ , and $g(x; \theta)$ is approximated by $g(x; \theta_0) + (\theta - \theta_0)[\partial g / \partial \theta]_{\theta=\theta_0}$. In the regression equation, the term $\beta g(x; \theta)$ is replaced by two terms $\beta u + \gamma v$, where $u = g(x; \theta_0)$, $v = [\partial g / \partial \theta]_{\theta=\theta_0}$ and $\gamma = \beta(\theta - \theta_0)$. Having fitted the regression equation to obtain estimates of β and γ , the estimate of θ is updated as $\theta_1 = \theta_0 + \hat{\gamma} / \hat{\beta}$, and the process repeated until it converges. This method has the advantage that not only can the parameters and their standard errors be calculated, but the parameter covariance matrix can also be estimated. However, convergence is not guaranteed if θ_0 is poorly chosen.

5. EXAMPLE

5.1. Data

To illustrate the differences between these models, data from 251 consecutive subjects from the Norfolk Arthritis Register (NOAR) [10] have been used. Subjects were recruited on presentation to a general practitioner with inflammatory polyarthritis, and followed up to determine the progression of disease, in relation to various risk factors recorded at recruitment. Five years after recruitment, all subjects had an X-ray taken to assess the extent of damage to the joints, which is one of the more serious consequences of arthritis. The outcome of interest is the severity of damage in the worst affected joint, measured on a widely used six-point ordinal scale [11]:

1. No damage.
2. Joint space narrowing.
3. Slight evidence of erosion.
4. Clear evidence of erosion.
5. Worse than 4, not as bad as 6.
6. No further damage to joint possible.

Only three risk factors are considered in this analysis: Age at onset of symptoms; rheumatoid factor, classified as present or absent; and shared epitope, a genetic factor of which an individual may have 0, 1 or 2 copies. Age and shared epitope were fitted as continuous variables, with age being measured in decades to get more meaningful parameter estimates. The aim was to see how these three risk factors are associated with the outcome.

The distributions of these variables are given in Table II.

Univariate models: The fit of the different models, measured by the log-likelihood, when rheumatoid factor and shared epitope are used as univariate predictors, are similar (see Table III). The AIC is therefore better for the models with fewer parameters: the additional complexity of the ML and ST models is of no advantage in these cases. In the CST model, rheumatoid factor distinguishes between three groups (1 & 2; 3; 4, 5, & 6), whilst shared epitope only distinguishes between 2 (1 & 2; 3, 4, 5, & 6) (see Table IV).

When age is the predictor, the fit is markedly better using the ST model than the PO or CR models. The AIC is considerably lower for the ML, ST and CST models, suggesting that the assumptions of proportional odds or proportional hazards are not satisfied. In the CST model, age only distinguishes between two groups: no erosions vs some damage. This is illustrated more clearly in Table IV, which shows the β and ϕ parameters from both the ST and CST models, with their confidence intervals. If no confidence interval is given, it is because the parameter was constrained to take a particular value.

Table II. Population characteristics.

Age	Mean (SD)	56 (14)
Rheumatoid factor (<i>n</i>)	Positive	79
	Negative	172
Copies of shared epitope (<i>n</i>)	0	101
	1	113
	2	37
Damage score	1	57
	2	64
	3	58
	4	32
	5	32
	6	8

Table III. Fit of univariate models.

Variable	Model	Deviance	D.F.	AIC
Rheumatoid factor	Null model	832.64	5	842.64
	PO	799.88	6	811.88
	CR	803.18	6	815.18
	ML	799.52	10	819.52
	ST	799.52	10	819.52
	CST	800.75	7	814.75
Shared epitope	PO	821.27	6	833.27
	CR	821.66	6	833.66
	ML	815.23	10	835.23
	ST	815.23	10	835.23
	CST	820.33	6	832.33
Age	PO	808.46	6	820.46
	CR	824.20	6	836.20
	ML	775.86	10	795.86
	ST	775.86	10	795.86
	CST	783.67	6	795.67

Considering rheumatoid factor as the predictor, it is clear from the confidence intervals that in the unconstrained model, ϕ_2 is not significantly different from ϕ_1 , and ϕ_5 and ϕ_4 are not significantly different from ϕ_6 . Applying these constraints leads the constrained model in the following column.

5.1.1. Multivariate models. When all three variables are fitted in a single model, neither the PO nor CR models provide a good fit to the data (see Table V). A one-dimensional stereotype model is again a poorer fit to the data than the ML model (LR $\chi^2 = 33.6$ on 8 d.f., $p < 0.0001$), although a two-dimensional model fits equally well (LR $\chi^2 = 2.6$ on 3 d.f., $p > 0.5$). Constraining this model produces a model that requires only nine parameters, and yet

Table IV. Coefficients and confidence intervals from the unconstrained & constrained univariate stereotype models.

Variable	Parameter	Unconstrained		Constrained	
		Coefficient	Standard error	Coefficient	Standard error
Rheumatoid factor	ϕ_1	0		0	
	ϕ_2	0.15	(-0.25, 0.54)	0	
	ϕ_3	0.47	(0.06, 0.88)	0.49	(0.14, 0.84)
	ϕ_4	0.78	(0.21, 1.35)	1	
	ϕ_5	0.94	(0.29, 1.60)	1	
	ϕ_6	1		1	
	β	2.32	(0.71, 3.93)	1.84	(1.17, 2.51)
Shared epitope	ϕ_1	0		0	
	ϕ_2	-0.16	(-0.64, 0.32)	0	
	ϕ_3	0.28	(-0.15, 0.70)	1	
	ϕ_4	0.65	(0.00, 1.30)	1	
	ϕ_5	0.36	(-0.16, 0.87)	1	
	ϕ_6	1		1	
	β	1.28	(0.18, 2.38)	0.66	(0.28, 1.03)
Age	ϕ_1	0		0	
	ϕ_2	1.51	(-0.07, 3.09)	1	
	ϕ_3	2.06	(-0.10, 4.22)	1	
	ϕ_4	1.43	(-0.12, 2.97)	1	
	ϕ_5	1.22	(-0.12, 2.55)	1	
	ϕ_6	1		1	
	β	0.51	(-0.04, 1.07)	0.79	(0.54, 1.04)

Table V. Fit of multivariate models.

Model	Deviance	D.F.	AIC
Null model	832.64	5	842.64
PO	770.59	8	786.60
CR	787.38	8	803.38
ML	732.41	20	772.41
ST(1)	765.96	12	789.96
ST(2)	735.22	17	769.22
CST(2)	744.30	9	762.30

the difference in fit between this model and the ML model is still not statistically significant (LR $\chi^2 = 11.9$ on 11 d.f., $p = 0.5$). The AIC for the CST(2) is markedly better than for any other model.

The parameter estimates from the ST(2) and CST(2) models are given in Table VI. It can be seen that the parameters ϕ_{21} , ϕ_{31} , ϕ_{41} , ϕ_{51} , and ϕ_{61} in the ST(2) model are not significantly different, and constraining them to be equal does not significantly reduce the fit of the model. Equally, the constraints $\phi_{12} = \phi_{22}$ and $\phi_{42} = \phi_{52} = \phi_{62}$ do not have a detrimental effect on the fit of the model. Thus outcomes 2–6 are indistinguishable with respect to age, and 1 & 2 are indistinguishable with respect to shared epitope and rheumatoid factor, as are outcomes 4–6.

Table VI. Coefficients and confidence intervals from multivariate stereotype models.

Parameter	ST(2)		CST(2)	
	Estimate	Confidence interval	Estimate	Confidence interval
ϕ_{11}	0		0	
ϕ_{21}	1.69	(-0.46, 3.84)	1	
ϕ_{31}	2.27	(-0.61, 5.15)	1	
ϕ_{41}	1.52	(-0.43, 2.46)	1	
ϕ_{51}	1.27	(-0.38, 2.92)	1	
ϕ_{61}	1		1	
$\beta_{\text{age}1}$	0.45	(-0.14, 1.04)	0.77	(0.51, 1.02)
$\beta_{\text{RF}1}$	0		0	
$\beta_{\text{ShE}1}$	0		0	
ϕ_{12}	0		0	
ϕ_{22}	-0.06	(-0.43, 0.32)	0	
ϕ_{32}	0.34	(0.00, 0.680)	0.50	(0.19, 0.80)
ϕ_{42}	0.71	(0.24, 1.18)	1	
ϕ_{52}	0.72	(0.25, 1.18)	1	
ϕ_{62}	1		1	
$\beta_{\text{age}2}$	0		0	
$\beta_{\text{RF}2}$	2.15	(0.74, 3.56)	1.68	(0.99, 2.38)
$\beta_{\text{ShE}2}$	1.07	(0.26, 1.87)	0.83	(0.35, 1.30)

RF=Rheumatoid factor

ShE = Shared epitope

5.2. Inference from stereotype models

Clearly, all three variables are related to the observed severity of erosions, but in different ways. Shared epitope and rheumatoid factor are associated with increased risk of damage at all levels. From the CST(2), we can conclude that the odds of being in outcome groups 4, 5 or 6, rather than groups 1 or 2, increase by a factor of approximately 5.4 ($e^{1.68}$) in the rheumatoid factor positive, with a 95 per cent confidence interval of ($e^{0.99}, e^{2.38}$) = (2.7, 10.8). The odds of being in outcome groups 4, 5 or 6, rather than groups 1 or 2, increase by a factor of approximately 2.3 ($e^{0.83}$) per copy of the shared epitope, with a 95 per cent confidence interval of ($e^{0.51}, e^{1.02}$) = (1.7, 2.8).

The odds of being in outcome group 3, rather than groups 1 or 2, increase by a factor of approximately 2.3 ($e^{(1.68 \times 0.50)}$) in the rheumatoid factor positive, and increase by a factor of approximately 1.5 ($e^{(0.83 \times 0.50)}$) per copy of the shared epitope. Calculating confidence intervals for these odds ratios is less straightforward, since they depend on a non-linear function of the parameters ($\beta \times \phi$). However, an approximation can be calculated using the delta method if the covariance between β and ϕ is known: $\text{var}(\beta\phi) = \hat{\beta}^2 \text{var}(\beta) + 2\hat{\beta}\hat{\phi}\text{cov}(\beta, \phi) + \hat{\phi}^2 \text{var}(\phi)$. For rheumatoid factor, the variance of β was 0.0021, the variance of ϕ was 0.0240 and their covariance was 0.0016, to give a variance of $\beta\phi$ of 0.015, and a 95 per cent confidence interval for $e^{\beta\phi}$ of (1.8, 2.9). The confidence interval for the odds ratio for the shared epitope is (1.0, 2.3) when calculated in the same way.

The effect of age on erosions is different to that of rheumatoid factor and the shared epitope, in that it brings an increased probability of having joint space narrowing (the mildest category of damage), but does not further increase the probability of erosions. The odds of being in

outcome groups 2–6, rather than 1, increase by a factor of approximately 2.2 ($e^{0.77}$) per decade increase in age, with a 95 per cent confidence interval of $(e^{0.51}, e^{1.02}) = (1.7, 2.8)$.

6. DISCUSSION

In a situation in which different variables are important predictors at different levels of an ordinal outcome variable, constrained multi-dimensional stereotype regression provides a well fitting, parsimonious model. In the example given, the CST(2) model with four parameters fitted as well as the ML model with 15, and far better than the PO model with 3.

In addition, the inference from the PO and CST(2) models are different. From the PO model we conclude that age, rheumatoid factor and the shared epitope all lead to a worse outcome. The conclusions concerning rheumatoid factor and the shared epitope from the CST(2) model are the same, but age is shown to only increase the prevalence of joint space narrowing, the mildest form of damage. Among subjects with joint space narrowing, there is no association between age and increasing severity of damage. This additional etiologic information is a major advantage of the stereotype model.

Generalizations of both the PO and CR models have been proposed [2]. These models are similar to the ML model, in that they assume that the predictors have different effects at different outcome levels, and have the same number of parameters. However, they will differ from the ML model slightly due to the different link functions.

It is possible to apply constraints to the generalized PO and CR models [2]. However, where indistinguishability corresponds to a very simple constraint in the stereotype model ($\phi_i = \phi_j$), this is not the case for the generalized PO and CR models, since at least one category in each of these models will consist of several groups combined. Thus selecting appropriate constraints to test for indistinguishability may be more difficult for these models.

We have seen that the stereotype model is more flexible than the PO and CR models commonly used for predicting ordinal outcomes. However, the assumption that the ϕ parameters are ordered implies that $P(Y=r)/P(Y=s)$ increases as the linear predictor increases, if $r > s$. This may be reasonable (for example, if the linear predictor is a prognostic index, one might expect the relative risk of a severe outcome rather than a less severe outcome to increase with the prognostic index), but is not strictly necessary for the expected value of Y to increase as the linear predictor increases. If this assumption is not true (i.e. the $\hat{\phi}$ parameters estimated by the stereotype model are not consistent with the hypothesis that the ϕ parameters are ordered), it may be more appropriate to fit a less restrictive model (such as treating Y as continuous in a generalized linear model).

This paper only considers selecting a regression model based on goodness of fit. It may be that knowledge of the physical process being modelled may suggest one model in preference to another. The parameters from different models have different meanings, and it may be that one particular model provides an estimate of a parameter of particular interest. Some examples of situations in which a particular model may be appropriate or inappropriate are given in Reference [6].

In a previous comparison of the ST(1) model and the PO model [5], it was shown, using simulated data, that the confidence intervals for the model parameters in the ST(1) model were wider than in the PO model, whether or not the PO assumption was true. However, it should be pointed out that the confidence intervals in the PO model are conditional upon the

PO assumption being true. In the stereotype model, there is uncertainty about this assumption, and this additional uncertainty leads to the wider confidence intervals.

Applying constraints to determine which groups are distinguishable is an important part of the modelling process. This is similar in principle to selecting variables to include in a regression model, and similar caveats apply. Several correlated hypothesis tests are involved, so the p -value of the entire process may be different from the p -value of the individual tests. Also, the standard errors of the variables left in the model are calculated on the assumption that the variables included in the model are selected randomly, which is not true (at each step, the least significant variable is selected for removal). Thus the true significance levels of the remaining parameters will not correspond to their nominal values, and may be much greater. Great caution must therefore be exercised when interpreting the results of a constrained model (when the constraints are data-driven, rather than pre-determined). However, this is true whenever variables are selected as part of the modelling process.

The selection of constraints to apply to a multidimensional stereotype model requires some thought. The univariate models will provide some insight: in the erosions example, age appeared to be affecting the outcome differently from rheumatoid factor and shared epitope in the univariate analyses. This suggested using age for one dimension of the ST model, and rheumatoid factor and shared epitope for the other. Indeed, it is only if different variables are important predictors a different levels of the outcome that a multi-dimensional model is appropriate, and this should be apparent in the univariate analyses.

In deciding which groups are indistinguishable, a process similar to backward elimination of variables in regression can be used. If there are k outcome levels, there are $k - 1$ possible constraints of the form $\phi_i = \phi_{i-1}$. Each of these can be tested and the one that causes the least change in the log-likelihood of the model adopted, provided the change in likelihood is less than some predetermined level. Then the remaining $k - 2$ constraints can be applied to this constrained model and again the one causing the least change in the log-likelihood adopted.

REFERENCES

1. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* 1997; **26**(6):1323–1333.
2. Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* 1989; **129**(1):191–204.
3. Anderson JA. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B* 1984; **46**(1):1–30.
4. Greenland S. An application of logistic models to the analysis of ordinal responses. *Biometrical Journal* 1985; **2**:189–197.
5. Holtbrugge W, Schumacher M. A comparison of regression models for the analysis of ordered categorical data. *Applied Statistics* 1991; **40**(2):249–259.
6. Greenland S. Alternative models for ordinal logistic regression. *Statistics in Medicine* 1994; **13**:1665–1677.
7. Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 1984; **71**(1):11–18.
8. Agresti A. *Categorical Data Analysis*, Chapter 8. Wiley: New York, 1990.
9. Box G, Tidwell P. Transformation of the independent variables. *Technometrics* 1962; **4**:531–550.
10. Symmons D, Barrett E, Bankhead C, Scott D, Silman A. The incidence of rheumatoid arthritis in the United Kingdom: results for the Norfolk arthritis register. *British Journal of Rheumatology* 1994; **33**:735–739.
11. Larsen A, Dale K, Eek M. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. *Acta Radiologica* 1977; **18**:481–491.