STATISTICS IN MEDICINE, VOL. 8, 785–794 (1989)

REGRESSION WITH AN ORDERED CATEGORICAL RESPONSE

T. J. HASTIE

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974, U.S.A.

J. L. BOTHA

Department of Community Health, University of Leicester, P.O. Box 65, Leicester LE2 7LX, U.K.

AND

C. M. SCHNITZLER

MRC Bone and Joint Research Unit, Department of Orthopaedic Surgery, University of the Witwatersrand and Johannesburg Hospital, Johannesburg, South Africa

SUMMARY

A survey on Mseleni joint disease in South Africa involved the scoring of pelvic X-rays of women to measure osteoporosis. The scores were ordinal by construction and ranged from 0 to 12. It is standard practice to use ordinary regression techniques with an ordinal response that has that many categories. We give evidence for these data that the constraints on the response result in a misleading regression analysis. McCullagh's[11] proportional-odds model is designed specifically for the regression analysis of ordinal data. We demonstrate the technique on these data, and show how it fills the gap between *ordinary regression* and *logistic regression* (for discrete data with two categories). In addition, we demonstrate non-parametric versions of these models that do not make any linearity assumptions about the regression function.

KEY WORDS Proportional odds Ordered categorical Non-parametric

1. INTRODUCTION

Mseleni joint disease, a crippling polyarthritic disease of unknown etiology, is localized to a small area in Northern Kwazulu, South Africa, and is particularly prevalent in women.[1–3] A recent suggestion is that Mseleni joint disease may not be a single entity, but may consist of more than one condition.[4,5] To investigate this hypothesis, X-rays of 273 women from the Mseleni area (X-rayed in earlier surveys) were reviewed, and the opportunity was used to screen pelvic X-rays for the presence or absence of osteoporosis, a demineralizing bone disease. The score (OP) is constructed as a sum of osteoporosis grades (0–3) for the sacrum, ilium, pubis and ischium with a minimum of 0 and a maximum of 12. Apart from comparisons with other communities in South Africa, data analysis focused on a possible difference in OP between women with and without osteoarthrosis (OA). The confounding effect of AGE weakens a direct comparison of OP between the two groups, since both OA and OP have positive associations with AGE. This sets the scene for an analysis that adjusts for age.

Table I presents a first attempt at an analysis by grouping the data into age categories. The two columns show the mean OP score in age categories for women with OA (OA positive) and women

0277–6715/89/070785–10$05.00

Received November 1986
Revised November 1988

Table I. Mean OP scores of AGE and OA classification. The number of subjects is given in parenthesis. For each age category, P-value corresponding to the Kruskal–Wallis two sample rank test is given

AGE	Mean OP score OA positive	OA negative	P-value (Kruskal–Wallis)
11–20	2·83 (6)	0·53 (139)	0·091
21–30	3·67 (3)	1·00 (13)	0·194
31–40	3·57 (14)	1·52 (27)	0·177
41–50	6·77 (30)	2·16 (31)	<0·0001
51–60	8·67 (6)	3·00 (4)	0·067

without OA (OA negative), as well as the number of subjects in each category. The third column gives p-values for the age specific Kruskal–Wallis tests between the OA positive and negative groups. Only the 41–50 year category shows a significant difference at the 5 per cent level. However, the mean OP scores in the OA positive groups are consistently higher than in the OA negative group. Has this grouped analysis helped us, or is it possible that the grouping entailed some loss of information?

We find that the grouped analysis has several deficiencies:

1. the mean values and the tests to compare them depend on the choice of age categories;
2. even if we choose the categories well, we can lose information by averaging within a category;
3. it does not provide a summary measure of the group separation.

There are probably a number of alternative grouped analysis techniques that would address problem 3, such as the Friedman test,[6] or a weighted average of differences. Rather than dwell on these, we look at some methods that address all three.

A natural candidate is linear regression, usually referred to as analysis of covariance in this context, which exploits the continuity of age when making the age adjustment. We will see, however, that an analysis of covariance may give misleading results due to the categorical nature of the response.

We have several goals in this paper:

1. to present smoothing as a non-parametric alternative to ordinary linear regression, and thereby deduce the correct form for subsequent parametric regressions;
2. to highlight possible problems in the analysis of ordinal data with conventional regression techniques;
3. to demonstrate the proportional-odds model as a solution to these problems;
4. to describe briefly a non-parametric version of this proportional-odds model.

2. METHODS AND RESULTS

2.1. Linear regression models

Figure 1 displays the results of several regression analyses on the observed scores (● denotes OA positive or OA+, ○ denotes OA negative or OA−). The most general model, having separate slopes and intercepts, has the form

$$\mathrm{OP} = \alpha_{\mathrm{OA}} + \beta_{\mathrm{OA}}\,\mathrm{AGE} + \mathrm{error}, \quad \mathrm{OA} = + \text{ or } -, \tag{1}$$

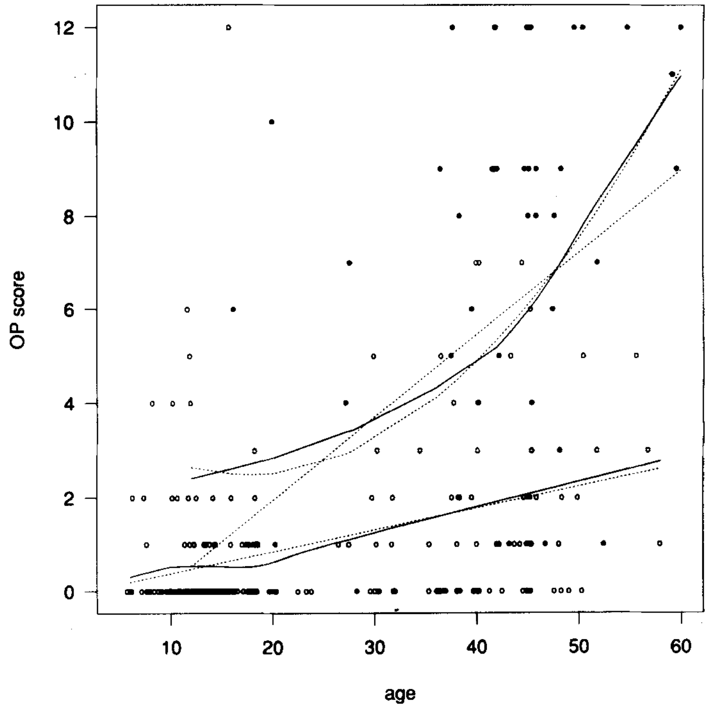

Figure 1. Regressions of OP score against AGE for OA positive (●) and OA negative (○) groups
The solid curves represent the non-parametric fit, model (4) in text. The dashed lines represent the separate slopes regression model (1), and the dashed curve the quadratic model for the OA positive group. There is a noticeable bunching of the OP scores in the lower left corner

which allows a different level for OA+ and −, as well as a different age effect. The alternative of interest is

$$OP = \alpha_{OA} + \beta\ AGE + error, \quad OA = +\text{ or } -, \tag{2}$$

the parallel slope model, which claims the same age effect in each group. Other more restrictive alternatives allow no group effect at all (same intercept, same slope), and no age effect (same or different intercept, no β).

Table II summarizes the results of fitting all these models by least squares. We see in particular that the effect of separate slopes over parallel lines is significant and indicates a different age effect for the two groups. Figure 1 also shows the fit of the model

$$\begin{aligned} OP &= \alpha_{OA+} + \beta_{OA+}\ AGE + \gamma_{OA+}\ AGE^2 + error \\ OP &= \alpha_{OA-} + \beta_{OA-}\ AGE + error, \end{aligned} \tag{3}$$

and Table II shows the quadratic effect is significant. The quadratic effect was also suggested by the non-parametric smoothing technique described in the next section, although admittedly one would likely fit a quadratic term as the first try upon suspicion of non-linearity.

The linear model (quadratic in fact) has become rather complicated; at least its message about the effect of age is. Looking at Figure 1, we might feel suspicious about the bunching of the zeros in the younger age groups. The question that begs itself is whether the X-ray rating of *zero* is too crude to describe the apparent continuous relationship we see in the figure; would the age effect for

Table II. ANOVA table for the regression models in Section 2. All F tests are based on Gaussian error assumption. Test 1–2 is more of an approximation than the others since for these non-parametric models: (1) RSS's are only approximately chi-squared distributed; (2) models are not strictly nested, and (3) RSS's are not strictly independent. Test 1–2 uses 2 as error, 3–4, 4–5, and 5–6 uses 6 as error term in F test

	Model	RSS	Residual d.f.	Effect RSS	Effect d.f.	F ratio
1	Non-parametric parallel curves	1 721·5	268·1			
2	Non-parametric separate curves	1 658·5	265·2			
1–2	Effect of separate versus parallel curves			63·0	2·9	3·5*
3	Linear – no OA effect	2 202·3				
4	Linear – parallel slopes	1 779·0	270·0			
3–4	Intercept effect (OA)			223·3	1·0	36·2†
5	Linear – separate slopes	1 686·9	269·0			
4–5	Effect of separate slopes			92·1	1·0	14·9†
6	OA+ quadratic OA – linear	1 655·0	268·0			
5–6	Effect of OA+ quadratic			31·9	1·0	5·2*

* $p<0{\cdot}05$
† $p<0{\cdot}01$

the OA-group have continued below the 0 line and parallel to the line for OA+? To investigate this question, we will need to transform the response scale in some way. Before we do this, however, let us have a look at the smoothing techniques mentioned above.

2.2. Smoothing and additive models

Although convenient, the linear model can be very restrictive as a means for modelling the effect of continuous scale variables such as age. A more general version of (1) is

$$\mathrm{OP} = \alpha_{\mathrm{OA}} + f_{\mathrm{OA}}(\mathrm{AGE}) + \mathrm{error}, \quad \mathrm{OA} = + \text{ or } -, \tag{4}$$

where f is some arbitrary (possibly smooth) function of AGE, with the subscript denoting a different function for each group. Since the models are completely separate in the two groups, let us see how to fit the model in either one, and thus drop the subscript OA.

If the function f is completely arbitrary, it seems we could fit each OP value exactly without error, so something is wrong. To estimate the function we need to assume that it is smooth in some sense, so that locally it is roughly constant. This motivates the k nearest neighbour smoother which would estimate the function at each AGE value by the average OP score for the k points in the sample closest in AGE. We can do this efficiently by moving a window from left to right, and updating the average as points enter and leave.

This is a very simple smoother; a variety of more sophisticated smoothers exist, all with the same goal. For example, instead of using nearest neighbours, kernel smoothers give points weights that die down smoothly with their distances in AGE from the target AGE_0, and then compute a weighted average of the response to obtain the fit at AGE_0. One does this at all values AGE_0 of interest. Other smoothers use the fitted values from straight lines in the k nearest neighbourhoods, or even locally weighted straight lines. Smoothing splines do not explicitly use neighbourhoods; they simply impose a restriction on a global measure of smoothness of the function, such as the integrated second squared derivative. There is already a large literature on smoothers; see Cleveland[7] for a description of the locally weighted running line smoother, and Silverman[8] for a review of spline smoothing techniques.

The two solid curves in Figure 1 are the estimates of the functions in (4) obtained by separately smoothing the OP scores against AGE for the OA+ and OA− groups. They are well approximated by the quadratic/linear model in Section 2.1. We used a locally weighted running line smoother with symmetric near-neighbourhoods, with a span of 50 per cent of the data in any given central neighbourhood (and because of the symmetric neighbourhoods, these shrink down to 25 per cent at the ends). Although our choice of span seems arbitrary, we have had empirical success with 50 per cent spans. Fixed span smoothers allow us to compute approximate 'degrees of freedom' for the non-parametric fits, which are used in Table II. A span of 50 per cent corresponds to approximately 3 to 4 d.f.[12]

It is also possible to fit non-parametric analogues of the alternative linear models described above, the most interesting being the model with parallel curves but different intercepts.[9,10] Approximate 'F' tests have been developed to compare the different non-parametric models as well, and are included in Table II; the details of their derivation, however, are beyond the scope of this article. To summarize the results of this and the last section, we conclude that either the linear/quadratic model, or equivalently the separate curve non-parametric model is appropriate if we model the scores themselves.

2.3. The proportional-odds model

We now address the question of whether the nature of the response variable might be responsible for the non-linear regression. At least in the grouped analysis (Table I) we used a non-parametric test, although a separate one in each age category. We require an analysis that enjoys the advantages of both the grouped analysis, as well as regression methods with their ability to incorporate effects.

Many models have been proposed to extend the logistic regression model for binary data to categorical data with more than two outcome categories. One needs special care with ordered categories. The proportional-odds model is one such extension, although in our mind ideally suited to these data.

The most appealing motivation for the proportional-odds model is in terms of a latent (and usually unobservable) continuous response variable.[11,12] Specifically we assume that the observed data are a categorization of the underlying continuous variable. In this case it would seem that many of the zeros in the OA negative group are truncated low OP scores. We then model the location parameter of this underlying response as a function of covariates. This has the same flavour as ordinary regression, and will allow us to summarize the age adjusted differences in a similar way. The model makes very mild distributional assumptions; we assume the observed counts have a multinomial distribution over the 13 categories.

Suppose the latent variable Z has a distribution $F_\eta(z) = F(z-\eta)$ where η is a location parameter. Given a vector of covariates $\mathbf{x}$, we customarily model $\eta(\mathbf{x}) = \boldsymbol{\beta}^t\mathbf{x}$, and thus the conditional distribution of $Z|\mathbf{x}$ by $F(z-\boldsymbol{\beta}^t\mathbf{x})$. We do not get to see realizations of Z; rather we see where Z lies in the categorization $(-\infty, \alpha_1], (\alpha_1, \alpha_2], \ldots, (\alpha_{k-2}, \alpha_{k-1}], (\alpha_{k-1}, \infty)$ of $\mathbb{R}$,[1] for some unknown α_j. This induces the observed random variable Y, where $Y=k$ if and only if $Z \in (\alpha_{k-1}, \alpha_k]$, and thus we can define the cumulative probabilities $\gamma_k(\mathbf{x}) = P(Y \leqslant k|\mathbf{x}) = F(\alpha_k - \boldsymbol{\beta}\mathbf{x})$. The choice of F is not crucial to the model; for ease of computation we use the logistic distribution $F(z) = e^z/(1+e^z)$ and

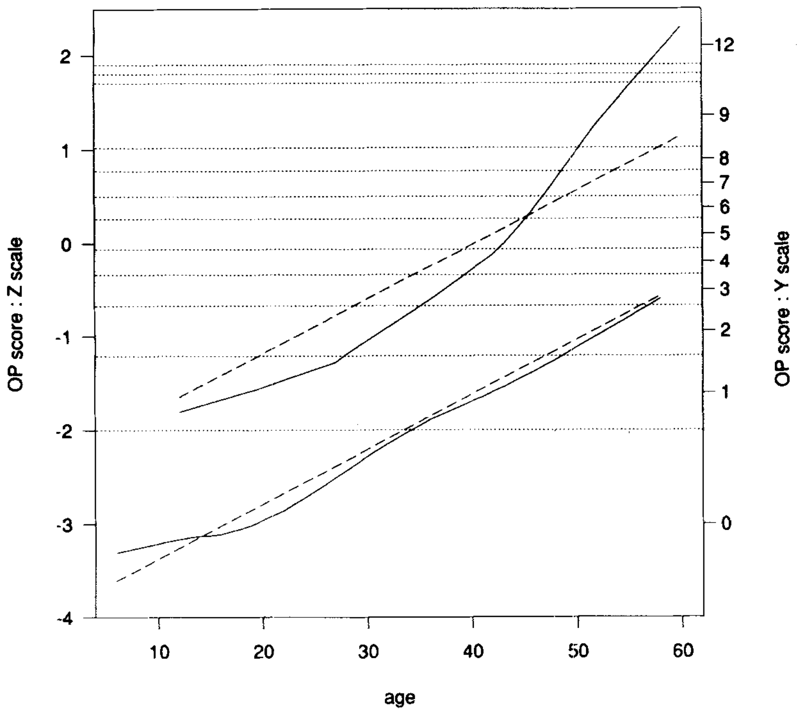

Figure 2. The proportional-odds model
The curves are plotted on the latent scale (Z or underlying continuous variate). The parallel dashed lines represent the final parallel slopes linear model; the solid non-linear functions represent the separate curve non-parametric model. These curves can be interpreted as the regression of Z against AGE. The dashed lines are drawn at the values $\hat{\alpha}_k$; each region between dashed lines represents the range of Z that corresponds to a given category of Y. The Y-scale is labelled on the right vertical axis

get the model:

$$\text{logit}\,(\gamma_k(\mathbf{x})) = \alpha_k - \boldsymbol{\beta}^t\mathbf{x}, \tag{5}$$

where $\text{logit}\,(\gamma_k(\mathbf{x})) = \log\,[\gamma_k(\mathbf{x})/(1-\gamma_k(\mathbf{x}))]$. In our application $Y = \text{OP}$, the score ranging between 0 and 12. We might think of the latent variable Z as the underlying 'continuous' measurement of the state of osteoporosis in the patient, measured on a suitable scale.

We make the following observations:

1. The odds ratio $\gamma_k(\mathbf{x}_1)(1-\gamma_k(\mathbf{x}_2))/(1-\gamma_k(\mathbf{x}_1))\gamma_k(\mathbf{x}_2) = \exp\,(\boldsymbol{\beta}^t(\mathbf{x}_2 - \mathbf{x}_1))$ is independent of k, and hence the name proportional odds. This gives the relative odds of having a score of k or less for two different values of the covariates, and has the same flavour as the proportional hazards model of Cox.[13]
2. The model specifies that the entire distribution of Z shifts linearly with $\mathbf{x}$. This in turn implies a smooth shift in the histogram of Y for changes in $\mathbf{x}$. In Figure 2 we see the model fitted to the OP data. The parallel lines show how the estimated median (or any quantile) of Z changes with age in the two OA groups.
3. The logistic assumption is not crucial; in fact any symmetric bell shaped distribution would give very similar results. In practice we can also use the Gaussian, as in Probit analysis, although the logistic is computationally more attractive. If we have reason to believe that the

Table III. ANOVA (analysis of deviance) table for the proportional-odds regression models

	Model	Deviance	D.F.	Effect deviance	Effect d.f.
1	Non-parametric parallel curves	788·0	257·1		
2	Non-parametric separate curves	786·3	254·2		
1–2	Effect of separate versus parallel curves			1·7	2·9
3	Linear – no OA effect	814·4	260·0		
4	Linear – parallel slopes	789·2	259		
3–4	Intercept effect (OA)			25·2*	1·0
5	Linear – separate slopes	786·5	258·0		
4–5	Effect of separate slopes			2·7	1·0
6	OA+ quadratic, OA− linear	786·2	257·0		
5–6	Effects of OA+ quadratic			0·3	1·0
7	Separate linear versus separate curves			0·2	3·8

* $p<0{\cdot}05$

underlying distribution is asymmetric, we can use alternative links such as loglog or complementary loglog.

For our data the linear model is

$$\text{logit}(\gamma_k)=\alpha_k-\kappa_{\text{OA}}-\beta_{\text{OA}}\text{AGE},\ \text{OA}=+,\ - \qquad (6)$$

and as before, we will test if we need separate slopes, and thus if we can replace β_{OA} with β.

We estimate the model (6) by maximum likelihood in the multinomial family. As in the case of logistic regression, the algorithm is iterative although somewhat more complicated. Hastie and Tibshirani[12] describe the algorithm in detail, and represent it in an intuitive form that also allows easier programming. They deal with both grouped as well as ungrouped data. Their algorithm iterates two steps, which we summarize briefly.

At any stage of the algorithm, current estimates of the α_k's and $\boldsymbol{\beta}$ allow us to compute the fitted cumulative proportions γ_k for each observation (and hence at each of the observed covariate vectors). We then compute a *vector* of $K-1$ standardized residuals per observation (as opposed to a single residual per observation in binary logistic regression). The two steps are then:

1. We update α_k by computing a weighted mean vector of these residual vectors. The weights (matrices) depend on the multinomial distribution as well as the derivative of the link function.
2. We update the regression coefficient by:
   (a) computing a single residual per observation. This is a weighted average of the vector of $K-1$ residuals that uses the row sums of the weight matrices as weights;
   (b) computing a single weight per observation, which is the sum of all the weights in the weight matrix;
   (c) performing a weighted linear regression of this residual onto the covariates.

This algorithm converges to the maximum-likelihood solution.

As is usual with maximum-likelihood methods, we can base tests of effect between various nested models on the likelihood ratio test statistic or *deviances*. We compare two models by computing the difference in their deviance scores, the difference in their degrees of freedom (d.f.),

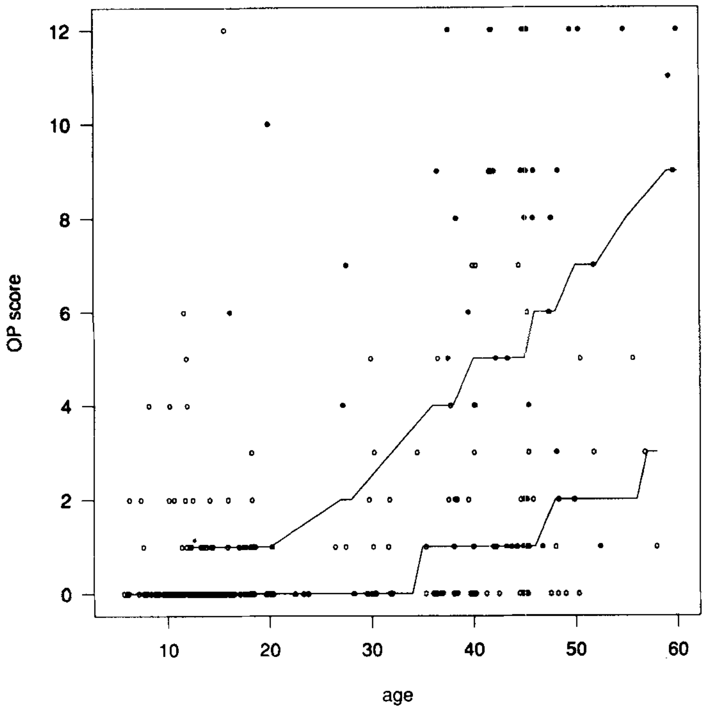

Figure 3. The fitted proportional-odds model on the Y-scale
Points on the curves represent the OP score corresponding to the fitted values from the *parallel linear* regression model. The categorization (or parallel fitted Z-values) causes the curves to bunch at the low AGE values, just as in Figure 1

and reference to the appropriate chi-square p-value. Further details, beyond the scope of this article, appear elsewhere.[11,12,14] Several standard statistical software packages have capabilities for fitting models of this kind.

Figure 2 and Table III summarize the fit of these models to the OP data. The test for separate curves (and later for separate slopes) is not significant. The constant OA effect (κ_{OA}) is highly significant. As is the case for logistic regression analysis on binary data, we have no goodness of fit statistic for this model if we do not group the data; we can only compare effects between models.

The dashed horizontal lines in the plot correspond to the fitted constants $\hat{\alpha}_k$, and hence the regions between these lines correspond to categories (labelled on the right of the plot). We can interpret the fitted regression line in two ways. On the Z-scale, we see how the median (or any quantile) of the underlying variable changes with the covariates. On the Y-scale, the 'fitted median category' for an observation is that category whose region contains the fitted values.

Although the Z-score generated by the proportional-odds regression (Figure 2) appears to allocate negative scores for osteoporosis, it is an estimate in standardized form and so the location and scale are irrelevant.

Figure 3 shows the fitted median categories for this parallel slopes linear model plotted against AGE, and we see the separating curve effect. It seems evident from Figure 2 that the truncation at category 0 and the squashing up of the transformed scores at the high categories indeed account for the apparent different curves seen in the analysis in the previous section and in Figure 3.

Although not shown here, we then used the proportional-odds model further to compare the OP scores between two geographical AREAS: Mseleni and Manguzi (a neighbourhood area with lower prevalence rates). We found that a parallel slope model was not significant (a deviance of 1·2 for 1 d.f.). This was a relatively 'clean' analysis by virtue of the parallel slopes. This means that we can adjust the scores for AGE and OA, and then compare the adjusted scores between AREAS.

3. THE ADDITIVE PROPORTIONAL-ODDS MODEL

Figure 2 also contains two non-parametric curves, which we fitted first in an exploratory fashion.

We calculated these with the use of a generalization of the proportional odds model (6). In the same spirit as the generalization from (1) to (4), we did this by replacing the linear component in (6) by a non-parametric one:

$$\operatorname{logit}(\gamma_k) = \alpha_k - \kappa_{OA} - f_{OA}(\text{AGE}),\ \text{OA} = +,\ -. \tag{7}$$

The estimation of (7) is also based on likelihood principles, and involves a suitable generalization of the techniques used in the previous section. Instead of a weighted linear regression in step 2 of the iterative pair, we perform a weighted additive regression. The details, even further beyond the scope of this paper, appear in Reference 11.

The fitted non-parametric curves do seem to suggest that the OA positive group increases slightly faster than the OA negative group. We can perform crude deviance tests that do not support this (see Table III). We summarize all the models tested in Table III. None of the fitted models performs significantly better than the parallel slopes linear model, for which the constant OA effect is highly significant.

4. DISCUSSION

Epidemiologic studies often involve the analysis of discrete variables. A wealth of analysis techniques is available for a dichotomous (0–1) response. In the regression context, standardization techniques and logistic regression are popular. When the response has more than two *ordered* categories, the common approach is to resort to the usual interval scale techniques, such as linear regression. We have presented an analysis that illustrates an alternative regression technique: the proportional-odds model. We have used the model on data with ungrouped covariates; one can similarly analyse data in the form of contingency tables, where one of the classifications (the response) has ordered categories.

For the example presented, we feel that the ordinary regression results mislead; we attribute the cause to the discrete nature of the data. In particular, the scores appear hemmed in at zero. They may all be recorded as zero because the X-ray screening method lacks sufficient sensitivity to differentiate among them. This suggests a badly categorized underlying 'continuous' variable, and hence the proportional-odds model seems appropriate.

If we were to view the response Y as continuous, we would conclude that the error structure is different at the low ages. A casual glance suggests the variance is increasing with AGE, and hence ordinary linear regression is not appropriate. One approach in situations such as these is to transform the Y values to force the error structure to be approximately independent of AGE; a log transform might do it here. From then on one fits models and makes interpretations on the transformed scale. There is an art in picking the appropriate transformation. The technique of Box and Cox[15] allows one to pick automatically an appropriate transformation, and removes some of the subjectivity. One can also view the proportional-odds model as a form of transformation model for categorical data, only more general. It has a similar flavour to the Box–Cox method. We estimate simultaneously the linear regression and the category cutpoints that define intervals for the underlying response Z. Rather than transform the discrete values of Y to another set of numbers, we transform them to *intervals*.

In contrast to the linear regression techniques, the proportional-odds regression indicates that the OA difference in OP scores is constant over age and significantly non-zero. This means that the age-related deterioration in osteoporosis occurs at the same rate in individuals with or without osteoarthrosis, although at different levels.

REFERENCES

1. Wittman, W. and Fellingham, S. A. 'Unusual hip disease in remote part of Zululand', *Lancet*, **i**, 842–843 (1970).
2. Lockitch, G. and Fellingham, S. A. 'Mseleni Joint Disease: The pilot clinical survey', *South African Medical Journal*, **47**, 2283–2293 (1973).
3. Yach, D. and Botha, J. L. 'Mseleni Joint Disease in 1981: Decreased prevalence rates, wider geographical location than before, and socio-economic impact of an endemic osteoarthrosis in an underdeveloped community in South Africa', *International Journal of Epidemiology*, **14**, 276–284 (1985).
4. Solomon, L., McLaren, P., Irwig, L., Gear, J. S. S., Schnitzler, C. M., Gear, A. and Mann, D. 'Distinct types of hip disorder in Mseleni Joint Disease', *South African Medical Journal*, **69**, 15–17 (1986).
5. Schnitzler, C. M., Solomon, L., Botha, J. L. and McLaren, P. 'Pelvic osteopenia associated with Mseleni Joint Disease. Radiological aspects', *South African Medical Journal*, **72**, 473–475 (1987).
6. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, Wiley, New York, 1981.
7. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).
8. Silverman, B. W. 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting. (with discussion)', *Journal of the Royal Statistical Society, Series B*, **47**, 1–52 (1985).
9. Hastie, T. and Tibshirani, R. 'Generalized additive models; some applications', *Journal of the American Statistical Association*, **82**, No. 398, 371–386, 66–81 (1987).
10. Hastie, T. and Tibshirani, R. 'Generalized additive models (with discussion)', *Statistical Science*, **1**, 297–318 (1986).
11. McCullagh, P. 'Regression models for ordinal data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
12. Hastie, T. and Tibshirani, R. 'Non-parametric logistic and proportional odds regression', *Applied Statistics*, **36**, 260–276 (1987).
13. Cox, D. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **74**, 187–220 (1972).
14. Thompson, R. and Baker, R. 'Composite link functions in generalized linear models', *Applied Statistics*, **30**, 125–131 (1981).
15. Box, G. E. P. and Cox, D. R. 'An analysis of transformations', *Journal of the Royal Statistical Society, Series B*, **26**, 211–252 (1964).