Ordinal Regression Models for Epidemiologic Data

Article *in* American Journal of Epidemiology · February 1989

DOI: 10.1093/oxfordjournals.aje.a115109 · Source: PubMed

2 authors, including:

Benedict Armstrong
London School of Hygiene and Tropical Medicine
466 PUBLICATIONS **35,780** CITATIONS

SEE PROFILE

AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 129, No. 1

Printed in U.S.A

ORDINAL REGRESSION MODELS FOR EPIDEMIOLOGIC DATA

BEN G. ARMSTRONG AND MARGARET SLOAN

Armstrong, B. G., and M. Sloan (School of Occupational Health, McGill U., Montreal, Quebec, Canada H3A 1A3). Ordinal models for epidemiologic data. *Am J Epidemiol* 1989;129:191–204.

Health status is often measured in epidemiologic studies on an ordinal scale, but data of this type are generally reduced for analysis to a single dichotomy. Several statistical models have been developed to make full use of information in ordinal response data, but have not been much used in analyzing epidemiologic studies. The authors discuss two of these statistical models—the cumulative odds model and the continuation ratio model. They may be interpreted in terms of odds ratios, can account for confounding variables, have clear and testable assumptions, and have parameters that may be estimated and hypotheses that may be tested using available statistical packages. However, calculations of asymptotic relative efficiency and results of simulations showed that simple logistic regression applied to dichotomized responses can in some realistic situations have more than 75% of the efficiency of ordinal regression models, but only if the ordinal scale is collapsed into a dichotomy close to the optimal point. The application of the proposed models to data from a study of chest x-rays of workers exposed to mineral fibers confirmed that they are easy to use and interpret, but gave results quite similar to those obtained using simple logistic regression after dichotomizing outcome in the conventional way.

asbestosis; regression analysis; statistics

Health status is often measured in epidemiologic studies on an ordinal scale. For example, extent of symptoms of physical disease may be measured on a scale such as "absent" to "mild" to "severe", and psychometric evaluations are frequently ordinal. In a particular example pursued further in this article, chest x-rays of men exposed to dust are commonly read for small opacities on a four- or 12-point ordered scale. Data of these type are sometimes analyzed as numerical scores, but such an approach is only strictly valid if intervals between consecutive points on the scale can be considered equivalent. To avoid this assumption, the scales are often dichotomized and analyzed using standard techniques for binary data. Although valid, this approach loses information by collapsing some categories of the original scale. Statistical methods which respect the ordinal nature of this kind of response data have been developed, and have recently received considerable attention in the statistical journals. Although some of these articles have used epidemiologic data to illustrate methods, the methods do not appear to have been widely used in substantive reports of epidemiologic studies. The purpose of the present article is to provide an introduction to these techniques which is accessible to epidemiologists, and to evaluate their usefulness for epidemiologic research.

Received for publication August 6, 1987, and in final form June 9, 1988.

From the School of Occupational Health, 1130 Pine Avenue West, Montreal, Quebec, Canada H3A 1A3. (Send reprint requests to Dr. Ben G. Armstrong at this address.)

Research supported by the Institut de recherche en santé et en securité du travail du Québec, through its funding of the Equipe Associée de Recherche en épidémiologie des lesions professionelles.

THE CUMULATIVE ODDS AND CONTINUATION RATIO MODELS

Statistical models of the dependence of an ordinal variable on one or more explanatory variables are termed "ordinal regression models" (1). Two of these statistical models, the cumulative odds model and the continuation ratio model, will be discussed in detail here. They will be illustrated first by the data presented in table 1 from a hypothetical survey of symptoms measured on a three category ordinal scale (none, mild, severe) in two groups.

Following the notation of McCullagh and Nelder (2), the integers 1,....,k act as labels for the k ordered response categories; and $\pi_j, j = 1,....,k$ are the multinomial probabilities of being in each category. In all ordinal regression models, the π_j depend on the values of a vector of explanatory variables x through regression parameters.

The cumulative odds model

Walker and Duncan (3), in a natural extension of the logistic model for binary response data, proposed the model:

$$\begin{aligned} \text{logit}(\gamma_j) &= \ln(\gamma_j/(1-\gamma_j)) \\ &= \theta_j - \beta^T x, \quad j = 1....k-1 \end{aligned} \tag{1}$$

where $\gamma_j = \pi_1 + ... + \pi_j$ are cumulative probabilities of being in one of the first j categories, $\ln(\cdot)$ is the natural logarithmic function, and θ_j and β are unknown parameters. McCullagh (1) considers this model in greater detail, calling it the "proportional odds" model. The parameters θ_j represent the baseline logits of cumulative response probabilities in a person for whom $x = 0$, and β represents the "regression" parameter through which the effect of the explanatory variables is mediated.

The cumulative odds model is sometimes described as "grouped-continuous", since it may be derived by assuming an unobserved underlying continuous response variable, the observed ordinal response variable being formed by taking contiguous intervals of the continuous scale, with cut-points unknown. This interpretation motivated the development of these models, and may be useful in some applications, but is not necessary. McCullagh (1) describes some further generalizations within the grouped continuous class of models.

Model 1 may be motivated from our illustrative data. The table may be simplified to a 2 × 2 table by dichotomizing the response scale either by collapsing levels 2 and 3, or levels 1 and 2 (these constituent tables are also shown in table 1). The former gives an estimated odds ratio of 1.71 (95 per cent logit confidence interval (CI) = 0.89–3.29) and a likelihood ratio against the null hypothesis of no association of 2.68 ($p = 0.05$, one-sided), the latter gives an odds ratio of 2.1 (95 per cent CI = 0.69–6.42) and a likelihood ratio of 1.83 ($p = 0.09$, one-sided). Either analysis is a valid approach to investigating the dependence of the response on group membership, but each one discards some information which is pertinent to that dependence. The cumulative odds model assumes that in the

TABLE 1

Hypothetical two group example of ordinal response data

Original data

	Response category			Total
	None	Mild	Severe	
Group 1	80	15	5	100
Group 2	70	20	10	100

Constituent tables of the cumulative odds model

	1		2	
	None	Mild or severe	None or mild	Severe
Group 1	80	20	95	5
Group 2	70	30	90	10

Constitutent tables of the continuation ratio model

	1		2	
	None	Mild or severe	Mild	Severe
Group 1	80	20	15	5
Group 2	70	30	20	10

hypothetical population from which this sample was drawn, the odds ratios from each of the two possible dichotomies are the same. If $x = 0$ for group 1, and $x = 1$ for group 2, this common odds ratio is obtained from expression 1 as e^{β}. Because the data of the 2 × 2 tables obtained from each dichotomy are correlated, estimation of the common odds ratio and inference concerning it cannot be done using techniques (Mantel-Haenszel, simple logistic regression, etc.) for obtaining summary information from *independent* strata. The parameters θ_j and β may, however, be estimated by maximizing the multinomial likelihood induced by the model 1, and inference effected using likelihood ratios, or the asymptotically equivalent score or Wald $(\{(\hat{\beta} - \beta_0)/SE(\hat{\beta})\}^2)$ chi-squared tests. Other adequate approaches to significance testing are possible, for example in the two-sample case the score statistic is the same as Wilcoxon's average rank statistic (1); these tests are not stressed here, since our emphasis is on estimation. Confidence intervals may be estimated most conveniently from standard errors on the assumption that the sampling distribution of parameters is normal. Computational issues are discussed below.

The data of table 1 give a maximum likelihood estimate for β of 0.56, and hence for the common odds ratio, e^{β}, of 1.74 (95 per cent CI = 0.92–3.33). This estimate lies between those obtained from each of the 2 × 2 tables after dichotomizing the three-point ordinal outcome, as is intuitively reasonable. The likelihood ratio against the null hypothesis that $\beta = 0$ is 2.91 ($p = 0.04$, one-sided), slightly higher than the largest of the two likelihood ratios from the 2 × 2 tables.

The continuation ratio model

Replace the cumulative probability of being in one of the first j categories in the cumulative model (γ_j) by the probability of being in category j conditional on being in category j or greater, $\delta_j = \pi_j/(1 - \gamma_{j-1})$. This gives:

$$\text{logit}(\delta_j) = \theta_j - \beta^T x, \qquad j = 1....k-1, \quad (2)$$

where

$$\text{logit}\ (\delta_j) = \ln\{\delta_j/(1 - \delta_j)\} = \ln\{\pi_j/(1 - \gamma_j)\}.$$

This model of conditional odds has been called the continuation ratio model (4), and it is essentially the proportional hazards model proposed by Cox (5) for survival data in discrete time. Note that, although the parameters θ_j and β in model 2 play approximately the same role as those given the same letters in model 1 (representing baseline category parameters and a regression parameter, respectively), they are not directly comparable, since one model predicts cumulative probabilities, and the other conditional probabilities ("hazards").

For the illustrative data, the continuation ratio model corresponds to partitioning the table into two 2 × 2 tables, but in a different manner than for the cumulative odds model. The tables are shown in table 1: the first is obtained by collapsing response levels 2 and 3 as before; but the second is obtained by discarding the responses in the first category, and comparing response category 3 to category 2. This latter gives an odds ratio of 1.5 (95 per cent CI = 0.42–5.32) and a likelihood ratio of 0.40 ($p > 0.20$, one-sided). In the continuation ratio model, it is assumed that the underlying population odds ratios from these two tables are the same and equal to e^{β}. Estimation of the parameters θ_j and β, and inference concerning them, may be carried out by maximum likelihood. Following the argument of Cox (5, 6), the strata induced by the continuation ratio model may be considered "conditionally independent", so that estimation and inference may proceed as for independent strata of binary response data.

The continuation ratio model fitted to the data in table 1 gives a maximum likelihood estimate for β of 0.51, and hence for the common odds ratio, e^{β}, of 1.67 (95 per

cent CI = 0.93–2.97). The likelihood ratio against H_0:$\beta = 0$, is 2.8 ($p = 0.05$, one-sided).

Generalizations

The cumulative odds model, 1, and the continuation ratio model, 2, have similar forms, and can be generalized in parallel ways.

First, the right hand side of expressions 1 and 2 can be generalized. As they stand, the models can include any number of explanatory variables ("covariates") in the vector x, and can thus be used to account for confounding variables and examine interactions between explanatory variables in the same way as the simple logistic regression model for dichotomous outcome (with analogous assumptions); explanatory variables may be numerical (discrete or continuous) as well as categorical. The basic assumption of the models, that covariate effects ($\beta^T x$) are additive (on the logistic scale) to "level" effects (θ_j) may be relaxed by allowing the regression parameter β to depend on level (i.e., allow interaction between covariates and levels). Thus, for the cumulative odds model:

$$\text{logit}(\gamma_j) = \theta_j + \beta_j{}^T x, \qquad (3)$$

and for the continuation ratio model:

$$\text{logit}(\delta_j) = \theta_j + \beta_j{}^T x. \qquad (4)$$

Fitting model 3 or 4 to a simple two group data set would give a "saturated" model (one in which all cell frequencies are fitted exactly), because there would be as many parameters as cells; however, with more complex covariate structures, this need not be the case.

In most applications, the usefulness of model 3 or 4 is to test the assumption of proportionality (also called parallelism) which may be formally stated as the hypothesis: $\beta_1 = \beta_2 = \ldots = \beta_{k-1}$. Where several explanatory variables are included in the model, so that β is a vector, it may be interesting to consider models in which the effect of some explanatory variables is parallel, but the effect of others is not; for example, we may wish to test for parallelism of variable effects separately.

Second, the logit "link" function of both models may be replaced by others, as with binary outcome data. The probit function is very similar to the logit; the complimentary log-log link, $f(\cdot) = \ln[\ln\{1 - (\cdot)\}]$ makes the cumulative odds equivalent to the continuation ratio model (7); the log-log link, $f(\cdot) = \ln\{-\ln(\cdot)\}$ has also been suggested. In some circumstances, log and identity links, corresponding to relative risk and risk difference models with dichotomous outcome (8), may also be useful.

Testing model assumptions

The goodness of fit of the cumulative odds or continuation ratio model to a data set may be examined in various ways. First, the likelihood ratio between the proposed model and the "saturated" multinomial model in which every cell count is fitted exactly (the "residual deviance") asymptotically follows a chi-squared distribution if all systematic variation has been accounted for by the proposed model. A residual deviance very much higher than the residual degrees of freedom is thus indicative of a poor fit. The approximation of the residual deviance to a chi-squared distribution is poor when cell frequencies are low, however, and in particular if one or more explanatory variables is continuous, this method should not be relied on (2). The Pearson chi-square ($\Sigma\{(O-E)^2/E\}$) for cell frequencies has similar properties. Second, a more directed test of model assumptions is achieved by comparing the fit of a model (usually via the likelihood ratio) with a specific generalization, such as one of those discussed in the previous section. Thus, choosing the appropriate combination of covariates and interactions between them may proceed in exactly the same way as for logistic regression with dichotomous outcome; testing proportionality (parallelism) may be done by comparing the log likelihood achieved by model 1 or 2 with that achieved by model 3 or 4. (In simple situa-

tions, such as the two group case, this is equivalent to examining the residual deviance). Comparing the fit of models with different links, or completely different ordinal regression models (e.g., the cumulative odds vs. the continuation ratio model), can be carried out informally by comparing likelihoods (9), or more formally by nesting the two models in a more general "mixture" model (10). Analysis of residuals, which may also be useful, is discussed by McCullagh (1) but not described here.

For the data of table 1, the fitted cumulative odds and the continuation ratio models gave residual deviances of 0.18 and 0.03, respectively, on one degree of freedom. Viewing these statistics as general indicators of goodness of fit, or of tests of proportionality or parallelism ($\beta_1 = \beta_2$ in models 3 and 4), we see that both models fit well, and that there is no evidence for different underlying odds ratios in the two constituent tables defined by the two possible cutpoints.

Computational issues

The cumulative odds model may be fitted to data with fairly general covariate structure (numerical or categorical, including interactions), using the SAS supplemental library program PROC LOGIST (11). The program output gives likelihood ratio and score statistics as well as maximum likelihood estimates (with standard errors) of the θ_j and β. Neither residual deviance nor a test for parallelism is provided with PROC LOGIST. However, these test statistics can be obtained with modest additional ad hoc calculations (Appendix 1).

The cumulative odds model may be shown to be a generalized linear model with a "composite link" (12). Thus, the GLIM package (13) may be adapted to fit this model, and Hutchison (14) has published a GLIM MACRO for this purpose. This method allows testing for parallelism, but is more cumbersome to use than PROC LOGIST, and works only for fairly simple covariate structures (it cannot be used with continuous covariates).

As mentioned above, the continuation ratio model may be fitted by considering the tables into which the data are partitioned as independent strata of binary responses. If the data are restructured in this way, any package capable of logistic regression can be used to estimate parameters (β and θ_j), and to calculate likelihood ratios. The GLIM package is well suited to this, allowing links other than the logit to be used, as well as having convenient mechanisms for handling categorical as well as numerical variables. Iyer (15) has given details of how continuation-ratio models may be fitted using GLIM. In addition to the explanatory variables of interest, a categorical variable ("LEVEL") indexing each of the strata formed by partitioning is required from which the level parameters, θ_j, are estimated. Parallelism may be tested by testing the interaction of "LEVEL" with the explanatory variable(s).

EXAMPLE: FIBROSIS IN MINERS

Data from a cross-sectional study of radiographic abnormalities in 244 past and present miners exposed to tremolite fibers (16) are used as an example. The response of interest was the profusion of small opacities on the chest x-ray, read according to the International Labour Office 1980 classification (17) on a 12-point ordered scale, conventionally labelled: 0/–, 0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2, 3/3, 3/4. A major objective of the study was to examine the relation between profusion of small opacities and exposure to fibers (here measured as a cumulative exposure index in units of fiber/ml-years). Potentially confounding variables included age and smoking habit.

Table 2 shows profusion score broken down by cumulative exposure, age, and smoking habit. As is common in this type of study, men are concentrated in the lower categories of the scale. Three of the profusion categories (0/–, 2/1, 3/4) were not used in the study, and thus could be dropped from consideration without losing information. Estimates of parameters from very sparse data may not be consistent (18).

TABLE 2

Example: profusion of small opacities in miners exposed to fibers

	No.	Profusion of small opacities										
		0/-	0/0	0/1	1/0	1/1	1/2	2/1	2/2	2/3	3/2	3/3	3/4
All men	244	–	182	17	16	16	5	–	3	3	1	1	–
By cumulative exposure (fiber/ml-years)													
0–10	92	–	79	6	4	2	1	–	0	0	0	0	–
–20	64	–	48	4	4	4	1	–	1	1	1	0	–
–100	53	–	38	5	3	3	2	–	1	1	0	0	–
–200	16	–	8	0	5	2	1	–	0	0	0	0	–
>200	18	–	9	2	0	5	0	–	1	1	0	1	–

Although inconsistency is unlikely to be a problem here, computational difficulties were experienced in maximizing likelihoods when the higher categories with very small frequencies were left separate; therefore, scores of 2/3 and above were collapsed into a single category, leaving seven. We fitted the cumulative odds, continuation ratio, and simple logistic model, the latter by dichotomizing the profusion score between 0/1 and 1/0, as is customary to define "normals" and "abnormals". Age and cumulative exposure were included in the models as continuous explanatory variables, and smoking as a factor of three levels (never smokers, ex-smokers, and current smokers).

Table 3 shows the components of the estimated regression parameter β (with all variables in the model), the corresponding odds ratios with 95 per cent confidence intervals, and the likelihood ratio against the null hypothesis that the relevant components of β are zero. As noted above, the regression parameter β, and hence the common odds ratio, has a different interpretation for the cumulative odds model than for the continuation ratio model; those for the simple logistic regression model are comparable to those of the cumulative odds model if the latter model holds.

Profusion score was significantly associated with age, smoking, and cumulative exposure under all three forms of analysis, although slightly less so (judging by the likelihood ratio tests) in the simple logistic than the two ordinal regression models. Likelihood ratios for smoking and cumulative exposure were similar for cumulative odds and continuation ratio models but that for age was higher in the continuation ratio model. The global likelihood ratio for all three factors simultaneously was 57.0, 75.1, and 76.8 for the simple logistic, cumulative odds, and continuation ratio models, respectively.

There was no consistent pattern of difference between the odds ratios estimated from the simple logistic and from the cumulative odds model. Standard errors of the regression parameters were slightly smaller in the cumulative odds model than in the simple logistic model, reflecting some increase in efficiency in the former. The odds ratios from the continuation ratio model were lower than the other two. This is an expected result due to the different meaning of β and hence the common odds ratio in this model (see "Discussion" section); the differences are small, however, relative to the width of the confidence intervals.

The overall multinomial log likelihood was very similar for the cumulative odds (−200.4) and the continuation ratio (−199.5) models, which have an equal number of parameters (10). The binomial log likelihood for the simple logistic model (−88.1) is not comparable. None of the overall likelihoods should be interpreted as measures of goodness of fit, since continuous exposure variables are used. Likelihood ratio tests of model assumptions against specific alternatives are shown in table 4. There appears to be no evidence for nonlinearity of age and cumulative exposure

TABLE 3

Example: fibrosis in miners—results of model fitting

Factor	Model	β (SE*)	Odds ratio	(95% CI†)	Likelihood ratio (df‡)	p value
Age (by 10-year age groups)	Simple logistic	0.84 (0.18)	2.32	(1.63–3.30)	27.2 (1)	<0.001
	Cumulative odds	0.76 (0.15)	2.14	(1.59–2.87)	31.2 (1)	<0.001
	Continuation ratio	0.68 (0.12)	1.97	(1.58–2.50)	39.1 (1)	<0.001
Smoking (relative to never smokers)						
Ex-smokers	Simple logistic	0.82 (0.69)	2.27	(0.59–8.78)		
	Cumulative odds	1.05 (0.60)	2.86	(0.88–9.26)		
	Continuation ratio	0.71 (0.44)	2.03	(0.86–4.82)		
Current smokers	Simple logistic	1.70 (0.71)	5.47	(1.36–22.01)	8.2 (2)§	0.02
	Cumulative odds	1.88 (0.62)	6.55	(1.94–22.09)	12.5 (2)§	0.002
	Continuation ratio	1.39 (0.45)	4.01	(1.66–9.70)	12.6 (2)§	0.002
Cumulative exposure (per 100 fiber/ml-years)	Simple logistic	0.35 (0.12)	1.42	(1.12–1.80)	10.9 (1)	<0.001
	Cumulative odds	0.38 (0.09)	1.46	(1.23–1.74)	17.3 (1)	<0.001
	Continuation ratio	0.24 (0.06)	1.27	(1.12–1.44)	17.8 (1)	<0.001

* SE, standard error. † CI, confidence interval. ‡ df, degrees of freedom.
§ Likelihood ratio for removing both smoking variables.

TABLE 4

Example: testing model assumptions

Additional term in the model	df†	Likelihood ratio		
		Simple logistic	Cumulative odds	Continuation ratio
Linearity of effects				
Age^2	1	0.3	0.4	2.2
Cumulative exposure^2	1	0.5	0.6	3.3
Interactions between explanatory variables				
Age*smoking	2	0.1	0.4	0.2
Smoking*cumulative exposure	2	0.1	0.1	0.2
Age‡*cumulative exposure	1	0.1	0.2	0.0
Interactions between level and explanatory variables (separate slopes models)				
Level*age	5	–	6.6§	8.0
Level*smoking	8	–	6.3§	8.0
Level*cumulative exposure	5	–	5.4§	2.5
Level*(age + smoking + cumulative exposure) (all effects allowed to depend on level)	18	–	20.1	18.8

† df, degrees of freedom.
‡ Age dichotimized at 60 years to define interaction.
§ Approximate values (see Appendix 1).

effects in the simple logistic or cumulative odds model. The likelihood ratio for a quadratic cumulative exposure effect in the continuation ratio model is on the borderline of conventional statistical significance ($p = 0.07$). There was no evidence for interaction between age, smoking, and cumulative exposure in any of the three models,

nor for interaction between outcome level and these variables in the cumulative odds or continuation ratio models (i.e., against parallelism of effects of age, smoking, or cumulative exposure).

Subsets of the data were also investigated using the three models, with similar general results. In no case did major issues of interpretation depend on choice of model. The reduction in standard error of β from using the cumulative odds rather than the simple logistic model was never greater than that in the analysis shown.

THE EFFICIENCY OF ANALYSES BASED ON THE CUMULATIVE ODDS MODEL RELATIVE TO THOSE BASED ON THE SIMPLE LOGISTIC MODEL WITH DICHOTOMIZED OUTCOME

The primary motivation for use of an ordinal regression model such as 1 or 2, rather than collapsing the outcome variable to just two groups and using conventional methods, is a desire to use all available information, and hence gain more power in investigating the effect of risk factors on outcome. However, in the example considered, this power gain was not very pronounced, and use of ordinal regression would not have changed interpretation of results. We have examined theoretical power gain by calculating, for various situations, the asymptotic relative efficiency of the logistic regression analysis using dichotomized outcome, compared with an analysis using the cumulative odds model.

Asymptotic relative efficiency is defined as the limit, as sample sizes increase, of the ratio of the sample sizes required for the two methods, in order that each achieve the same power (or equivalently the same precision) when close to the null hypothesis (19). In this case, this is also equal to the ratio of variances of the two estimates of β. For simplicity, we consider the comparison of response in two equally sized groups. Because we are initially trying to show the maximum power gain by using the cumulative odds model, we look first at the situation in which under the null hypothesis each category of response is equally probable, since it is in this situation that the cumulative odds model is most powerful. Under these assumptions, a simple formula for asymptotic relative efficiency can be derived (see Appendix 2). The value of this, expressed as a percentage, is shown in table 5 for ordinal outcome variables with 3–10 levels. Cumulative odds regression with k outcome levels is compared with each of $(k - 1)$ simple logistic regressions, where the dichotomous outcome for the latter analyses are created by taking the single cut-point after each of levels 1,..., $(k - 1)$ in turn. Thus, for example, if the outcome has four levels, simple logistic regression has 60 per cent of the efficiency of cumulative odds

TABLE 5

Asymptotic relative efficiency (per cent) of simple logistic regression compared with cumulative odds regression, comparison of two groups, each category of outcome equally probable

Cut-point*	No. of levels of outcome (k)							
	(2)†	3	4	5	6	7	8	9
>1	(100)	75	60	50	43	38	33	30
>2		75	80	75	69	63	57	53
>3			60	75	77	75	71	68
>4				50	69	75	76	75
>5					43	63	71	75
>6						38	57	68
>7							33	53
>8								30

* Definition of positive outcome for simple logistic regression.

† At two levels of outcome, cumulative odds and simple logistic regression are identical.

regression if the dichotomy is made by cutting after level 1 or level 3, but 80 per cent relatively efficient if the cut is after level 2. This exercise shows that if the dichotomy for simple logistic regression is close to its optimal point (creating equal numbers of "positive" and "negative" responders), then power gain using the cumulative odds model is modest, since the relative efficiency of the simple logistic regression is between 75 and 80 per cent, depending on the numbers of categories used.

Also discernible indirectly from table 5 is the fact that power gain on increasing the number of categories of outcome suffers from a law of quickly diminishing returns. In fact, if we assume an underlying continuous outcome, so that cut-points may be chosen arbitrarily, we may directly investigate the efficiency of cumulative odds models as a function of the number of levels used (again, assumed chosen to given equal probability in each category). Maximum precision of estimated β (equivalent to maximum power) in large samples may be considered as the limit as number of outcome levels increase. Relative to this limit, simple logistic regression (dichotomizing at the median outcome level) is asymptotically 75 per cent efficient, and the relative efficiency of ordinal regression using from three to nine groups is 89, 94, 96, 97, 98, 98, and 99 per cent, respectively.

Since the variance of β estimated from the cumulative odds model is minimum if $\pi_{ij} = 1/k$ (see Appendix 2), potential gain from using ordinal regression methods would be expected to be smaller if the π_{ij} vary. For example, if the probability of falling in each of seven levels was the same as for the real example presented above (i.e., 75, 7, 7, 7, 2, 1, and 2 per cent), then the efficiency of simple logistic regression relative to cumulative odds ordinal regression is 97 per cent if the cut-point for the simple logistic regression is after the first level used, 77 per cent if the cut-point is after the second level (the conventional ≥1/0 used in the example), and 54, 26, 16, and 10 per cent if the cut-point is after levels 3, 4, 5, and 6, respectively.

To test the applicability of asymptotic efficiencies to small data sets with substantial risk differences between groups, a small simulation study was carried out. One hundred data sets, each containing 100 subjects divided into two exposure groups, were generated according to the cumulative odds model. The parameters for the simulation were based on those found in a subset of data from our example. Specifically, five response categories, 0/0, 0/1, 1/0, 1/1, 1/2+, were assumed, with probabilities of 0.904, 0.040, 0.027, 0.029, 0.007, respectively, in the less exposed group. The effect parameter β was taken as 1.10, equivalent to an odds ratio of three between exposure groups. Each data set was then analyzed using a simple logistic model (defining "positive" as ≥1/0 as before), a cumulative odds model, and a continuation ratio model.

Results of the simulations, shown in table 6, broadly confirm the applicability of asymptotic efficiencies to this situation. The empirical efficiency of the simple logistic regression estimate of β relative to

TABLE 6

Results of simulations comparing analyses by the simple logistic, cumulative odds, and continuation ratio models

Model	Mean of estimated β* (CI†)	Standard deviation of estimated β (CI†)	Mean of test statistic‡ (CI†)	% significant at $p < 0.05$
Simple logistic	1.19 (1.08–1.30)	0.55 (0.48–0.63)	6.7 (5.7–7.7)	75
Cumulative odds	1.16 (1.06–1.26)	0.50 (0.44–0.58)	8.8 (7.6–10.0)	84
Continuation ratio	0.85 (0.77–0.93)	0.42 (0.37–0.49)	7.5 (6.3–8.7)	76

* True $\beta = 1.1$.

† CI, 95 per cent confidence interval.

‡ Likelihood ratio test against null hypothesis: $\beta = 0$.

that obtained from the cumulative odds model (informally estimated as the ratio of the variances of the two estimates of β over the 100 simulations) is 83 per cent, very close to the asymptotic efficiency of 82 per cent. A comparison of the two mean likelihood ratios against the null, $\beta = 0$, also reflect a modest information gain from using the cumulative odds model. If data were simulated to follow a model other than the cumulative odds, we would generally expect to see results less favorable to this model.

Direct comparison of analyses based on the continuation ratio model with those based on the other two is not possible, because of differences between the models reflected in the different interpretation of the parameter β; however, we expect the continuation ratio model to have properties similar to the cumulative odds model in this respect. Results for the continuation ratio model from the simulations (which show efficiency intermediate between the two other methods) would be more favorable if the data had been simulated to follow this model.

TABLE 7

Example: application of polytomous logistic regression model for nominal data

Level	Regression parameter (SE*) (per 100 fiber/ml-years)	Odds ratio (95% CI†) (relative to level 0/0)
0/0	-	1.0 (baseline)
0/1	0.01 (0.26)	1.01 (0.61–1.68)
1/0	0.05 (0.25)	1.05 (0.64–1.72)
1/1	0.46 (0.15)	1.58 (1.18–2.13)
1/2	−0.04 (0.64)	0.96 (0.27–3.37)
2/1	-	-
2/2	0.46 (0.23)	1.58 (1.01–2.49)
2/3+	0.74 (0.21)	2.10 (1.38–3.16)

* SE, standard error.

† CI, confidence interval.

DISCUSSION

Other models for ordinal response data

Polytomous logistic regression models for nominal (unordered categorical) responses are gaining popularity among biostatisticians and epidemiologists (20, 21) and may also be applied to ordinal responses. In its general form, this type of model may be fitted with repeated simple logistic regressions (22) and is available in standard statistical packages such as SAS and SPSSX. A polytomous regression model for x-ray category on age, smoking, and cumulative exposure was fitted (using SAS PROC LOGIST) to the example described above. The model required 30 parameters. The six representing the effects of cumulative exposure (for increase in log odds per 100 fiber/ml-years of having profusion score 0/1, 1/0, 1/1, 1/2, 2/2, 2/3+ relative to having score 0/0) are shown in table 7. The (Wald) chi-square for cumulative exposure was 15.1 on six degrees of freedom—a clearly significant effect ($p = 0.02$), but less strongly so than for the more sparsely parameterized ordinal models discussed above. The ordinality of the scale is suggested by the trend in odds ratios, but the confidence intervals are wide, so that it would be inappropriate to pay great attention to deviations from this trend. The overall log likelihood was −191.5—almost exactly the improvement in fit over the cumulative odds and continuation ratio models that would be expected by chance, given the 20 extra parameters.

To overcome the problem of overparameterization (in general, $(p + 1)(k - 1)$ parameters are required for p explanatory variables and k response categories), most authors have suggested reduced forms of the general model. Some of these, while not implying ordinality, are especially suitable for reflecting patterns typical in ordinal response data. Cox and Chuang (23) explore a number of possibilities along these lines, and they also illustrate the use of cumulative odds and continuation ratio models. Anderson (24) proposes a family of "stereotype regression models" for ordinal response data, which lie within the polytomous logistic regression model framework. Particular advantages of these models are that order of response categories need not be specified a priori, and tests for distin-

guishability of response categories are immediate.

We concur with the view of McCullagh (1, 25) and McCullagh and Nelder (2) that if order of categories may be specified with confidence a priori, models making this ordering a strong assumption (such as the cumulative odds and continuation ratio models) are preferable to the more flexible logistic models such as that of Anderson (24) because of their simpler interpretation. However, we acknowledge that this view is contentious among statisticians. Perhaps the attitude that data analysts take on this issue will reflect the relative emphasis that they wish to give to a priori considerations and to information contained in the data. A more practical impediment to wider use of general polytomous logistic models is the absence of easy to use software for fitting the more interesting reduced forms.

A substantial body of statistical literature on the analysis of ordinal data has been published in the contingency table tradition (26). Since these models do not generally distinguish between explanatory and dependent variables, they are less suitable for the regression context that is considered here. One approach in this tradition that could nevertheless be applied is to assign a score to each category of outcome, and to include this score in the specification of a log-linear model for cell probabilities. The assignation of scores introduces an undesirable arbitrary element into the analysis, but, to avoid this, Chuang and Agresti (27) proposed taking the scores as unknown parameters that are estimated from the data.

Choice of ordinal regression model

The cumulative odds and continuation ratio models have several properties which we think make them attractive to epidemiologic data analysts:

1) Both models may be interpreted in terms of odds ratios, which are familiar to epidemiologists.

2) The basic underlying assumption of each model—equality of β_js in models 3 or 4—can in principal be tested simply (although this requires some computing beyond that available in the major packages for the cumulative odds models), and can be intuitively grasped as the equality of odds ratios in 2×2 tables.

3) Packages exist (SAS PROC LOGIST and GLIM) which can estimate parameters and carry out significance tests.

4) Statistical models should be plausible biologically. However, in few epidemiologic contexts is enough known about biologic mechanisms for this criterion to be evaluated. The grouped continuous interpretation of the cumulative odds model is consistent with a mechanism in which regressor variables shift the location of an underlying continuous response that follows a logistic distribution.

Nature may not always be so obliging as to present investigators with data that conform to convenient models. However, data will often be insufficient to distinguish between models—see for example Cox and Chuang (23) or our example—and interpretation will depend little on the model used, in which case, choice of the most convenient model seems justified.

We turn now to the choice between the cumulative odds and continuation ratio models. If the strict cumulative odds model, 1, holds (i.e., *cumulative* log odds ratios are constant, say β^*), then the *conditional* log odds ratios (β_j) of the more general continuation ratio model 4, will start at $\beta_1 = \beta^*$, but then tend toward 0 as j increases. Thus, the cumulative odds model has been proposed as an alternative to the proportional hazards (i.e., continuation ratio) model for survival data when hazard rates of groups are thought likely to converge with time (28). (Essentially the same argument explains why if the cumulative odds and continuation ratio models 1 and 2 are both fitted to a set of data, the estimated continuation odds ratio will be less than the cumulative odds ratio, as seen in the fibrosis example.) In many data sets, the fit of the two models (models 1 and 2) is similar,

however, so that choice between them must be on a priori grounds.

A feature of the cumulative odds model but not of the continuation ratio model is that it is unchanged if the ordering of the categories is reversed. In some applications (such as to survival data), the direction of ordering appropriate for the continuation ratio (proportional hazards) model is clear from the context. In other applications, this is not so, so that the fact that different results may be obtained according to this arbitrary choice is an undesirable feature of the model (although differences will often be marginal; on reversing the order for the data in table 1, the estimate of β from the continuation ratio model changed from 0.51 to 0.52, and the likelihood ratio against the null changed from 2.8 to 2.9).

Concluding remarks

We concluded from our study of efficiency that simple logistic regression applied to a dichotomized response measure has only slightly less power than the ordinal regression methods discussed if the cut-point is made close to its optimal point. However, there may be reasons other than efficiency for not dichotomizing (for example, investigating whether level-specific effects are present). Furthermore, choice of exactly where to make the cut-point for simple logistic regression constitutes an undesirable arbitrary element in the analysis. Sometimes (for example, with chest x-rays), the choice is strongly suggested by convention, but the conventional cut-point might not be close to that optimal for achieving power. Thus, where avoiding loss of information is a major consideration, or if choosing a single cut-point to define a positive response involves an arbitrary choice, the use of ordinal regression methods are indicated.

References

1. McCullagh P. Regression models for ordinal data (with discussion). J R Stat Soc [B] 1980;42:109–27.
2. McCullagh P, Nelder JA. Generalized linear models. London: Chapman Hall, 1983.
3. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967;54:167–79.
4. Fienberg SE. The analysis of cross-classified data. 2nd ed. Cambridge, MA: MIT Press, 1980.
5. Cox DR. Regression models and life tables (with discussion). J R Stat Soc [B] 1972;74:187–220.
6. Cox DR. Partial likelihood. Biometrika 1975; 62:269–76.
7. Laara E, Matthews JNS. The equivalence of two models for ordinal data. Biometrika 1985;72:206–7.
8. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol 1986;123:174–84.
9. Walker AM, Rothman KJ. Models of varying parametric form in case-referent studies. Am J Epidemiol 1982;115:129–37.
10. Atkinson AC. A method for discriminating between models (with discussion). J R Stat Soc [B] 1970;32:323–53.
11. SAS Institute Inc. SUGI supplemental library user's guide, version 5 edition. Cary, NC: SAS Institute Inc., 1986.
12. Thompson R, Baker RJ. Composite link functions in generalized linear models. Appl Stat 1981; 54:167–79.
13. Payne CD, ed. The GLIM system release 3.77 manual. Oxford: Numerical Algorithms Group, 1985.
14. Hutchison D. Ordinal variable regression using the McCullagh (proportional odds) model. GLIM Newsletter 1985;9:9–17.
15. Iyer R. Continuation-odds model in ordinal variable regression. GLIM Newsletter 1985;10:4–8.
16. McDonald JC, Sebastien P, Armstrong B. Radiological survey of past and present vermiculite miners exposed to amphibole fibres in the tremolite series. Br J Ind Med 1986;43:445–9.
17. International Labour Office. Guidelines for the use of ILO international classification of radiographs of pneumoconioses. Geneva: ILO, 1980. (Occupational Safety and Health, series 22, rev. 80).
18. McCullagh P. On the elimination of nuisance parameters in the proportional odds model. J R Stat Soc [B] 1984;46:250–6.
19. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall, 1974:304–97.
20. Marshall RT, Chisholm EM. Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer. Stat Med 1985;4:337–44.
21. Thomas DC, Golberg M, Dewar R, et al. Statistical methods for relating several exposure factors to several diseases in case heterogeneity studies. Stat Med 1986;5:49–60.
22. Begg CB, Gray R. Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika 1984;71:11–18.
23. Cox C, Chuang C. A comparison of chi-square partitioning and two logit analyses of ordinal pain data from a pharmaceutical study. Stat Med 1984;3:273–85.
24. Anderson JA. Regression and ordered categorical variables (with discussion). J R Stat Soc [B] 1984;46:1–30.

25. McCullagh P. In: Discussion of Anderson JA. Regression and ordered categorical variables. J R Stat Soc [B] 1984;46:22–3.
26. Agresti A. Analysis of ordinal categorical data. New York: John Wiley and Sons, 1984.
27. Chuang C, Agresti A. A new model for ordinal pain data from a pharmaceutical study. Stat Med 1986;5:15–20.
28. Bennett S. Analysis of survival data by the proportional odds model. Stat Med 1983;2:279–85.

APPENDIX 1

Calculating log likelihoods, residual deviances, and tests for parallelism in the cumulative odds model

It is useful to define a general multinomial model in which probability π_{ij} of being in category j is a function of a vector of explanatory variables x_i and unknown parameters. A complete multinomial observation vector is denoted $y_i = (y_{i1},...,y_{ik})$, where y_{ij} is the count in the jth category of all persons with the same vector of explanatory variables x_i. (When x includes a continuous variable, it is usually sensible to index each person $i = 1...n$, so that for all i, $y_{ij} = 0$ for all but one j, for which it is 1.) The log likelihood for the total data $\{y_i\}$ is then

$$l = \sum_i \sum_j y_{ij} \ln (\pi_{ij}) \tag{1.1}$$

where $\ln(\cdot)$ denotes the natural logarithmic function. In the cumulative odds model, π_{ij} is dependent on parameters θ_j and β through the cumulative probabilities γ_{ij} according to the equations:

$$\begin{aligned} \pi_{i1} &= \gamma_{i1}, \\ \pi_{ij} &= \gamma_{ij} - \gamma_{i(j-1)}, \qquad \text{for } j = 2 \ldots k-1 \\ \pi_{ik} &= 1 - \gamma_{i(k-1)}, \end{aligned} \tag{1.2}$$

and γ_{ij} is given by

$$\gamma_{ij} = 1/\{1 + \exp(-\theta_j + \beta^T x_i)\}. \qquad j = 1 \ldots k-1 \tag{1.3}$$

The log likelihood can be maximized using any general maximization routine, for example, the BMDP program P3R (see reference 23 and the BMDP manual for some further details). Since use of such general programs requires greater mathematical and computing expertise than does the use of SAS PROC LOGIST, it is worth considering ad hoc methods for calculating the useful statistics which PROC LOGIST does not provide—the residual deviance and the test for parallelism.

The residual deviance is equal to twice the difference between the log likelihood in the model under consideration and that in the "saturated" model in which all cells are fitted exactly. Since PROC LOGIST provides the former, it is necessary only to calculate the latter. This is obtained by replacing π_{ij} in expression 1.1 by $y_{ij}/\sum_j y_{ij}$, i.e., the actual observed proportion of observations with this combination of explanatory variables which fall into category j. Thus, for the illustrative data in table 1 the saturated log likelihood (l) is

$$-80 \times \ln(80/100) - 15 \times \ln(15/100) + ... = -141.47.$$

Since the log likelihood for the cumulative odds model 1 is (from PROC LOGIST) 141.55, the residual deviance is $2 \times (141.55 - 141.47) = 0.16$.

To obtain the test for parallelism of βs in the general cumulative odds model, we must first fit the "separate slopes" model 3, and compare the log likelihood obtained using this model with that (given by PROC LOGIST) for the more restricted cumulative odds model 1. Fitting the model in which no variable effects are assumed parallel, and hence all components of β are allowed to depend on level j, may be carried out by fitting $k - 1$ separate simple logistic regression equations with the dependent variable defined at each possible cut-point in turn. Since the intercept and slope terms of the simple logistic regression dichotomizing between levels j and $j + 1$ correspond to θ_j and $-\beta_j$, respectively, in model 3, equations 1.3, 1.2, and 1.1 can be used to calculate a log likelihood.

For the data in table 1, simple logistic regression dichotomizing between levels 1 and 2 gives intercept $\theta_1 = 1.386$ and slope $\beta_1 = 0.539$; dichotomizing between levels 2 and 3 gives $\theta_2 = 2.944$ and $\beta_2 = 0.747$. Applying 1.3 and 1.2, we obtain from 1.1:

$$\pi_{11} = \gamma_{11} = 1/\{1 + \exp(-1.386 + 0.539 \times 0)\} = 0.80,$$

$$\pi_{12} = \ldots., \text{ etc.}$$

$$\text{and } l = -80 \times \ln(0.80) - .... = -141.47.$$

Hence, the likelihood ratio test for parallelism is $2 \times (141.55 - 141.47) = 0.16$. This is the same as the residual deviance, because, in this simple case, the separate slopes model is the "saturated" model.

In practice, it is usually easier to deal with fitted cumulative counts from the simple logistic regressions

$$(s_{ij} = \sum_{m=1}^{j} y_{im})$$

and to calculate the log likelihood as

$$\sum_i \sum_j y_{ij} \ln(\hat{y}_{ij}/\sum y_{ij}),$$

where

$$\hat{y}_{ij} = s_{ij} - s_{i(j-1)}.$$

We may wish also to obtain the fitted parameters and the log likelihood for models in which some components of β are assumed the same for all levels, whereas others depend on level—for example, to test parallelism for specific variables only. Exact ad hoc calculations for this are tedious, but approximate estimates may be obtained by keeping those components of β which are assumed the same for all levels fixed at the estimated value given for model 1 by PROC LOGIST, and offsetting these effects in fitting separate estimates for remaining components of β as above.

The calculations above may be performed in simple cases by hand held calculators, or programmed quite simply using a high level language. We have found the GLIM language suitable for this.

Finally, we should note a theoretical problem with model 3 that can sometimes hinder estimation. Since the β_j vary with j, graphs of logit (γ_j) against x will intersect at some value of x, so that cumulative probability cannot be a monotone increasing function of j for all x. Thus, negative probabilities π_j will be predicted for some x values. Usually, the points of intersection do not occur within the domain of the observed x, so that this does not present a problem.

APPENDIX 2

Derivation of asymptotic relative efficiencies

For the cumulative odds model 1 applied to the comparison of outcome in two groups (of size n_1 and n_2, $n_1 + n_2 = n$), the asymptotic variance of β (here, the log odds ratio) close to the null hypothesis, $\beta = 0$, may be shown (for example, by McCullagh (1)) to be equal to

$$(n_1 n_2/3n)\{1 - \sum_{j=1}^{k} \pi_j^3\}. \tag{2.1}$$

(Since $\beta \cong 0$, π_j is approximately equal in the two groups.)

Simple logistic regression after dichotomizing between levels m and $m + 1$ may be seen as a special case of the cumulative odds model, giving asymptotic variance near the null

$$(n_1 n_2/3n)/\{1 - (\sum_{j=1}^{m} \pi_j)^3 - (\sum_{j=m+1}^{k} \pi_j)^3\}. \tag{2.2}$$

Asymptotic relative efficiency of simple logistic regression relative to the cumulative odds model is thus expression 2.2 divided by expression 2.1.

As McCullagh (1) observes, it is clear from expression 2.1 that variance is minimum if $\pi_1 = \pi_2 = \ldots = \pi_k = (1/k)$. Substituting this in 2.2 and 2.1 and dividing, we obtain the asymptotic relative efficiency for this special case:

$$\{1 - (m/k)^3 - [(k - m)/k]^3\}/\{1 - k(1/k)^3\} = 3m(k - m)/(k^2 - 1).$$