STATISTICS IN MEDICINE, VOL. 16, 2873–2882 (1997)

SIMPLE MODELS FOR REPEATED ORDINAL RESPONSES WITH AN APPLICATION TO A SEASONAL RHINITIS CLINICAL TRIAL†

J. K. LINDSEY[1], B. JONES*[1], AND A. F. EBBUTT[2]

[1]*Department of Medical Statistics, School of Computing Sciences, De Montfort University, The Gateway, Leicester LE1 9BH, U.K.*
[2]*European Clinical Statistics, Glaxo Wellcome Ltd., Greenford Road, Middlesex UB6 0HE, U.K.*

SUMMARY

In contrast to other models for ordinal data, the continuation ratio model can be fitted with standard statistical software. This makes it particularly appropriate for large clinical trials with ordinal response variables. In addition, when the trials are longitudinal, this model can be applied to individual responses instead of frequencies in contingency tables. Dependence can be incorporated by conditioning on the previous response, yielding a form of Markov chain. This approach is applied to the analysis of a large seasonal rhinitis trial, where patients were observed over 28 days and six symptoms recorded as ordinal responses. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2873–2882 (1997)
No. of Figures: 0 No. of Tables: 5 No. of References: 13

1. INTRODUCTION

Ordinal responses are very common as response variables measured in clinical trials. Most often, these involve the evaluation of the state of each patient. Many models have been proposed for such responses, but many require specialized software which may not be widely available or special programming for the problem at hand. Even when one of these two conditions if fulfilled, very large data sets cannot usually be handled. In this paper, we look at some standard techniques for modelling ordinal variables which can be applied to large longitudinal data sets using widely available software.

As an example of the problem, consider a seasonal rhinitis clinical trial, where interest centres on sustained relief from symptoms. Patients were randomized to one of three treatments: A; B; and placebo. Responses on six types of symptoms (blockage on waking, blockage during the day, sneezing, nasal itching, runny nose, eye watering) were recorded for 28 consecutive days. All symptom responses were scored on a 0–3 scale, with 0 being no symptom and 3 being bad. The first type of symptom was recorded in the morning and the others in the afternoon. Occasionally,

†Presented at the International Society for Clinical Biostatistics, Sixteenth International Meeting, Barcelona, Spain, July–August 1995.

*Correspondence to: B. Jones, Department of Medical Statistics, School of Computing Sciences, De Montfort University, The Gateway, Leicester LE1 9BH, U.K.

CCC 0277–6715/97/242873–10$17.50

a supplementary rescue medication was administered, but this is not of interest in the present analysis and will not be considered further here. A total of 416 patients were randomized to give a potential total of 11,648 observations. Because of dropouts and some randomly missing observations, about 10,650 were in fact available, with slight variations among symptoms. This is an average of about 2·5 missing observations per patient. Many are due to a few patients dropping out about half way through, with some others having observations missing on the first day or two. In most cases, these can be assumed to be random, not depending on treatment.

Common procedures, in the pharmaceutical industry as elsewhere, might be to reduce the 28 days worth of data on a subject to a single summary measure, such as the median response, the percentage of symptom-free (score equal to zero) days, or the response at the end of the period. (This last one would only be of interest in the study of a cure for a disease, which is not the case here.) Treatment differences might then be tested by a Wilcoxon non-parametric test. A slightly more sophisticated analysis would dichotomize the responses as symptom (that is, greater than zero) or not and apply a logistic model, often ignoring dependence among responses on the same patient. None of these methods uses the information in the data efficiently, and some, such as ignoring independence, are clearly wrong. The procedures recommended here require little more work than fitting a standard logistic model and do not involve the loss of information in such dichotomization nor the approximation of assuming independence.

The three commonly used models for ordinal responses are the log multiplicative,[1] continuation ratio or stereotype,[2] and proportional odds[3] models. All three are generally applied to frequency data in contingency tables, and hence are not directly applicable to longitudinal data. The first and last require more specialized software, as mentioned above. However, the continuation ratio model can be fitted by standard logistic regression techniques after a simple restructuring of the data. It can also be fitted directly to individual responses, not regrouped in a contingency table. Thus, it is particularly suitable for longitudinal data models.

2. THE CONTINUATION RATIO MODEL

Each of the three models mentioned above may be preferable in particular situations,[4] although the proponents of each seem to think that their model is best for all situations! Thus, the proportional odds model is characterized by being easily interpretable if the number of ordinal levels is changed, for example by collapsing some adjacent categories. The log multiplicative model yields a scale showing how close the categories are to each other and can also detect scores which are not properly ordered. Although closely related to the proportional odds model, the continuation ratio model has the advantage of being a simple decomposition of a multinomial distribution, with the result that it can be fitted by standard software. However, in spite of these differences, in general, all three give rise to very similar conclusions about treatment effects in practical applications. Ashby *et al.*[5] provide a clear discussion of ordered polytomous regression applied to single outcomes, especially with respect to the proportional odds model.

The continuation ratio model makes a series of comparisons of all lower categories on a scale to the next succeeding one. Thus, the first category is compared to the second, the first two to the third, and so on. Obviously, this comparison is asymmetric, and one could start from the top of the scale, instead of the bottom, if this made more sense. This model describes the probability of moving one step further on the scale, given present position. In many longitudinal studies, this may be an appropriate response to consider.

For a score with I possible categories, there will be $I-1$ comparisons. Thus, if the multinomial distribution of four ordinal responses has probabilities π_i, $i = 1, 2, 3$, with $\pi_0 = 1 - \pi_1 - \pi_2 - \pi_3$, we can reparameterize as the series of conditional probabilities

$$\lambda_1 = \frac{\pi_1}{1 - \pi_2 - \pi_3}$$

$$\lambda_2 = \frac{\pi_2}{1 - \pi_3}$$

$$\lambda_3 = \pi_3.$$

Thus, for example, λ_1 will be the conditional probability of a response at level 1, given that it is no higher than 1, that is, neither 2 nor 3. In terms of these new parameters, the original multinomial distribution becomes a product of $I-1$ (here three) binomial distributions. The usual assumption for this model is that all λ_i are equal for a given level of the explanatory variables, that is, that these new probabilities do not interact with explanatory variables such as treatments.

Thus, a contingency table with an I category ordinal response variable can be reconstructed as $I-1$ subtables making the comparisons described above. Then, a logistic regression, using any standard software package, can be applied to the resulting binary response variable. However, individual responses can also be reconstructed in the same way.[6,7] In the first subtable, only individuals with responses in the first two categories are compared. In the second subtable, those in the first three categories are included, with the binary variable comparing third to first or second. For example, a person with a response 2 will be coded 1 in the second subtable and 0 in the third, but will not appear in the first subtable, while a response 3 will only appear in the third subtable, with code 1.

This procedure can result in rather large variable vectors, close to $I-1$ times the number of observations. The exact number depends on how many individuals have responses at the higher levels, because these are eliminated from the earlier subtables. Thus, for the rhinitis trial data, the vectors will be of length about 30,000. For certain software packages, logistic regression with such large vectors may take some time, but with modern computer power this should generally not be a problem.

Up until now, our discussion of ordinal variable models has not taken into account the dependence among successive responses of an individual in repeated measurement situations. Two types of such dependence might be expected:[8] those arising from heterogeneity among individuals, often called frailty, and those from serial correlation in time. Unfortunately, the first, although often important, is also very difficult to model realistically with the present state of statistical software. In keeping with our goal to provide simple, practical methods for the analysis of ordinal longitudinal clinical trial data, we must ignore this problem here, but discuss it further in the last section. On the other hand, serial correlation can easily be accommodated by conditioning; indeed, this will usually account for a considerable proportion of the heterogeneity.

It is now well-known that Markov chains can be fitted to binary data by standard logistic models.[9,10] In such a model, the present response at any time point is made conditional on that in the previous time period. However, little or no work seems to have been done on applying such methods to ordinal responses. With the transformed binary response of the continuation ratio model, these techniques can be applied directly. One could condition on the lagged value of the binary response, but interpretation is usually easier if the lagged ordinal response variable is used directly as an explanatory factor variable instead of recoding it.

3. RESULTS

Several standard analyses were carried out by the statisticians at the end of the study on seasonal rhinitis. For the percentage of symptom-free days, each of placebo and B was compared to A using a Wilcoxon rank sum test. For each of the six symptoms, the result was a very significant difference with placebo but not with B. For the median symptom score over the 28 days, a proportional odds model was fitted to compare the three treatments. The 95 per cent confidence intervals for the relative odds comparing A and placebo did not include one, but that comparing A and B did include one for each of the six symptoms. Neither of these procedures used the longitudinal aspects of the data.

3.1. Models assuming independence

In order to determine if there is dependence among repeated responses for a patient, it is necessary to have available the results from the simpler models that assume independence. These will provide a base null model with which our more sophisticated procedures can be compared. We thus begin by studying some of these models, although they are not expected to be appropriate for the data at hand.

If we only consider the relationship between treatment and the ordinal response, aggregating over time and hence ignoring serial dependence, we can summarize the results in contingency tables from which summary statistics such as the median response or percentage of symptom free days just mentioned are directly available. The contingency tables for the six different response symptoms are presented in Table I. Missing responses have been eliminated, assuming them to be random.

However, the data presented in this way ignore completely that several responses have been observed on each patient, so that any inferences based on such measures will be misleadingly precise. Let us, nevertheless, continue and inspect these tables, applying standard log-linear or logistic modelling techniques, as well as the continuation ratio model,[2] looking at the results as simple descriptive statistics. These will provide a basis for comparison with the models to follow; however, we would not recommend that anyone only perform these naive independence-based analyses.

The second column of Table II gives the usual deviance (minus two log-likelihood), with six degrees of freedom, for independence of response from treatment for each contingency table, ignoring the ordinal character of the response, based on the log-linear model

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

where i indexes treatment and j score with mean frequency, μ_{ij}. Evidently, by any inference criterion, there is indication of a strong difference in response score with treatment. This treatment deviance can be decomposed into an ordinal part (with 2 degrees of freedom) from the continuation ratio model and a lack of fit of this model (with 4 degrees of freedom), as in the third and fourth columns of Table II. For the continuation ratio model, each table is reconstructed, as described above, into three subtables with binary responses and a logistic regression without interaction between subtable and treatment is applied

$$\log\left(\frac{\pi_{ij}}{\pi_{ij-}}\right) = \theta + \delta_i + \gamma_j$$

Table I. Contingency tables for each of the six ordinal symptom scores crossed with treatment

Symptom	Treatment	Score			
		0	1	2	3
1	A	1679	1601	345	74
	B	1858	1446	289	84
	Placebo	730	1458	676	365
2	A	2057	1307	302	41
	B	2133	1255	226	72
	Placebo	1099	1319	566	270
3	A	1801	1447	371	88
	B	1968	1277	372	70
	Placebo	746	1506	789	213
4	A	2282	1094	262	68
	B	2227	1056	326	78
	Placebo	1043	1311	645	255
5	A	2213	1051	340	104
	B	2252	1044	314	77
	Placebo	800	1340	779	333
6	A	1882	1143	422	259
	B	1796	1100	614	177
	Placebo	1284	1141	592	237

Table II. Analysis of deviance for treatment effect in the continuation ratio and logistic models for the six symptoms. Only the last column allows for response dependence on each individual. All deviances, except column 4, refer to treatment differences

Symptom	Multinomial			Logistic 2 d.f.	Ordinal Markov chain 2 d.f.
	Total 6 d.f.	Ordinal 2 d.f.	Lack of fit 4 d.f.		
1	1025·2	945·6	79·6	649·5	167·8
2	769·2	659·0	110·2	485·0	108·5
3	944·0	936·0	8·0	780·2	201·0
4	927·1	902·7	24·4	770·3	164·8
5	1386·0	1377·6	8·4	1200·8	206·6
6	154·4	102·2	52·2	99·9	19·2

where $j-$ indicates the combined group of scores less than j and π_{ij} is the conditional probability of score $j(=1, 2, 3)$ under treatment i given that the score is no higher than j. This is the deviance of the third column, while the deviance for interaction is the lack of fit.

Table III. Parameter estimates for treatment differences with respect to placebo from the logistic and continuation ratio models for the six symptoms

Symptom	Treatment	Logistic	Continuation ratio	
			Independence	Markov chain
1	A	−1·05	−1·02	−0·56
	B	−1·25	−1·19	−0·63
2	A	−0·89	−0·89	−0·49
	B	−0·98	−0·98	−0·53
3	A	−1·16	−1·01	−0·56
	B	−1·36	−1·16	−0·64
4	A	−1·22	−1·11	−0·62
	B	−1·27	−1·05	−0·56
5	A	−1·51	−1·29	−0·63
	B	−1·57	−1·35	−0·68
6	A	−0·46	−0·30	−0·19
	B	−0·38	−0·38	−0·18

The fifth column of Table II gives the changes in deviance for removing treatment from a standard logistic model

$$\log\left(\frac{v_i}{1-v_i}\right) = \varepsilon + \xi_i$$

where v_i is the probability of zero score under treatment i. This model contrasts a zero response (no symptoms) with the three others combined, that is, compares the frequencies of zero score with the sum of frequencies of non-zero scores for a given symptom in Table I. These deviance values are even smaller than those for the continuation ratio (third column), demonstrating decreased precision in the estimation of treatment effect, resulting from the loss of information due to collapsing the tables. In other words, likelihood-based confidence or credibility intervals for treatment effects, derived from the log-likelihood or deviance, are wider for the logistic than for the continuation ratio model. (Asymptotic standard errors are not shown because they are generally not reliable for such sparse binary data.)

In Table III, the parameter estimates from the continuation ratio and logistic models are compared. We can see that they are reasonably similar and that the two treatments have similar effects. However, if treatments A and B are combined, the deviances (not shown) increase, by up to 18 for some responses, with only symptoms 4 and 5 showing no difference between treatments. For symptom 6, the relationship between treatments A and B is reversed between the two models (B has a larger effect when estimated using the logit model, whereas A has the larger effect when estimated by the continuation ratio model). Similar differences for symptom 6 can also be seen in the original data in Table I: the odds of a 1, 2 or 3 versus 0 are 1824/1882 for A and 1891/1796 for B, that is, B is the larger. However, A has the larger effect when estimated using the continuation ratio model due to its relatively large count for score 3. Thus, the main difference between the models seems to be the increased precision of estimation of treatment effect in the ordinal model

Table IV. Dependence of response on day in the continuation ratio model. Changes in deviance have one d.f.

Symptom	Independence		Markov chain	
	Slope	Change in deviance	Slope	Change in deviance
1	0·0000	0·0	0·0040	2·2
2	0·0051	5·5	0·0019	0·4
3	−0·0154	53·3	−0·0050	3·8
4	−0·0151	48·1	−0·0012	0·2
5	−0·0036	2·9	0·0033	1·4
6	−0·0014	0·5	−0·0001	0·0

as compared to the logistic, as indicated by the larger changes in deviance in the former in Table II, discussed above. However, for these data, with such large deviances for treatment effects in both models, this is not an important consideration.

We could continue in this way, creating larger contingency tables. For example, we could include the day on which each response is recorded. However, this would already yield a table with 336 ($= 3 \times 4 \times 28$) cells. Thus, it is preferable to proceed by using the raw data. Obviously, the preceding results can also be obtained directly from the same models applied to the raw data. Contingency tables were used above for ease of visual presentation.

We shall concentrate on the continuation ratio model. Introduction of a linear trend in time gives a decrease in the frequency of serious symptoms over time for symptoms 3 and 4, with some indication of an increase for symptom 2, as shown in the second and third columns of Table IV. For example, the slope of $-0{\cdot}0154$ for response 3 corresponds to the odds of being one point higher on the ordinal scale decreasing to 0·9 ($= \exp(7 \times -0{\cdot}0154)$) of what it had been one week before. For symptom 3, there is some slight indication of departure from linearity, as indicated by fitting a factor variable in time (not shown in the table).

3.2. Markov chain models

Up until now, we have analysed the data as if the 28 responses of a patient on consecutive days were independent. Time dependence or serial correlation can easily be introduced by using the response on the previous day as an explanatory variable. Of course, we could also use previous values for all responses, but we shall not pursue this here. With such lagged variables, we lose the first observation. If a response is missing in the middle of the 28 days, we lose not only that response but the next one after it as well.

The deviances for treatment are given in the last column of Table II. With dependence on previous response included in the model, the deviances are much reduced with respect to all previous models because this dependence among observations on a patient has now been taken into account. In other words, the precision of the estimate of treatment effect is reduced; again, the likelihood-based confidence or credibility intervals will be wider. The estimates themselves are given in Table III. These are treatment effects given the state on the previous day. We see that they also are much smaller than those from the independence models. (Markov dependence can also be added to the logistic model used above, but the precision of treatment effects is still less

Table V. Changes in deviance for adding interactions with treatment to the Markov chain model

Symptom	Subtable 4 d.f.	Day 3 d.f.	Previous response 6 d.f.
1	16	3	45
2	16	0	29
3	7	8	27
4	3	0	6
5	11	2	13
6	38	0	34

than that for the ordinal Markov model. There is no real advantage to doing this, with the attendant loss of information, because of the ease of fitting the ordinal Markov model.)

Once previous response is entered into the model, the linear trend in response can be removed (in those cases where it was present) as can be seen in the last two columns of Table IV; in no case is the change in deviance for eliminating it (last column) significant at the 5 per cent level. In addition, there now appears no longer to be any significant difference between the treatments A and B for any symptom.

Our further element remains to be considered. In Table II, we noticed a considerable lack of fit of the continuation ratio model in the simple case when only treatment was present in the model. We may ask if this persists when additional explanatory variables are present. More generally, we may ask if treatment interacts with subtable, day, or response on the previous day. The deviances for adding these interactions individually to the previous Markov chain model are given in Table V. We see that they are generally much smaller than those for treatment differences in Table II, except for symptom 6. No general pattern appears in the parameter values.

In conclusion, either treatment A or B provides relief from rhinitis symptoms. The odds of being one category of severity higher under treatment, as compared to placebo, is about $\exp(-0{\cdot}55)/(1+\exp(-0{\cdot}55)) = 0{\cdot}37$ for the first five symptoms (taking the average of the values in the last column of Table III) and somewhat higher for the sixth. When dependence among successive responses of a patient is taken into account in the way just done, the difference between treatment and placebo remains constant over the 28 day observation period for all six symptoms.

Analysis using the aggregate data, as was originally done (described above), does not provide this confirmation. Percentages and median scores could be hiding changing effects over time, whereby, say, treatment A was better early in the period and treatment B later. Although the independence model did provide some indication of a trend over time, this disappeared when we allowed for dependence through the conditional Markov model.

4. DISCUSSION

At present, the biggest handicap in the modelling of non-normal data is the lack of appropriate software to handle frailty. The most common methods require numerical integration which is still impossible with such large data sets. For the present data set, the Markov chain model fits very much better than a random effects model (but this requires specialized software; we used Sabre[11]). Of course, no general conclusions for other data sets can be drawn from this result, but our view is

that a simple Markov model will generally take reasonable account of dependence among repeated ordinal responses.

Another possibility is to use a conditional argument, a generalization of the Rasch model.[12] This involves counting the number of possible combinations of states of a patient and classifying each patient according to his or her specific combination. For large repeated measures data sets, this has a number of drawbacks. For the combinations to have meaning, all patients must be observed the same number of times. If the series are long, there will be a large number of combinations and few patients in each. For longitudinal data, the combination into which a patient is classed depends on the whole series of observations; conditioning on this means that early observations are explained by later ones.

Finally, a fixed effect model might be used, with patients as blocks nested in treatments. For a small data set, this usually yields results similar to the previous methods, although the estimates are not consistent. For large data sets, the problem is more serious, and, in addition, means introducing the equivalent of a factor variable with a large number of levels (about 420 for the rhinitis data). (Unfortunately, even with the new `$eliminate` instruction in GLIM4, this is not practical, because eliminating the nested blocks means that the treatment contrast parameters are not displayed.)

Marginal models also have found considerable appeal. However, again no software suitable for large unbalanced data sets is available. More seriously, most generalized estimating equations (GEE) have no basis in a stochastic model generating the data. Those marginal models with a stochastic basis generally imply a complex and unrealistic dependence structure among observations on a patient.[13]

Thus, the methods suggested here seem to be the most realistic for treatment of large clinical trials involving ordinal response variables. Both continuation ratio models and Markov chains are well known methodologies. Their combination provides a practical and understandable modelling approach to longitudinal ordinal responses.

All of the models were fitted with GLIM4, but any logistic regression package, which allows easy data manipulation to create the extended vectors, should serve as well. Berridge[6] provides appropriate GLIM4 instructions.

REFERENCES

1. Anderson, J. A. 'Regression and ordered categorical variables', *Journal of the Royal Statistical Society*, *Series B*, **46**, 1–30 (1984).
2. Lindsey, J. K. *Modelling Frequency and Count Data*, Oxford University Press, Oxford, 1995.
3. McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society*, *Series B*, **42**, 109–142 (1980).
4. Greenland, S. 'Alternative models for ordinal logistic regression', *Statistics in Medicine*, **13**, 1665–1677 (1994).
5. Ashby, D., Pocock, S. J. and Shaper, A. G. 'Ordered polytomous regression: an example relating serum biochemistry and haematology to alcohol consumption', *Journal of the Royal Statistical Society*, *Series C*, **35**, 289–301 (1986).
6. Berridge, D. M. 'Fitting the continuation ratio model using GLIM4', *in* Fahrmeir, L., Francis, B., Gilchrist, R. and Tutz, G. (eds), *Advances in GLIM and Statistical Modelling*, Springer, Berlin, 1992, pp. 27–33.
7. Berridge, D. M. 'Modelling ordinal recurrent events', *Journal of Statistical Planning and Inference*, **47**, 71–78 (1995).
8. Lindsey, J. K. *Models for Repeated Measurements*, Oxford University Press, Oxford, 1993.
9. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. *Discrete Multivariate Analysis*: *Theory and Practice*, MIT Press, Cambridge, 1975.

10. Cox, D. R. *Analysis of Binary Data*, Chapman and Hall, London, 1970.
11. Barry, J., Francis, B. and Davies, R. *Sabre. Software for the Analysis of Binary Recurrent Events*, Centre for Applied Statistics, Lancaster, 1990.
12. Conaway, M. R. 'Analysis of repeated measurements with conditional likelihood methods', *Journal of the American Statistical Association*, **84**, 53–62 (1989).
13. Azzalini, A. 'Logistic regression for autocorrelated data with application to repeated measures', *Biometrika*, **81**, 767–775 (1994).

