FITTING TRANSITION MODELS TO LONGITUDINAL ORDINAL RESPONSE DATA USING AVAILABLE SOFTWARE

M. Ganjali
Department of Statistics, Shahid Beheshti University, Iran
m-ganjali@sbu.ac.ir

In many areas of medical and social research, there has been an increasing use of repeated ordinal categorical response data in longitudinal studies. Many methods are available to analyze complete and incomplete longitudinal ordinal responses. In this paper a general transition model is presented for analyzing complete and incomplete longitudinal ordinal responses. How one may obtain Maximum Likelihood (ML) estimates for the transition probabilities by existing software is also illustrated. The approach is implemented on a real application. For this data set, two important results are underlined: (1) some transition probabilities may be estimated to be zero and (2) the model for current response, which conditions on previous response may reduce the effects of some covariates that had previously been strongly significant.

INTRODUCTION

Statistical activity starts with a scientific question which has to be answered by scientific methods. Our question of interest in this paper is how an ordinal response changes according to treatment or some time-varying explanatory variables. Answers to questions such as “Does the previous ordinal response affect the current ordinal response?” or “Does knowledge of a previous state reduce the effects of other explanatory variables?” are of interest. To answer such scientific questions we have to collect some data. This collection of data may be done by an observation study or a designed experiment. In each of these, the response of interest has to be observed for each subject repeatedly, at several times. That is why, in health-related and social science applications, we have to learn about longitudinal or panel studies where repeated ordinal response data commonly occur. For example, in such studies, a physician might evaluate patients at baseline and at weekly follow-ups regarding whether a new drug treatment is successful. Another example is where the assessment of side effects of radiation therapy for cancer treatment is recorded on the ordinal scale of ‘no problems’, ‘minor problems’, and ‘severe problems’ for patients who may be followed at regular intervals for some years.

After collecting the data, the first and most important step in learning from the data about the process generating them is exploratory data analysis. This step may lead us to decide which statistical model is the most appropriate to use in order to answer the scientific questions of interest. Questions such as “Does the chosen approach allow us to answer our scientific question in an appropriate manner? or “Does the model fit well?” are also very important to investigate. After assessing goodness of fit of the model, what remains is the sensible interpretation of what the data and the model reveal.

In longitudinal studies, there will be a sequence of responses recorded on each individual. In the current context, we have to take into account not only the fact that responses are ordinal in nature but also the possibility of dependence or correlation between responses given by the same individual. Different models can be used to handle such dependence. Agresti (1999) and Lall et al. (2002) conducted a comprehensive survey of models for ordered categorical data, in which the need for model interpretation is emphasized. One possibility is marginal modelling, which can be used to study the population-average pattern or trend over time (Ten Have et al., 1996; Kim, 1995; Liang et al., 1992). A second possibility is conditional random effects modeling which makes inferences about variability between subjects. In this approach, individual behaviour is often of scientific interest (Harvile & Mee, 1984; Verbeke & Lesaffre, 1996; Tutz and Hennevogl, 1996; Verbeke & Molenberghs, 1997; Tutz, 2005). However, both of these approaches are generally appropriate for longer sequences of measurements than those examined here. These approaches are not appropriate for the primary question of interest here which is how transitions from one level of response to another are made between consecutive time points. For such a scientific question, a more appropriate approach would be to use Markov (transition) models (see Garber, 1989; Francom et al., 1989; Rezaee & Ganjali, 2009, Rezaee et al., 2009) where we can consider the effect of previous response on current response. For reviews of transition and other models for

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

longitudinal ordinal data, see McCullagh (1980), Agresti (2002), Diggle et al. (2002) and Song (2007).

In this paper, the use of a first order transition model for repeated ordinal responses is presented. It is shown how to use existing software to fit the model. The insomnia data are introduced and the initial exploratory data analysis is presented. This leads us to make two important points about transition probabilities. Then, the model and the likelihood are given and the results of applying the model to the insomnia data are discussed. Finally, conclusions are presented.

INSOMNIA DATA

The data in Table 1, extracted from Francom et al. (1989), show the results of a randomized, double-blind clinical trial comparing an active drug with a placebo in 239 patients who have insomnia problems. The measure of interest is the patient's response to the question ‘How quickly did you fall asleep after going to bed?’ The response was categorized as: ‘less than 20 minutes’; ‘20 to 30 minutes’; ‘more than 30 and less than or equal to 60 minutes’; and ‘greater than 60 minutes’. Patients were asked this question after a one week placebo washout period (baseline measurement) and following a two-week treatment period.

Table 1. Time to falling asleep obtained from the question, ‘How quickly did you fall asleep?’ in grouped minutes (follow-up response, Y_2, by treatment and initial response, Y_1, observed counts and row percentages)

		Follow-up (Y_2)				
Treatment	Initial (Y_1)	<20	20-30	30-60	>60	Total
Active	<20	7 58.3%	4 33.3%	1 8.3%	0 0.0%	12 100.0%
	20-30	11 55.0%	5 25.0%	2 10.0%	2 10.0%	20 100.0%
	30-60	13 32.5%	23 57.5%	3 7.5%	1 2.5%	40 100.0%
	>60	9 19.1%	17 36.2%	13 27.7%	8 17.0%	47 100.0%
Placebo	<20	7 50.0%	4 28.6%	2 14.3%	1 7.1%	14 100.0%
	20-30	14 70.0%	5 25.0%	1 5.0%	0 0.0%	20 100.0%
	30-60	6 17.1%	9 25.7%	18 51.4%	2 5.7%	35 100.0%
	>60	4 7.8%	11 21.6%	14 27.5%	22 43.1%	51 100.0%

For an exploratory analysis of these data, one has to think about how to answer questions like (1) What kind of association measure should we use to calculate the association between two ordinal responses? (2) Is there any correlation between the two responses? (3) If there is any correlation between the two responses, is the correlation the same in each of the two treatments?

The answers to these questions are important since, if there is no correlation between responses, one may fit separate marginal models to each response to examine the treatment effect. For the insomnia data, the answer to question (1) is the gamma association measure (Goodman & Kruskal, 1954). This measure is the difference between the concordant and the discordant pairs divided by the sum of the concordant and the discordant pairs and it takes values in the range [-1,1]. In answer to question (2), the estimate of gamma for the two responses is 0.546 (S.E. =0.063, P-value=0.000) which shows a strong association between the two responses and consequently any statistical analysis of these data should take this association into account. Partial gamma (gamma for a specific treatment) may be used to answer question (3). This is 0.461 (S.E. =0.105, P-value=0.000) for the active drug and 0.635 (S.E. =0.075, P-value=0.000) for the placebo. As the association between the two responses is not the same for the two treatments, we need to choose a longitudinal approach which is able to take into account the fact that the covariance structure of the responses is dependent on treatment.

Table 2 displays empirical marginal distributions for the initial and follow-up responses for the two treatments.

Table 2. Empirical marginal distributions of initial and follow-up responses for two treatments

		Response category			
Response	Treatment	<20	20-30	30-60	>60
Initial	Active	0.101	0.168	0.336	0.395
	Placebo	0.117	0.167	0.292	0.425
Follow-up	Active	0.336	0.412	0.160	0.092
	Placebo	0.258	0.242	0.292	0.208

From Table 2, we can conclude that, initially, the two groups had similar distributions, but at the follow-up, those patients on the active treatment tended to fall asleep more quickly.

Let us give an example which highlights the difference between the two treatments. The sample probability of a patient who initially took more than 60 minutes to fall asleep but who, having taken the active drug, took less than or equal to 30 minutes to fall asleep by the follow-up is 0.553 (see Table 1). The same probability is just 0.294 for a patient on the placebo. This shows the level of improvement on using the active drug for an insomnia patient initially required more than 60 minutes falling asleep. An important question is whether this significant difference between the two treatments on follow-up response remains the same for all initial response levels. The model in the next section can answer this question using existing software.

As we have seen the treatment effect may be reduced if we condition on the value of previous response. Another important point in analyzing the insomnia data using a transition model is that some of the transition probabilities may be estimated to be zero. Table 1 confirms this fact by showing the zero empirical transition probability of $Y_2 \,|Y_1 < 20, Treatment = Active$. When the treatment is the active drug, there is no observation with $Y_2 > 60$ given $Y_1 < 20$.

ORDERED TRANSITION MODEL USING CUMULATIVE LOGITS

The best approach to analyzing longitudinal data is to start with marginal modelling of responses where one assumes independence between responses. Results of this initial marginal model can be compared with a subsequent model which takes into account the correlation between responses.

Perhaps the most popular method for the analysis of univariate ordered categorical data is that based upon the cumulative logit regression model which was first proposed by Snell (1964) and further generalized by McCullagh (1980) to allow link functions other than the logit. The model estimates the effects of explanatory variables on the log odds of selecting lower, rather than higher, response categories. This model for a univariate response can be expressed in terms of a latent variable model of the form:

$$y_i^* = \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i$$

which gives

$$y_i = \begin{cases} 1 & if\ y_i^* \le \alpha_1 \\ 2 & if\ \alpha_1 < y_i^* \le \alpha_2 \\ 3 & if\ \alpha_2 < y_i^* \le \alpha_3 \\ \vdots & \vdots \\ J-1 & if\ \alpha_{J-2} < y_i^* \le \alpha_{J-1} \\ J & if\ \alpha_{J-1} < y_i^* \end{cases}$$

If we assume a logistic distribution for the error term (ε_i) this gives the logistic model of the form

$$\ln[\frac{pr(Y_i \le b;\alpha,\beta)}{pr(Y_i > b;\alpha,\beta)}] = \alpha_b - \sum_{k=1}^{K}\beta_k X_{ik} \qquad b = 1,\ldots,J-1. \qquad (1)$$

In the above equations, Y_i (with observed value y_i and latent variable y_i^*) is the response of the i-th individual, X_{ik} is the k-th explanatory variable for the i-th individual, *J* is the number of ordered categories of the dependent variable, α_b 's are the partition-specific intercepts (cut-points) indicating the logarithms of odds of selecting lower, rather than higher, categories when all explanatory variables are set to zero, $\alpha = (\alpha_1,\cdots,\alpha_{J-1})$ is the vector of cut-point parameters in which $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{J\text{-}1}$, $\beta = (\beta_1,\cdots,\beta_K)$ is the vector of regression coefficients for the explanatory variables and K is the number of explanatory variables. In equation (1), as the linear predictor, $\sum_{k=1}^{K}\beta_k X_{ik}$, is subtracted from, rather than added to, the intercepts, a positive coefficient indicates increased likelihood of selecting a higher response category. The cumulative logit model assumes that the effects of different explanatory variables are fixed across all (J-1) partitions of the ordinal response. This model can be implemented readily in software such as SPSS (ordinal regression) and STATA (ordered logit regression).

In transition models, the probability distribution of the outcome of individual *i* at time t , Y_{it} is a function of the individual's covariates at time t , X_{it} , and the individual's outcome history $Y_{i1},\ldots,Y_{it-1}$, t>1. Such models are appropriate when there is a natural sequencing of the responses, as in longitudinal studies. Examples of this approach include Bonney (1987), the binary Markov model of Muenz and Rubenstein (1985) and Kalbfleisch and Lawless (1985).

The form of the transition model for T response variables with missing responses (which will be applied to the insomnia data where $T = 2$) is:

$$\ln[\frac{pr(Y_{\mathrm{i}1} \le b;\alpha_1,\beta_1)}{pr(Y_{\mathrm{i}1} > b;\alpha_1,\beta_1)}] = \alpha_{b1} - \sum_{k=1}^{K}\beta_{k1}X_{i1k} \qquad b = 1,\ldots,J\text{-}1$$

$$(2)$$

$$\ln[\frac{pr(Y_{it} \le b \,|\, Y_{it-1} = a;\alpha_{at},\beta_{at})}{pr(Y_{it} > b \,|\, Y_{it-1} = a;\alpha_{at},\beta_{at})}] = \alpha_{abt} - \sum_{k=1}^{K}\beta_{akt}X_{itk} \quad b = 1,\ldots,J-1 \quad a = 1,\ldots,J \quad t = 2,\ldots,\mathrm{T}$$

where $Y_{i1}$and Y_{it} for $t = 2,\ldots,T$ are the responses given by the *i*-th individual at the initial time and at (T-1) follow-up times, respectively. The vectors $\alpha_1 = (\alpha_{11},\ldots,\alpha_{J-1,1})$ and $\beta_1 = (\beta_{11},\ldots,\beta_{J-1,1})$ are defined as before, $a_{at} = (a_{a1t},\ldots,a_{a(J-1)t})$ is the vector of cut-point parameters for $a = 1,\ldots,J$, $t = 2,\ldots,T$ and $b_{at} = (b_{a1t},\ldots,b_{aKt})$ is the vector of regression coefficients for the explanatory variables for $a = 1,\ldots,J$ and $t = 2,\ldots,T$. This model forms a general model which includes interactions of the previous response with all covariates and hence the response correlation structure is dependent on the covariates. Using the transition model, the likelihood for two time points with complete data on the initial responses and possible randomly missing responses at time 2 is:

$$L = \prod_{i=1}^{n} P(Y_{i1} = y_{i1};\alpha,\beta)\prod_{i=1}^{m} P(Y_{i2} = y_{i2} \big| Y_{i1} = y_{i1};;\alpha_{a2},\beta_{a2})$$

$$= \prod_{i=1}^{n} P(Y_{i1} = y_{i1};\alpha,\beta)\prod_{a_1}^{J}\prod_{i\in A_1} P(Y_{i2} = y_{i2} \big| Y_{i1} = a_1;;\alpha_{a2},\beta_{a2})$$

where $A_1 = \{i : Y_{i1} = a_1, i = 1,...,m\}$, m is the number of individuals without any missing data and n is the total number of individuals. For the insomnia data, there are no missing values and hence $m = n$.

The vectors of parameters, in system of equations (2), are assumed to be distinct at different times and hence parameter estimation can be carried out using existing software, such as SPSS, as follows:

(1) modeling the probability of Y_1 by going to analyze of SPSS and then using ordinal regression,
(2) separately modeling the conditional probability of Y_2 given Y_1 at each level of Y_1 by going to data, and selecting the part of data with the chosen level of Y_1 and then using analyze and ordinal regression,
(3) continuing in the same way until separately modeling the conditional probability of Y_T given Y_{T-1} at each level of Y_{T-1}.

RESULTS OF APPLYING TRANSITION MODEL TO INSOMNIA DATA

Results from the marginal model for the initial response (not reported here) show that there is no significant effect of treatment on the cumulative probability of initial response. Results of the conditional components of the transition model are given in Table 3.

Table 3. Results for the transition model where Y_2 is the follow-up response (parameters significant at the 5% level are highlighted in **bold**)

	$Y_2\|\text{initial} < 20$		$Y_2\|20 < \text{initial} \le 30$		$Y_2\|30 < \text{initial} \le 60$		$Y_2\|\text{initial} > 60$	
Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
α_1	-.089	0.522	0.909	0.490	**-2.235**	0.436	**-.561**	0.385
α_2	**1.478**	0.621	**2.377**	0.623	-0.083	0.330	**-.926**	0.288
α_3	**3.007**	1.058	**3.397**	0.833	**2.592**	0.610	0.317	0.272
Treatment baseline: placebo								
Active	-.507	0.765	0.792	0.651	**-1.705**	0.477	**-.161**	0.381

Now, we have gained more insight into the process generating the data. In Table 4, for different values of the initial response, the parameters α_j for j=1,2,3 are intercepts indicating the log-odds of lower, rather than higher, times to falling asleep when patients use the placebo. For example, when the initial time to falling asleep is less than 20 minutes, for follow-up response log-odds of less than 20 rather than time more than 20 is -0.089+0.507 =0.418, or the odds are 1.519 when patients use the active drug. These log-odds, when the initial response is more than 60 minutes, is -0.561+0.161=-0.400, or the odds are 0.670.

When the initial response is ‘less than 20’ or ‘20-30’ there is no significant effect of the active drug. But, for an initial value of ‘30-60’ or ‘more than 60’ there is a positive effect of the active drug. This means that the drug is less likely to be effective for patients who previously took less than 30 minutes to fall asleep and so knowledge of the initial response may inform practitioners when considering prescribing this particular treatment.

CONCLUSIONS

In this paper, exploratory analyses of the insomnia data have revealed that: (a) some transition probabilities were zero and (b) conditioning on previous response has reduced the effect of treatment on current response. After initial exploratory data analysis, we used a Markov (transition) model for longitudinal ordinal response data. Existing software was used to estimate model parameters for an insomnia patient's ordinal response to the question ‘How quickly did you fall asleep?’. For these data, we found that the effectiveness of the active drug at follow-up depends on the initial response. The longer the time it took to fall asleep, the more likely the effect of the active drug was to be significant. One important step we have not discussed here is assessing the goodness of fit of the chosen model. Nagelkerke’s pseudo R^2 may be used to investigate the goodness of fit of an ordinal model (for details see Rezaee & Ganjali, 2009).

REFERENCES

Agresti, A. (1999). Modeling ordered categorical data: recent advances and future challenges. *Statist. Med., 18,* 2191-2207.

Agresti, A. (2002). *Analysis of categorical data.* John Wiley and Sons, New York.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics, 43,* 951-973.

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of longitudinal data.* Oxford: University Press.

Francom, S. F., Chuang-Stein, C., & Landis, J. R. (1989). A log-linear model for ordinal data to characterize differential change among treatments. *Statist. Med., 8,* 571-582.

Garber A. M. (1989). A discrete-time model of the acquisition of antibiotic-resistant infections in hospitalized patients. *Biometrics, 45,* 797-816.

Goodman, L. A, & Kruskal ,W. H. (1954). Measures of association for cross-classification. *J. Am. Statist. Ass., 49,* 732-804.

Harvile, D. A., & Mee, R. W. (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics, 40,* 393-408.

Kalbfleisch, J. D., & Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *J. Am. Statist. Assoc., 80,* 863-871.

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statist. Med., 14,* 1341-1352.

Lall, R., Campbell, M. J., Walters, S. J., & Morgan, K. A. (2002). Review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research, 11,* 49-67.

Liang, K. Y., Zeger, S. L., & Qaqish, B. F. (1992). Multivariate regression analyzes for categorical data (with discussion). *J. R. Statist. Soc. B., 54,* 3-40.

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

Muenz, L. R., & Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequence. *Biometrics, 41,* 91-101.

Rezaee, Z., Ganjali, M., & Berridge, D. (2009). A Transition Model for Ordinal Response Data with Random Dropout: An Application to the Fluvoxamine Data. *Journal of Biopharmaceutical Statistics, 19*(4), 658-671.

Rezaee, Z., & Ganjali, M. (2009). Testing Homogeneity in Markov Models for Analyzing Longitudinal Ordinal Response Data with Random Dropout. *Journal of Statistical Theory and Applications, 8*(2), 125-139.

Snell, E. J. (1964). A scaling procedure for ordered categorical data. *Biometrics, 2,* 592-607.

Song, P. X. K. (2007). *Correlated Data Analysis*. Springer.

Ten Have, T. R., Landis, J. R., & Hartzel, J. (1996). Population-average and cluster-specific models for clustered ordinal response data. *Statist. Med., 15,* 2573-2588.

Tutz, G. (2005). Modeling of repeated ordered measurements by isotonic sequential regression. *Statistical Modeling, 5,* 269-287.

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis, 22,* 537-557.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Am. Statist. Ass., 91,* 217-221.

Verbeke, G., & Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach.* Springer.