Biostat/Stat 571
Winter 2010
P. Heagerty

Exercise #2
Due: January 20, 2010

Reading:
- McCullagh & Nelder, Chapter 5, "Polytomous Data"
  - Ananth & Kleinbaum (1997) *International Journal of Epidemiology*
  - Fahrmeir & Tutz, sections 3.1, 3.3, 3.4 (multinomial models)

1. The data `tenhave.data` on the class web page are taken from TenHave and Uttal (1994). The data are from a psychology experiment in which children were shown a map giving the location of a hidden toy, and then allowed up to three chances to successfully find the object. The goal of the experiment was to compare a group of children presented with a correctly oriented map to a second group that was shown a rotated map. Are children presented with a rotated map able to compensate and find the hidden object as well as the control children? By repeating the test (with a new maze/ location) ten times per subject, the investigators were able to evaluate whether children presented with a rotated map are able to adjust and improve performance over time.

(a) Display the distribution of the response over time for each treatment group. Comment on the evidence for a "learning effect".

(b) Consider the outcome from the first maze (trial). Construct a 2×4 table that displays the distribution of the number of attempts (4 categories) by treatment group.

(c) For trial=1 fit a logistic regression to the binary response $Y_{i1}(1) = \mathbf{1}(T_{i1} > 1)$ using tx group as the predictor. Interpret the regression results. How does this summarize the table in 1(b)?

(d) For trial=1 fit a logistic regression to the binary response $Y_{i1}(2) = \mathbf{1}(T_{i1} > 2)$ using tx group as the predictor. Interpret the regression results. How does this summarize the table in 1(b)?

(e) For trial=1 fit a logistic regression to the binary response $Y_{i1}(3) = \mathbf{1}(T_{i1} > 3)$ using tx group as the predictor. Interpret the regression results. How does this summarize the table in 1(b)?

(f) Write the proportional odds model for the ordinal response T_{i1} using tx group as the predictor. How do the parameters in this model relate to the parameters given in 1(c)-1(e)? Use the maximum likelihood **R/S+** function provided on the class web page to fit the proportional odds model. Compare the inference regarding treatment obtained with this model to that obtained in 1(c)-1(e).

(g) Evaluate whether the proportional odds model is reasonable for these data.

(h) Now consider the outcome from trial=8. Construct a 2×4 table that displays the distribution of the number of attempts (4 categories) by treatment group. Compare this table to that for trial=1. Is there evidence for a "learning effect"?

(i) Write the proportional odds model for the ordinal response T_{i8} using tx group as the predictor.

How do the estimated parameters from trial=8 compare to the parameters for trial=1? Interpret the evidence for a "learning effect".

(j) From 1(a)-1(g) we can see that the proportional odds model is equivalent to simultaneous logistic regression models for $Y_{i1}(1) = \mathbf{1}(T_{i1} > 1)$, $Y_{i1}(2) = \mathbf{1}(T_{i1} > 2)$ and $Y_{i1}(3) = \mathbf{1}(T_{i1} > 3)$ (for trial=1). Is the maximum likelihood estimate of $\boldsymbol{\beta} = (\beta_{(0,1)}, \beta_{(0,2)}, \beta_{(0,3)}, \beta_1)$ equivalent to the estimate obtained by pooling ($Y_{i1}(1), Y_{i1}(2), Y_{i1}(3)$) and performing logistic regression? Justify your answer by representing, and comparing the score equations (estimating functions) used for maximum likelihood estimation and by "pooling logistic regressions". Can you comment on the validity of simply pooling the logistic regressions?

2. The experimental outcome for the `tenhave data` is naturally considered a "continuation" outcome. A child will attempt the maze a second time only if they are unsuccessful on their first attempt. Similarly, a child will attempt the maze a third time only if unsuccessful on their first and second attempt. This type of outcome may also be thought of as a discrete-time outcome, T_{ij} defines the length of time required for success (measured in attempts). The coding $T_{ij} = 4$ simply means that all three attempts failed.

(a) In lecture we showed how the score equations for the proportional odds model can be derived via a linear transformation of the score equations for the multinomial model using simple category indicators. That is, we showed that there is an $\boldsymbol{L}$ such that $\boldsymbol{L}(\boldsymbol{Y} - \boldsymbol{\pi}) = (\boldsymbol{Y}^* - \boldsymbol{\gamma})$ where $\boldsymbol{Y}$ is the vector of category indicators, $E(\boldsymbol{Y}) = \boldsymbol{\pi}$, and $\boldsymbol{Y}^*$ is the vector of cumulative indicators, $E(\boldsymbol{Y}^*) = \boldsymbol{\gamma}$. For the continuation ratio model show that there is a matrix, $\boldsymbol{B}$ (that may depend on π's), such that $\boldsymbol{B}(\boldsymbol{Y} - \boldsymbol{\pi}) = (\boldsymbol{Y} - \boldsymbol{\mu} \cdot \boldsymbol{H})$ where $\boldsymbol{H} = \text{vec}(H_j)$, $H_j = 1 - \sum_{k=1}^{j-1} Y_k$ is the "at-risk" indicator, and $\boldsymbol{\mu} = \text{vec}(\mu_j)$, $\mu_j = E(Y_j \mid H_j)$.

(b) What does the transformation $\boldsymbol{B}$ do to the covariance matrix of $\boldsymbol{Y}$? Justify your answer.

(c) Again consider the first trial (trial=1). Formulate and fit a continuation ratio logit model using single odds ratio for treatment group. Formally check that a common (single) tx odds ratio is appropriate (versus three parameters for tx).

(d) Now consider trial=8. Formulate and fit a continuation ratio logit model using a single odds ratio for treatment group. Formally check that a common (single) tx odds ratio is appropriate (versus three parameters for tx).

(e) Summarize the evidence for a learning effect and compare your results in 2(a)-2(b) with the summaries obtained in question (1).

3. In questions (1) and (2) we performed two separate cross-sectional analyses. That is, we analyzed the treated versus the control group at fixed times. A complete analysis of these data requires methods for longitudinal or correlated categorical data.

(a) Based on the EDA you have for 1(a) what form of mean model might you propose for the response variables $T_{i1}, T_{i2}, \ldots, T_{i10}$ that uses the trial variable $t_j = j$ and treatment group? Give a POM and/or CRM specification for the 10 means. Which parameters address the primary scientific question?

(b) Another method for analyzing repeated measures of longitudinal data is to take the data for each subject, calculate a summary of each subject's data (such as a mean, or a slope for a regression over time), and then compare the treatment groups using the summary statistics. Suggest a summary method for T_{ij} that might be used to compare treatment groups regarding evidence for a "learning effect".