Paper SA09-2012

Models for Ordinal Response Data

Robin High, University of Nebraska Medical Center, Omaha, NE

Abstract

Ordinal responses are commonly collected in medical studies, yet the various types of analyses possible with these data with SAS procedures are not well known. Interpretation of ordinal logistic regression output requires consideration of how data are coded and the computing options or formats invoked. Although the cumulative and generalized logit models can easily be run with statistical procedures such as LOGISTIC, CATMOD, GENMOD, or GLIMMIX, these and several other models can also be developed through programming statements entered in the NLMIXED procedure. This talk provides a survey of the various models for ordinal data that NLMIXED can run such as the proportional odds and partial proportional odds models, adjacent logits, continuation ratio, and the stereotype model. Odds ratios and predicted values from these models also require programming statements that help to interpret the results and to test the goodness of fit. Comparisons with odds ratios computed from the more common ordinal response models will also be given. Implementing these models assumes a background with categorical data analysis including maximum likelihood equations and computing odds ratios with binary data.

Ordinal Data

It is not recommended that existing numerical data be recoded into ordinal categories; working with actual numerical data almost always preferable to avoid unnecessary measurement error. However, it is relatively common to not be able to measure variables with a numerical value. Instead, the data are coded into distinct categories where their inherent order is meaningful. Ordinal data analysis in this paper is defined as methods to evaluate responses that have a well-defined order with 3, 4, or perhaps as many as 5 or 6 possible values. Working with six or more levels of an ordinal response is less common and perhaps more difficult to work with ordinal data modeling procedures.

In biomedical studies, ordinal responses commonly occur when a subject progresses gradually through the levels of a response that is difficult to quantify, such as specific durations or severity levels of a disease. Other common examples are opinions collected at one point in time that are not expected to change suddenly and are measured on a Likert scale, such as having a choice of five responses ranging from total disagreement to total agreement. Scales, the sum of several Likert responses producing a reasonably large alpha coefficient, often have an adequate range and nearly symmetric distribution of values so that the normality condition is a reasonable choice. Almost any single response variable where the underlying interpretation has an increasing or decreasing order of levels which are not easy to quantify would qualify as ordinal data. Though often coded numerically (e.g., as increasing non-negative integers 1, 2, 3, ..) the actual values of the integers do not imply that a difference between any pair of them has a meaningful quantitative interpretation. For data analysis statistical methods designed for continuous data based on a normal distribution of residuals may be routinely invoked by assuming these conditions also apply for ordinal responses. The sample size, range of values, and the central limit theorem all work together and in some cases may make a normal theory model a reasonable choice. However, the purpose of this paper is to explore alternatives, in particular how SAS/STAT software may be applied to ordinal response data of the types described in Chapters 3 and 4 of "Analysis of Ordinal Data, " 2nd Ed. by Agresti.

SAS Procedures for Analyzing Ordinal Data

SAS/STAT procedures with capabilities for analyzing ordinal response data with one or more covariates are LOGISTIC, GENMOD, and GLIMMIX (among a few others). For ease of interpretation the odds ratios will be the desired output which also gives predicted probabilities of the response for stated values of the covariates. With these three procedures, the cumulative logit (proportional odds) and generalized logit models are available by specifying link=clogit or link=glogit on the MODEL statement. The first model represents the most general situation, the proportional odds, and the second, the most specific, the generalized logit (designed for nominal data, which usually ignores ordinal nature of the data). PROC CATMOD also is designed for categorical response models, but generally works with weighted least squares, not maximum likelihood and is optimized for use with categorical explanatory data (numerical explanatory variables are entered through a DIRECT statement), and for some is cumbersome and difficult to apply. Methods to analyze ordinal response models that allow computation of odds ratios in between these two extremes are not well known or not readily available. These types of ordinal data models include the partial proportional odds, the adjacent category logit, continuation ratio, and stereotype models.

Since maximum likelihood estimation (MLE) techniques are often a preferred method of estimation, the ideal condition of sufficiently large sample sizes will be assumed for application of these models. For MLEs to work reasonably well, the majority of combinations of categorical factors in a contingency table should have 10 or more counts for most of the cells and even larger counts are desirable for the study to be mathematically sound.

When coded as single digit integers, the actual text value of the response can easily be attached with a format. When working with SAS procedures and formats it is also necessary to apply the order=internal option on the PROC statement so that the inherent ordering of the ordinal responses remains as required for the analysis interpretation.

Example

As a simple introduction to modeling ordinal categorical data, four levels of an ordered response variable will be evaluated with one categorical explanatory variable having two levels. The data come from a paper by Ananth and Kleinbaum concerning a clinical trial of a single-dose, post-operative analgesic drug.

```
---------------------------------------------------------------------------
|            |    Rating of the drugs     |      |      |      |      |      |
|            |----------------------------|      |      |      |      |      |
|            |      |      |      | VGd/ |      |      |      |      | Very |
|            | Poor | Fair | Good | Excl | All  | Poor | Fair | Good | Good |
|            |------+------+------+------+------+------+------+------+------|
|            |  N   |  N   |  N   |  N   |  N   |Row % |Row % |Row % |Row % |
|------------+------+------+------+------+------+------+------+------+------|
|Drug        |      |      |      |      |      |      |      |      |      |
|C15 & C60   |    17|    18|    20|     5|    60|  28.3|  30.0|  33.3|   8.3|
|Z100 & EC4  |    10|     4|    13|    34|    61|  16.4|   6.6|  21.3|  55.7|
---------------------------------------------------------------------------
```

The SAS data set that produces this table consists of 8 records defined by the 4 responses and 2 levels of the treatment drug. The treatment variable is coded as drug=1 for “C15 & C60” and drug=2 for Z100 & EC4. The four ordinal ratings are coded as y= 1 (Poor), =2 (Fair), =3 (Good), and =4 (VGd/Excl) (the fourth and fifth levels were combined due to sparse cell counts). With discrete data, the cell frequencies printed in the table are more efficiently stored with a variable called count, though the results would be the same when working with individual responses.

Generalized Logit Model

A generalized or baseline logit treats the response as nominal (the ordered nature of the response categories is not maintained) and provides a perfect fit to these data. the results are the same as computing individual 2x2 tables for Poor, Fair, and Good compared with VGd/Excl as the reference level. The statements for a generalized logit model with PROC LOGISTIC produce the following maximum likelihood estimates and odds ratios:

```
PROC LOGISTIC DATA=cltr order=internal;
CLASS drug / PARAM=ref;
FREQ count;
MODEL y = drug / link= glogit;
ODDSRATIOS drug / cl=pl;
FORMAT drug drg. rsp yy. ;
RUN;
```

```
Parameter              Rating   Estimate   Odds Ratios
Intercept               Poor    -1.2238
Intercept               Fair    -2.1401
Intercept               Good    -0.9614
drug      C15 & C60     Poor     2.4475     11.56  = (17/5) / (10/34)
drug      C15 & C60     Fair     3.4210     30.60  = (18/5) / (4/34)
drug      C15 & C60     Good     2.3477     10.46  = (20/5) / (13/34)
```

The entries in the column for odds ratios are computed from the Estimate column, EXP(estimate), or can be derived directly from the respective cell counts in the table. Since a format is applied to both the response and explanatory data, it is important to enter the order=internal to maintain the coded order, not alphabetical of the formatted values.

PROC NLMIXED

This paper will focus on how statements from the SAS procedure NLMIXED are written to estimate several types of ordinal response models. Other SAS procedures such as NLIN or MODEL may also be possible to estimate them, though the necessary components for maximum likelihood estimation are readily available in NLMIXED. The general layout of statements for NLMIXED with a four level ordinal response variable includes the following statements:

```
PROC NLMIXED DATA=indat;
PARMS < enter initial values of parameters >;

* Statements for the three linear predictors (eta1, eta2, eta3);
* Statements to compute four response probabilities (p1, p2, p3, p4);

* Enter one of these Maximum Likelihood Equations (with 4 response levels);

lk = (p1**(y=1)) * (p2**(y=2)) * (p3**(y=3)) * (p4**(y=4)); * ascending;
lk = (p4**(y=1)) * (p3**(y=2)) * (p2**(y=3)) * (p1**(y=4)); * descending;

* Compute loglikelihood;
IF (lk > 1e-8) then llk = log(lk); else llk=-1e100; * computational safety ;

* Estimate model;
MODEL y ~ general(llk);

REPLICATE count;  * for categorical data entered as counts;
ESTIMATE statements for odds ratios;
PREDICT statements for predicted probabilities;
RUN;
```

Model Fit

The model fit based on comparing predicted counts with the actual values can be obtained from the output data sets derived from the PREDICT statements, one for each level of the response. Each file contains a variable called pred for its predicted probability of the respective values of the response. Append all files together, sort by drug and response levels, and then multiply the predicted value, pred, by the respective row total for each level of the explanatory variable (i.e., merge in the row totals for drug computed with PROC FREQ). For a categorical explanatory variable both the Pearson and log-likelihood chi-square statistics can then be accumulated across all records:

```
pearsonchisq + ((count - pred_c)**2)/pred_c;
loglikechisq + (2*count*(LOG(count/pred_c)));
```

Significance (e.g., goodness of fit) is determined by comparing the computed values to the given critical value for the degrees of freedom.

Cumulative Logit Model

The cumulative logit model can be fit with PROC LOGISTIC by entering link=clogit in the MODEL statement. The PROC NLMIXED statements listed below also estimate the cumulative logit or proportional odds model and is a helpful starting point to begin. In particular, to maintain ordinality, note the starting values of the PARMS and the construction of the statements for the linear predictors:

```
PROC NLMIXED DATA=cltr ;
PARMS Int_1 -2 Int_2 -1 Int_3 1 _d0 .1 ;

* linear predictors;
eta1 = Int_1 + _d0*(drug=1); * drug=2 (bottom row) as reference;
eta2 = Int_2 + _d0*(drug=1);
eta3 = Int_3 + _d0*(drug=1);

* ordered logit model cumulative probabilities;
cp1 = 1 / (1 + exp(-eta1));
cp2 = 1 / (1 + exp(-eta2));
cp3 = 1 / (1 + exp(-eta3));

p1 = cp1;
p2 = cp2-cp1;
p3 = cp3-cp2;
p4 = 1-cp3;

lk = (p1**(y=1)) * (p2**(y=2)) * (p3**(y=3)) * (p4**(y=4)); * ascending order ;
IF (lk > 1e-8) then llk = log(lk); else llk=-1e100;
```

```
MODEL y ~ general(llk);
REPLICATE count;
ESTIMATE ‘Odds Ratio: C15 & C60” vs Z100 & EC4’ EXP(_d0);

PREDICT p1 OUT=p1(where=(y=1)); * contains predicted value called pred ;
PREDICT p2 OUT=p2(where=(y=2));
PREDICT p3 OUT=p3(where=(y=3));
PREDICT p4 OUT=p4(where=(y=4));
RUN;
```

Initial values for the three intercepts are entered in the PARMS statement with an increasing order from negative to positive. These entries are necessary to reflect the ascending order of the cumulative probability for the response (i.e., the order of the response of the variables listed in the likelihood statement). It is also crucial to structure the model and choose starting values for the parameters (on a log scale) that will compute valid probabilities – p1, p2, p3, p4 – that is, all values lie between 0 and 1.

The cumulative odds ratio from this model is 5.88 (also produced with PROC LOGISTIC with link=clogit). It implies that if the four responses are grouped in pairs -- 1 vs 234, 12, vs 34, and 123 vs 4 -- the odds ratio formed from the predicted counts for each of the three aggregated 2 x 2 tables is 5.88. With the POM inference is extended to underlying continuum of responses and is invariant with respect to choice of categories to compare.

```
-------------------------------------------------------------------------
|NLMIXED      |              Rating               |      |      |      |      |
|clogit:      |-----------------------------------|      |      |      |      |
|predicted    |   1   |   2   |   3   |   4   |  1   |  2   |  3   |  4   |
|counts and   |-------+-------+-------+-------+------+------+------+------|
|row percents | Pred  | Pred  | Pred  | Pred  |Row % |Row % |Row % |Row % |
|-------------+-------+-------+-------+-------+------+------+------+------|
|Drug         |       |       |       |       |      |      |      |      |
|C15 & C60    |   20.6|   13.9|   16.2|    9.3|  34.3|  23.2|  27.1|  15.4|
|Z100 & EC4   |    5.0|    6.4|   18.0|   31.6|   8.2|  10.5|  29.5|  51.8|
-------------------------------------------------------------------------
```

```
------------------------------------------ ------------------------------------------
|           | y = 1 | y= 234| odds | odrt | |           |y = 12 | y= 34 | odds | odrt |
|-----------+-------+-------+------+------| |-----------+-------+-------+------+------|
|Drug       |       |       |      |      | |Drug       |       |       |      |      |
|C15 & C60  |   20.6|   39.4| 0.522|      | |C15 & C60  |   34.5|   25.5| 1.353|      |
|Z100 & EC4 |    5.0|   56.0| 0.089| 5.877| |Z100 & EC4 |   11.4|   49.6| 0.230| 5.877|
------------------------------------------ ------------------------------------------
```

The fit for this model is Chi-sq = 12.2 with 4 d.f., indicating the fit is not adequate. With the clogit model, PROC LOGISTIC also provides a score test for the proportional odds assumption:

```
Chi-Square       DF     Pr > ChiSq

   14.2431        2        <0.001
```

One way to observe why this test rejects the proportional odds assumption can be observed by comparing separate odds ratios for each 2x2 table are estimated with PROC LOGISTIC:

```
 2x2                    Odds
Table     Estimate     Ratio

1 vs 234     0.701      2.02
12 vs 34     1.548      4.70
123 vs 4     2.628     13.85
```

Adequacy of the proportional odds model would imply the three values under the Estimate column would be reasonably similar. Also, the estimates are nearly linear when plotted against the three comparisons (coded 1,2,3).

Partial Proportional Odds

The unequal values of the individual logits for aggregated 2x2 tables indicate an alternative to this proportional odds model could be the partial proportional odds model. The NLMIXED statements are the same as above except for revised linear predictors and to include three ESTIMATE statements for the odds ratios:

```
eta1 = Int_1 + _d0*(drug=1) + 0*_d1*(drug=1); * for a linear trend of the estimate,
eta2 = Int_2 + _d0*(drug=1) + 1*_d1*(drug=1); * add an extra amount to the logit;
eta3 = Int_3 + _d0*(drug=1) + 2*_d1*(drug=1); * when drug = 1 ;

ESTIMATE 'y: 1 v 234' _d0 + 0*_d1;
ESTIMATE 'y: 12 v 34' _d0 + 1*_d1;
ESTIMATE 'y: 123 v 4' _d0 + 2*_d1;
```

```
                        Standard
Parameter    Estimate    Error     Probt

 int_1       -1.6357     0.3440    0.001
 int_2       -1.2429     0.2959    0.003
 int_3       -0.2213     0.2573    0.415
 _d0          0.6899     0.4495    0.163
 _d1          0.9216     0.2661    0.008
```

```
--------------------------------------------------------------------------
|NLMIXED:     |            Ratings            |      |      |      |      |
|Predicted    |-------------------------------|      |      |      |      |
|values and   |   1   |   2   |   3   |   4   |  1   |  2   |  3   |  4   |
|row percents |-------+-------+-------+-------+------+------+------+------|
|             | Pred  | Pred  | Pred  | Pred  |Row % |Row % |Row % |Row % |
|-------------+-------+-------+-------+-------+------+------+------+------|
|Drug         |       |       |       |       |      |      |      |      |
|C15 & C60    |   16.8|   18.7|   19.1|    5.4|  28.0|  31.1|  31.9|   9.0|
|Z100 & EC4   |    9.9|    3.7|   13.5|   33.9|  16.3|   6.1|  22.1|  55.5|
--------------------------------------------------------------------------
```

The fit for this model is Chi-sq = 0.139 with 3 d.f., indicating an extremely good fit. For The modified logits (Estimates) and odds ratios can be compared to the proportional odds model above.

```
                                   odds_
      rsp         Estimate         ratio

  rsp 1 v 234       0.690          1.994
  rsp 12 v 34       1.612          5.010
  rsp 123 v 4       2.533         12.593
```

In Section 15.13 of Stokes, et. al. (2000) a method with PROC GENMOD is described that computes a Partial Proportional Odds Model by first restructuring the data set to include a dichotomous variable for the p-1 (=3) comparisons and then adding a REPEATED statement (to work with the multiple observations in the data file that result from each subject). PROC LOGISTIC (Version 12.1) provides the partial proportional odds logistic regression (unconstrained) with the UNEQUALSLOPES option in the MODEL statement. A detailed example of how LOGISTIC works with this model is described in Chapter 9.2.4 of Stokes, et. al. (2012). Details of these approaches will not be given here.

NLMIXED also allows the addition of individual effects to the betas (constrained) rather than a linear fit:

```
eta1 = _Int_1 + _d0*(drug=1) + d1*(drug=1); * add separate components, d1, d2,d3 ;
eta2 = _Int_2 + _d0*(drug=1) + d2*(drug=1); * to the linear predictor;
eta3 = _Int_3 + _d0*(drug=1) + d3*(drug=1);

ESTIMATE 'rsp 1 v 234' _d0 + d1;
ESTIMATE 'rsp 12 v 34' _d0 + d2;
ESTIMATE 'rsp 123 v 4' _d0 + d3;
```

Resulting in these Odds Ratios:

```
                                  odds_
     rsp         Estimate         ratio

 Rsp 1 v 234       0.701          2.016
 Rsp 12 v 34       1.548          4.700
 Rsp 123 v 4       2.628         13.852
```

Adjacent Logits

Adjacent logit models (ALM) provide comparisons of responses for specified pairs of adjacent categories, that is, it compares the ratio of two adjacent probabilities across the two levels of the explanatory variable (rather than the underlying continuum of all responses). For two adjacent columns the linear predictors constructed below compare the ratios of probabilities right-to-left for both rows 1 and 2.

```
        Log(p2/p1) = α1 + α2 + ß*x
        Log(p3/p2) = α1 + α3 + ß*x
        Log(p4/p3) = α1      + ß*x
```

The NLMIXED statements to estimate the adjacent logit model are the same as above with these modifications for the linear predictors and probability statements:

```
Linear predictors

eta1 = alpha1 + alpha2 + drg*(drug=1); * drug = 2 is reference category;
eta2 = alpha1 + alpha3 + drg*(drug=1);
eta3 = alpha1          + drg*(drug=1);
```

To compute probabilities, with these equations work backwards from three values of eta to find p4/p3, p3/p2, and p2/p1. The necessary total for the probabilities to sum to 1 is:

```
Total = (1 + exp(eta1) + exp(eta1+eta2) + exp(eta1+eta2+eta3));
```

The probability calculations are then:

```
p1 = 1 / total;
p2 = exp(eta1)*p1;
p3 = exp(eta2)*p2;
p4 = exp(eta3)*p3;
```

The parameter estimates from this model are

```
                       Standard
Parameter   Estimate      Error    EXP(Estimate)

alpha1         0.445      0.242
alpha2       -0.0663      0.395
alpha3         0.396      0.444
drg           -0.809      0.188     0.445
```

To compare this result with the proportional odds model, which assumes an increasing order of the response categories, the sign of the estimate is reversed or take ratio of the exponentiated coefficient, since ALM compares probs right to left:

```
Exp(Estimate) = 1/0.445 = 2.25
```

The parameter estimates produce the following table of predicted values:

```
      ----------------------------------------------------------------------
      |Predicted   |           Counts           |         Proportions       |
      |Values      |---------------------------|---------------------------|
      |            |      |      |      |VGd/  |      |      |      | VGd/ |
      |            | Poor | Fair | Good | Excl | Poor | Fair | Good | Excl |
      |------------+------+------+------+------|------+------+------+------|
      |Drug        |      |      |      |      |      |      |      |      |
      |C15 & C60   |  21.5|  14.0|  14.4|  10.0| 0.359| 0.233| 0.241| 0.167|
      |Z100 & EC4  |   5.5|   8.0|  18.6|  29.0| 0.090| 0.131| 0.304| 0.475|
      ----------------------------------------------------------------------
```

The fit for this model is Chi-sq = 15.0 with 4 d.f., indicating the fit is not adequate. To see how these predicted proportions are computed, the ratios of adjacent values all equal the exponentiated value of the coefficient.

```
        Fair vs Poor:      [0.233 / 0.359] / [0.131 / 0.090] = 0.445
        Good vs Fair:      [0.241 / 0.233] / [0.304 / 0.131] = 0.445
        VGd/Excl vs Good: [0.167 / 0.241] / [0.475 / 0.304] = 0.445
```

If these ratios are allowed to vary freely (consider how the linear predictors would be formed), the predicted values would produce a perfect fit, the same end result as the generalized logit.

The analogous statements in CATMOD would be:

```
PROC CATMOD DATA= indat ;
WEIGHT count;
MODEL rsp = _response_ drug / param=ref ;
RESPONSE alogit out=acr;
RUN; QUIT;
```

Drug is treated as a class effect with the request for reference category coding (param=ref) rather than the default effect coding. The addition of _response_ to the model statement estimates a single set of coefficients for each pair of adjacent probabilities. One difference between the CATMOD and the NLMIXED solutions is that for the 4 levels of the ordinal response, CATMOD estimates a drug coefficient = -0.681, so the ratios of adjacent probabilities are still held fixed, but when examining the contents of the output data set acr the adjacent cell probabilities are conditional and do not sum to 1.

Continuation Ratio

The continuation ratio (CR) model estimates the probability of a response at a specific level of the ordinal value, given the number of responses at that level or higher (one can also evaluate from higher to lower). An important assumption is that the subject proceeds gradually through the levels of the response and will not reverse the trend. As the restructured table of the example data shown below, assuming a four level ordinal response variable, the continuation ratio model examines the binary nature of the four responses divided into three stages. The mathematical derivation of the CR model is beyond the scope of this paper, but is outlined in Chapter 6, Problem 3 in Agresti (1984). It computes the probability of the number of ratings at stage i (i = 1,2,3) given the number of counts that exist in the levels greater than i. The process essentially groups the table into multiple sets of 2x2 tables where y2 models the binary nature of the responses at each stage.

```
-----------------------------------------------
|Rating         |     y2     |     |     |     |
|Comparisons    |------------|     |     |     |
|               |  0  |  1  | All |  0  |  1  |
|               |-----+-----+-----+-----+-----|
|i vs y > i     |  N  |  N  |  N  |Row %|Row %|
|---------------+-----+-----+-----+-----+-----|
|1 vs 2,3,4     |     |     |     |     |     |
|  C15 & C60    |   17|   43|   60| 28.3| 71.7|
|  Z100 & EC4   |   10|   51|   61| 16.4| 83.6|
|2 vs 3,4       |     |     |     |     |     |
|  C15 & C60    |   18|   25|   43| 41.9| 58.1|
|  Z100 & EC4   |    4|   47|   51|  7.8| 92.2|
|3 vs 4         |     |     |     |     |     |
|  C15 & C60    |   20|    5|   25| 80.0| 20.0|
|  100 & EC4    |   13|   34|   47| 27.7| 72.3|
-----------------------------------------------
```

The mathematical derivation of the CR model and equivalence to the stage-wise process described above indicates how NLMIXED code works directly with the original data set (no restructuring necessary), with these specific commands:

```
PARMS int_1 .1 int_2 .1 int_3 .1 drg .7  ;
eta1 = int_1 + drg*(drug=1);
eta2 = int_2 + drg*(drug=1);
eta3 = int_3 + drg*(drug=1);
p1 =  1 / (1 + exp(-eta1));
p2 = (1 / (1 + exp(-eta2)))*(1 - p1);
p3 = (1 / (1 + exp(-eta3)))*(1 - p1 - p2);
p4 =  1 - (p1 + p2 + p3);
```

```
Parameter    Estimate    Odds Ratio

Int_1        -2.216
int_2        -2.084
int_3        -0.723
drg           1.597       4.94
```

With the restructured dataset PROC LOGISTIC produces the same results (notice the param=GLM coding in the CLASS statement and NOint option on the MODEL statement):

```
ODS SELECT responseprofile parameterestimates oddsratiospl;

PROC LOGISTIC DATA=cltrCR order=internal;
FREQ count;
CLASS stg drug / param=GLM;
MODEL rsp2 = stg drug / NOINT aggregate scale=none expb;
ODDSRATIO drug / cl=pl;
RUN;
```

```
Parameter               DF    Estimate   Odds Ratio

stg       1 vs 2,3,4     1     -2.216
stg       2 vs 3,4       1     -2.084
stg       3 vs 4         1     -0.723
drug      1              1      1.597        4.94
drug      2              0          0
```

The parameter estimates for the intercepts due to the stages are the same as the CR and the odds ratio is also found to be 4.94 though the increasing wider differences in proportions between drugs observed at each stage in the table above may make it suspect.

If the comparisons of the response levels were made in the decreasing order [i.e., 4 vs 1,2,3 / 3 vs 12 / 2 vs 1], the resulting odds ratio of 0.341 does not have the equivalent inverse relation found with the proportional odds model (i.e., 0.341 is not equal to 1/4.94=0.202). The methodology for implementation (with a restructured data set) and how to interpret and diagnose the adequacy of the Continuation Ratio model with PROC LOGISTIC is described in Chapter 6, Section 7 of Paul Allison's book on Logistic Regression (2012).

Another interesting relationship with the CR model occurs when the cloglog link is chosen. The observed coefficient for drug is the same when running PROC LOGISTIC with either the original data or the restructured data sets.

```
PROC LOGISTIC DATA=cltr ; * original data set;
CLASS drug / param=glm;
FREQ count;
MODEL y = drug / link = cloglog; * ordinal responses;
RUN;

PROC LOGISTIC DATA=cltrCR ; * restructured data set;
CLASS stg drug / param=GLM;
FREQ count;
MODEL y2 = stg drug / NOINT link = cloglog; * binary data;
RUN;
```

Both approaches compute the coefficient for drug as 1.345 (the intercepts are necessarily different). The cloglog link has been recommended as a way to avoid data restructuring for the CR model, though as shown in this example, the logit link can also be applied with NLMIXED to the original data set.

Stereotype Model

First, consider the generalized logit model with four response levels (J=4) and three explanatory variables (p=3). The linear predictors are coded as:

$$\mathrm{Log}(\pi_1(x) / \pi_4(x)) = \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3$$
$$\mathrm{Log}(\pi_2(x) / \pi_4(x)) = \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2 + \beta_{23}x_3$$
$$\mathrm{Log}(\pi_3(x) / \pi_4(x)) = \alpha_3 + \beta_{31}x_1 + \beta_{32}x_2 + \beta_{33}x_3$$

Estimation of all 12 parameters produces a perfect fit. Consider how computation of the beta coefficients can be modified:

$$\mathrm{Log}(\pi_1(x) / \pi_4(x)) = \alpha_1 + \varphi_1{*}(\beta_1x_1 + \beta_2x_2 + \beta_3x_3)$$
$$\mathrm{Log}(\pi_2(x) / \pi_4(x)) = \alpha_2 + \varphi_2{*}(\beta_1x_1 + \beta_2x_2 + \beta_3x_3)$$
$$\mathrm{Log}(\pi_3(x) / \pi_4(x)) = \alpha_3 + \varphi_3{*}(\beta_1x_1 + \beta_2x_2 + \beta_3x_3)$$

$\varphi_1 {*}\beta_1$ vs β_{11} $\varphi_1 {*}\beta_2$ vs β_{12} $\varphi_1 {*}\beta_3$ vs β_{13}
$\varphi_2 {*}\beta_1$ vs β_{21} $\varphi_2 {*}\beta_2$ vs β_{22} $\varphi_2 {*}\beta_3$ vs β_{23}
$\varphi_3 {*}\beta_1$ vs β_{31} $\varphi_3 {*}\beta_2$ vs β_{32} $\varphi_3 {*}\beta_3$ vs β_{33}

For identifiability set $\varphi_1 = 1$ ($\alpha_4 = 0$ and $\varphi_4 = 0$ are assumed from the fourth unwritten equation) resulting in the estimation of 8 parameters, 4 fewer than the generalized logit. For a stereotype model, the number of parameters to be estimated depends on the number of response categories and the number of explanatory variables:

```
J = number of response levels
p = number of explanatory variables

Generalized Logit: (J-1) + (J-1) * p
       Stereotype:  2*J - 3 + p
```

This table compares the efficiency possible for 2 or 3 explanatory variables with a stereotype model (ST) versus a generalized baseline (GL):

```
--------------------------------
|      |   Explanatory Vars    |
|      |-----------------------|
|      |   1   |   2   |   3   |
|      |-------+-------+-------|
|      |GL |ST |GL |ST |GL |ST |
|------+---+---+---+---+---+---|
|Resp  |   |   |   |   |   |   |
|3     |  4|  4|  6|  5|  8|  6|
|4     |  6|  6|  9|  7| 12|  8|
|5     |  8|  8| 12|  9| 16| 10|
|6     | 10| 10| 15| 11| 20| 12|
--------------------------------
```

The example data presented above with other types of models does not work for the stereotype model, namely because as the table indicates, with one categorical explanatory variable, the stereotype model requires the same number of parameters, thus making it equivalent to the generalized logit.

As an alternative, consider this example data with J = 4 responses, p = 2 explanatory variables (coded 1/2) for a total 400 observations.

```
----------------------------------------
| Counts |           y           |     |
|        |-----------------------|     |
|        |  1  |  2  |  3  |  4  | All |
|--------+-----+-----+-----+-----+-----|
|x1  x2  |     |     |     |     |     |
|1   1   |   21|   12|   18|   34|   85|
|    2   |   44|   19|   35|   40|  138|
|2   1   |    9|    6|   17|   41|   73|
|    2   |   20|   11|   29|   44|  104|
----------------------------------------
```

The following NLMIXED commands estimate the stereotype model:

```
PARMS theta1 .1 theta2 .1 theta3 .1  phi2 .3 phi3 .7  b_x1 .1 b_x2 .1 ;
phi1=1;  * fix these parameters as specified values ;
phi4=0; theta4=0;
eta1 = theta1 + phi1*( b_x1*(x1=1) + b_x2*(x2=1) );
eta2 = theta2 + phi2*( b_x1*(x1=1) + b_x2*(x2=1) );
eta3 = theta3 + phi3*( b_x1*(x1=1) + b_x2*(x2=1) );
eta4 = theta4 + phi4*( b_x1*(x1=1) + b_x2*(x2=1) );
tot = exp(eta1) + exp(eta2) + exp(eta3) + exp(eta4);
p1 = exp(eta1) / tot;
p2 = exp(eta2) / tot;
p3 = exp(eta3) / tot;
p4 = exp(eta4) / tot;
```

Coefficients are estimated for the stereotype model and compared with the proportional odds and generalized logit models:

Parameter		POM	Stereo Type	Generalized Logit
Intercept	1	-1.419	-0.828	-0.822
Intercept	2	-0.816	-1.389	-1.44
Intercept	3	0.237	-0.567	-0.411
x1	1	0.674	0.946	0.939
	2		0.691	0.735
	3		0.412	0.275
x2	1	-0.456	-0.627	-0.634
	2		-0.458	-0.389
	3		-0.273	-0.487
			Φ_2 = 0.731	
			Φ_3 = 0.436	

For ordinal response data with four levels, the two estimated Φ coefficients need to maintain the specific ordering of values bounded between 0 and 1 and also that:

$$\Phi_1 = 1 > \Phi_2 > \Phi_3 > \Phi_4 = 0$$

That is, estimation of the values of Φ should not determine order of response levels, that is, how to interpret the model if Φ_3 is greater than Φ_2. The equality of these coefficients can be tested with ESTIMATE statements in NLMIXED (not shown) which may lead to a decision to collapse two or more response categories into one.

With one categorical explanatory variable (drug), the stereotype coding of the linear predictors has an interesting relationship to other logistic regression models. The phi coefficients of the linear predictors set as 1/0:

```
eta1 = int_1 + 1 * (d0*(drug=1));
eta2 = int_1 + 1 * (d0*(drug=1));
eta3 =     0 + 0 * (d0*(drug=1));
eta4 =     0 + 0 * (d0*(drug=1));
```

produces the same results as recoding the ordinal responses 1,2 as 0 and 3,4 as 1 and running binary logistic regression with PROC LOGISTIC (d0=1.548, odds ratio=4.70).

Entering descending integer values for the phi coefficients:

```
eta1 = theta1 + 3*d0*(drug=1);
eta2 = theta2 + 2*d0*(drug=1);
eta3 = theta3 + 1*d0*(drug=1);
eta4 =      0 + 0*d0*(drug=1);
```

produces the results found when comparing worst to better conditions as computed above with adjacent logits (d0=.809, exp(d0)=2.247).

References

Ananth, Cande V and David G Kleinbaum, (1997) *GRegression Models for Ordinal Responses: A Review of Methods and Applications,* International Journal of Epidemiology, Volume 26, No. 6, pp. 1323-1333.

Agresti, Alan. (1984) Analysis of Ordinal Categorical Data, John Wiley & Sons, New York, NY.

Agresti, Alan. (2010) Analysis of Ordinal Categorical Data, Second Ed., John Wiley & Sons, New York, NY.

Allison, Paul D. (2012) *Logistic Regression Using SAS@: Theory and Application*, Second Edition. Cary, NC: SAS Institute Inc.

Kuss, Oliver, *On the estimation of the stereotype regression model*, Computational Statistics & Data Analysis 50 (2006) 1877 – 1890

Stokes, Maura E., Charles S. Davis, and Gary G. Koch. (2000) *Categorical Data Analysis Using the SAS@ System, Second Edition*, Cary, NC: SAS Institute Inc.

Stokes, Maura E., Charles S. Davis, and Gary G. Koch. (2012) *Categorical Data Analysis Using the SAS@ System, Third Edition*, Cary, NC: SAS Institute Inc.

Your comments and questions are valued and encouraged. Contact the author at:

Robin High
Statistical Coordinator
College of Public Health
Department of Biostatistics
University of Nebraska Medical Center
984375 Nebraska Medical Center
Omaha, NE 68198-4375
Phone: (402) 559-2929
email: rhigh@unmc.edu