Ordered/Ordinal Logistic Regression with SAS and Stata[1]

This document will describe the use of Ordered Logistic Regression (OLR), a statistical technique that can sometimes be used with an ordered (from low to high) dependent variable. The dependent variable used in this document will be the fear of crime, with values of:

1 = not at all fearful
2 = not very fearful
3 = somewhat fearful
4 = very fearful

Ordered logit model has the form:

$$\text{logit}(p_1) \equiv \log \frac{p_1}{1 - p_1} = \alpha_1 + \beta' x$$

$$\text{logit}(p_1 + p_2) \equiv \log \frac{p_1 + p_2}{1 - p_1 - p_2} = \alpha_2 + \beta' x$$

.

.

$$\text{logit}(p_1 + p_2 + \cdots + p_k) \equiv \log \frac{p_1 + p_2 + \cdots + p_k}{1 - p_1 - p_2 - \cdots - p_k} = \alpha_k + \beta' x$$

$$\text{and } p_1 + p_2 + \cdots + p_{k+1} = 1$$

This model is known as the proportional-odds model because the odds ratio of the event is independent of the category *j*. The odds ratio is assumed to be constant for all categories.

Source:
http://www.indiana.edu/~statmath/stat/all/cat/2b1.html

Syntax and results using both SAS and Stata will be discussed.

OLR models cumulative probability. It simultaneously estimates multiple equations. The number of equations it estimates will the number of categories in the dependent variable minus one. So, for our example, three equations will be estimated. The equations are:

	Pooled Categories	compared to	Pooled Categories
Equation 1:	1		2 3 4
Equation 2:	1 2		3 4
Equation 3:	1 2 3		4

Each equation models the odds of being in the set of categories on the left versus the set of categories on the right.

OLR provides only one set of coefficients for each independent variable. Therefore, there is an assumption of parallel regression. That is, the coefficients for the variables in the equations would not vary significantly if they were estimated separately. The intercepts would be different, but the slopes would be essentially the same. (In Stata there is a way to test whether this assumption is being met. See “Testing the assumption of Parallel Regression” later in this document.)

The following syntax in Stata can be used to estimate an OLR model.

```
. ologit  nfear_in female educ
```

And this is the output for that equation.

```
Iteration 0:   log likelihood = -15065.131
Iteration 1:   log likelihood = -14925.462
Iteration 2:   log likelihood = -14925.243
```

[1] Prepared by Karen Snedker, Patty Glynn, Chiachi Wang, University of Washington, 10/25/02

```
Ordered logit estimates                           Number of obs   =      12261
                                                  LR chi2(2)      =     279.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -14925.243                       Pseudo R2       =     0.0093

------------------------------------------------------------------------------
    nfear_in |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .5454657   .0336724    16.20   0.000     .4794691    .6114624
        educ |  -.0180993   .0062164    -2.91   0.004    -.0302832   -.0059154
-------------+----------------------------------------------------------------
       _cut1 |   -1.11355   .0924472          (Ancillary parameters)
       _cut2 |   .6022201    .092258
       _cut3 |   2.977691   .0984957
------------------------------------------------------------------------------
```

The syntax used to estimate the same OLR equation in SAS follows.

```
proc logistic descending ;
model nfear_in = female educ ; run ;
```

And the results follow.

```
                              The LOGISTIC Procedure
                                Model Information
             Data Set                      SNEDKER
             Response Variable             nfear_in             fear
             Number of Response Levels     4
             Number of Observations        12261
             Model                         cumulative logit
             Optimization Technique        Fisher's scoring

                                 Response Profile
                          Ordered                     Total
                            Value     nfear_in    Frequency

                                1            4          641
                                2            3         3858
                                3            2         4793
                                4            1         2969

              Probabilities modeled are cumulated over the lower Ordered Values.

                              Model Convergence Status
                       Convergence criterion (GCONV=1E-8) satisfied.
                       Score Test for the Proportional Odds Assumption

                          Chi-Square       DF     Pr > ChiSq
                            252.5987        4         <.0001

                                Model Fit Statistics
                                                      Intercept
                                       Intercept         and
                       Criterion          Only        Covariates

                       AIC             30136.263       29860.486
                       SC              30158.505       29897.557
                       -2 Log L        30130.263       29850.486
```

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	279.7770	2	<.0001
Score	277.1848	2	<.0001
Wald	277.6892	2	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 4	1	-2.9777	0.0970	942.0679	<.0001
Intercept 3	1	-0.6023	0.0900	44.7414	<.0001
Intercept 2	1	1.1135	0.0905	151.4820	<.0001
female	1	0.5455	0.0337	262.6788	<.0001
educ	1	-0.0181	0.00612	8.7416	0.0031

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
female	1.725	1.615	1.843
educ	0.982	0.970	0.994

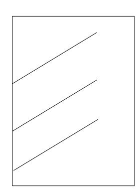
Association of Predicted Probabilities and Observed Responses

Percent Concordant	50.3	Somers' D	0.131
Percent Discordant	37.3	Gamma	0.149
Percent Tied	12.4	Tau-a	0.090
Pairs	51624633	c	0.565

You may interpret the coefficients as you would interpret logistic regression coefficients - except in this case, there are three transitions estimated instead of one transition - as there would be with a dichotomous dependent variable. Being female increases the likelihood of being in a higher fear category, while being more highly educated reduces the likelihood of being in a higher fear category. A positive coefficient indicates an increased chance that a subject with a higher score on the independent variable will be observed in a higher category. A negative coefficient indicates that the chances that a subject with a higher score on the independent variable will be observed in a lower category. SAS reports the odds ratio estimates, but Stata does not. Odds ratios can easily be derived from the coefficients by taking the exponent of the coefficient. (For example, In Excel, **=exp(coef)**)

Note that Stata reports "Ancillary parameters", and SAS reports Intercepts. The numbers are the same, but the signs are reversed. Consider that OLR restrains estimation of the coefficients so that they cannot vary between transitions. That is, the slope for education for Equation 1, must be the same as the slopes for Equations 2 and 3 (as described above under "OLR models cumulative probability"). Only the Intercepts are allowed to vary.

	Pooled Categories	compared to	Pooled Categories
Equation 1:	1		2 3 4
Equation 2:	1 2		3 4
Equation 3:	1 2 3		4



The Intercepts and Cut Points can be used to calculate predicted probabilities for a person with a given set of characteristics of being in a particular category. The formula used with SAS Intercepts and Stata cut points will be slightly different. Information about calculating the probabilities for the output Stata provides can be found at the following URL. http://www.stata.com/support/faqs/stat/ologit_con.html "Example 20: Predicted Probability Computation" in the following

URL provides information about calculating predicted probabilities with SAS.
http://www.indiana.edu/~statmath/stat/all/cat/2b1.html

For information on testing the model for explanatory power of a model, please refer to:
http://www.ats.ucla.edu/stat/stata/library/logit_wgould.htm

Testing the assumption of Parallel Regression (Drawn from Regression Models for Categorical Dependent Variables Using Stata, Long and Freese, 2001)

J. Scott Long and Jeremy Freese have created an add-in file for Stata which allows the easy testing of the assumption of Parallel Regression. For complete information on how to install this ado file, see:
http://www.indiana.edu/~jsl650/spostinstall.htm#Heading03

Once you have installed these useful additions, estimate your Ordinal Logistic Regression model (for example):

```
. ologit  nfear_in female educ
```

And then issue the command:

```
. brant, detail
```

You will get results like:

```
Estimated coefficients from j-1 binary regressions

               y>1         y>2         y>3
female   .62821863   .48924177   .55179906
  educ   .04872347  -.04745798  -.15864904
 _cons   .14818344  -.16496036  -1.1202262

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |    256.77    0.000     4
-------------+--------------------------
      female |     10.44    0.005     2
        educ |    250.30    0.000     2
----------------------------------------

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```

First you will see the results of each binary regression that was estimated when the OLR coefficients were calculated. These represent the equations represented above under the heading **"OLR models cumulative probability".** The model " y>1" represents Equation 1, "y>2" is Equation 2, and "y>3" is Equation 4. (We proved this to ourselves by estimating logistic regression models for each of these.) For the Assumption of Parallel Regression to be true, the coefficients across these equations would not vary very much. But, in this example, they do vary. In fact, for education, the slope even changes directions. You are also provided with the results of a Chi-square test, which, in this case, shows that the parallel regression assumption has been violated. Note: This test is sensitive to the number of cases. Samples with larger numbers of cases are more likely to show a statistically significant test, and evidence that the parallel regression assumption has been violated.