# Gologit2: A Program for Generalized Logistic Regression/ Partial Proportional Odds Models for Ordinal Dependent Variables

Richard Williams, Richard.A.Williams.5@ND.Edu

Last revised May 12, 2005

[This document is a work in progress. Comments are welcome. Parts of this paper are adapted from the documentation for Vincent Fu's original `gologit` command and are used with permission. Those who learn best by examples may wish to skim over the early sections.]

**Overview.** `gologit2` is a user-written program that estimates generalized logistic regression models for ordinal dependent variables. The actual values taken on by the dependent variable are irrelevant except that larger values are assumed to correspond to "higher" outcomes.

A major strength of `gologit2` is that it can also estimate two special cases of the generalized model: the *proportional odds model* and the *partial proportional odds model*. Hence, `gologit2` can estimate models that are less restrictive than the proportional odds /parallel lines models estimated by `ologit` (whose assumptions are often violated) but more parsimonious and interpretable than those estimated by a non-ordinal method, such as multinomial logistic regression (i.e. `mlogit`). The `autofit` option greatly simplifies the process of identifying partial proportional odds models that fit the data, while the `pl` (parallel lines) and `npl` (non-parallel lines) options can be used when users wish to specify the model themselves.

An alternative but equivalent parameterization of the model that has appeared in the literature is reported when the `gamma` option is selected. Other key advantages of `gologit2` include support for linear constraints, Stata 8.2 survey data (`svy`) estimation, and the computation of estimated probabilities via the `predict` command.

`gologit2` is inspired by Vincent Fu's `gologit` program and is backward compatible with it but offers several additional powerful options. `gologit2` was written for Stata 8.2 and many of the references in the help file are for Stata 8 manuals.

**Description.** The `ologit` command included with Stata imposes what is called the *proportional odds assumption* on the data. This is also known as the *parallel lines/ parallel regressions assumption*. The proportional odds/parallel lines model is a special case of the generalized model estimated by `gologit2`. By default, `gologit2` relaxes the proportional odds assumption and allows the effects of the explanatory variables to vary with the point at which the categories of the dependent variable are dichotomized. However, if the `pl` option is specified without parameters, `gologit2` estimates the proportional odds model, e.g. the commands

```
ologit y x1 x2 x3
```

and

```
gologit2 y x1 x2 x3, pl lrforce
```

will produce equivalent results.

In practice, the proportional odds assumption is often violated by the data. Standard advice in such situations is to go to a non-ordinal model, such as `mlogit`. Unfortunately, such models can be far less parsimonious and more difficult to interpret than the proportional odds model. `gologit2` provides an alternative by estimating partial proportional odds models. With such models, the parallel lines/ proportional odds assumption can be relaxed for some explanatory variables while being maintained for others. For example, the command

```
gologit2 y x1 x2 x3, npl(x1)
```

would relax the proportional odds/parallel lines assumption for x1 while maintaining it for x2 and x3. An equivalent command is

```
gologit2 y x1 x2 x3, pl(x2 x3)
```

which forces x2 and x3 to meet the proportional odds/ parallel lines assumption while not imposing the assumption on x1.

More formally, suppose we have an ordinal dependent variable Y which takes on the values 1, 2, ..., m. The generalized ordered logit model estimates a set of coefficients (including one for the constant) for each of the m - 1 points at which the dependent variable can be dichotomized. The probabilities that Y will take on each of the values 1, ..., m is equal to

$$P(Y = 1) = F(-XB_1)$$
$$P(Y = j) = F(-XB_j) - F(-XB_{j-1}) \qquad j = 2, ..., m - 1$$
$$P(Y = m) = 1 - F(-XB_{m-1})$$

The generalized ordered logit model uses the logistic distribution as the cumulative distribution, although other distributions may also be used. The logistic distribution allows researchers to interpret this model in terms of logits:

$$\log[P(Y > k) / P(Y <= k)] = XB_k \qquad k = 1, ..., m-1$$

The proportional odds model (estimated by Stata's `ologit` command and by `gologit2` with the `pl` option) restricts the $B_k$ coefficients to be the same for every dividing point k = 1, ..., m-1. The partial proportional odds model (estimated in `gologit2` via the `npl()` and `pl()` options) restricts some $B_k$ coefficients to be the same for every dividing point while others are free to vary.

Note that unlike models such as OLS regression and binary logit, the generalized ordered logit model imposes explicit restrictions on the range of the X variables. Since probabilities are by definition constrained to be in the range [0,1], valid combinations of the X variables must satisfy the following inequalities:

$$XB_1 >= XB_2 >= XB_3 ... >= XB_{m-1}$$

Other scholars (e.g. Peterson & Harrell, 1990) have proposed an alternative but equivalent parameterization of the partial proportional odds model in which there is only one set of Betas but a second set of coefficients, called Gammas, can vary across the dividing points. The gammas indicate the extent to which the proportional odds assumption does not hold for a variable; if the gammas for a variable equal 0, then the parallel lines assumption holds for that variable. This parameterization can be displayed by using the `gamma` option.

Key Options. `gologit2` supports many standard Stata options, which work the same way as they do with other Stata commands. Options which are unique to `gologit2` are described below. See the help file for other options. The complete syntax is

```
gologit2 depvar [indepvars] [weight] [if exp] [in range] [, lrforce pl pl(varlist) npl
       npl(varlist) autofit autofit(alpha) gamma nolabel store(name)
       constraints(clist) robust cluster(varname) level(#) score(newvarlist|stub*) or
       log v1 svy svy_options maximize_options ]
```

`pl, npl, npl(), pl(), autofit` and `autofit()` provide alternative means for imposing or relaxing the proportional odds/ parallel lines assumption. Only one may be specified at a time.

- `autofit(`*alpha*`)` uses an iterative process to identify the partial proportional odds model that best fits the data. *alpha* is the desired significance level for the tests; alpha must be greater than 0 and less than 1. If `autofit` is specified without parameters, the default alpha-value is .05. Note that, the higher alpha is, the easier it is to reject the parallel lines assumption, and the less parsimonious the model will tend to be. This option can take a little while because several models may need to be estimated. The use of `autofit` is highly recommended but other options provide more control over the final model if the user wants it.

- `pl` specified without parameters constrains all independent variables to meet the proportional odds assumption. It will produce results that are equivalent to `ologit`.

- `npl` specified without parameters relaxes the proportional odds/ parallel lines assumption for all explanatory variables. This is the default option and presents results equivalent to the original `gologit`.

- `pl(varlist)` constrains the specified explanatory variables to meet the proportional odds/ parallel lines assumption. All other variable effects do not need to meet the assumption. The variables specified must be a subset of the explanatory variables.

- `npl(varlist)` frees the specified explanatory variables from meeting the proportional odds/ parallel lines assumption. All other explanatory variables are constrained to meet the assumption. The variables specified must be a subset of the explanatory variables.

`lrforce` forces Stata to report a Likelihood Ratio Statistic under certain conditions when it ordinarily would not. Some types of constraints can make a Likelihood Ratio chi-square test invalid. Hence, to be safe, Stata reports a Wald statistic whenever constraints are used. But, Likelihood Ratio statistics should be correct for the types of constraints imposed by the `pl` and `npl` commands. Note that the `lrforce` option will be ignored when robust standard errors are specified either directly or indirectly, e.g. via use of the `robust` or `svy` options. Use this option with caution if you specify other constraints since these may make a LR chi- square statistic inappropriate.

`gamma` displays an alternative but equivalent parameterization of the partial proportional odds model used by Peterson and Harrell (1990) and Lall et al (2002). Under this parameterization, there is one Beta coefficient and M-2 Gamma coefficients for each explanatory variable, where M = the number of categories for Y. The gammas indicate the extent to which the proportional odds assumption is violated by the variable, i.e. when the gammas do not significantly differ from 0 the proportional odds assumption is met. Advantages of this parameterization include the fact that it is more parsimonious than the default layout. In addition, by examining the test statistics for the Gammas, you can get a feel for which variables meet the proportionality assumption and which do not.

`store(`*`name`*`)` causes the command `estimates store` *`name`* to be executed when `gologit2` finishes. This is useful for when you wish to estimate a series of models and want to save the results.

`nolabel` causes the equations to be named eq1, eq2, etc. The default is to use the first 32 characters of the value labels and/or the values of Y as the equation labels. Note that some characters cannot be used in equation names, e.g. the period (.), the dollar sign ($), and the colon(:), and will be replaced with the underscore (_) character. The default behavior works well when the value labels are short and descriptive. It may not work well when value labels are very long and/or include characters that have to be changed to underscores. If the printout looks unattractive and/or you are getting strange errors, try changing the value labels of Y or else use the `nolabel` option.

`v1` causes `gologit2` to return results in a format that is consistent with `gologit` 1.0. This may be useful/necessary for post-estimation commands that were written specifically for `gologit` (in particular, some versions of the Long and Freese `spost` commands support `gologit` but not `gologit2`). However, post-estimation commands written for `gologit2` (including `predict`) may not work correctly if `v1` is specified.

`log` displays the iteration log. By default it is suppressed.

`or` reports the estimated coefficients transformed to relative odds ratios, i.e., exp(b) rather than b; see [R] `ologit` for a description of this concept. Options `rrr`, `eform`, and `irr` produce identical results (labeled differently) and can also be used.

`constraints(`*`clist`*`)` specifies linear constraints to be applied during estimation. Constraints are defined with the `constraint` command. `constraints(1)` specifies that the model is to be constrained according to constraint 1; `constraints(1-4)` specifies constraints 1 through 4; `constraints(1-4,8)` specifies 1 through 4 and 8. Keep in mind that the `pl`, `npl` and `autofit` options work by generating across-equation constraints, which may affect how any additional constraints should be specified. When using the `constraint` command, refer to equations by their equation #, e.g. #1, #2, etc.

`svy` indicates that `gologit2` is to pick up the svy settings set by `svyset` and use the robust variance estimator. Thus, this option requires the data to be `svyset`; see help `svyset`. When using `svy` estimation, use of `if` or `in` restrictions will not produce correct variance estimates for subpopulations in many cases. To compute estimates for subpopulations, use the `subpop()` option. If `svy` has not been specified, use of other Stata 8.2 svy-related options (e.g. `subpop`, `deff`, `meff`) will produce an error.

*Other standard Stata options supported by gologit2:* `robust cluster level score`

*Other standard svy-related options supported by gologit2:* `subpop nosvyadjust prob ci deff deft meff meft`

*Options available when replaying results:* `gamma store or level prob ci deff deft`

`prob`, `ci`, `deff` and `deft` are only available when `svy` estimation has been used.

*Options available for the predict command:* `xb stdp stddp p`

p gives the predicted probability. Note that you specify one new variable with `xb`, `stdp`, and `stddp` and specify either one or k new variables with `p`. These statistics are available both in and out of sample; type "`predict ... if e(sample) ...`" if wanted only for the estimation sample.

## Examples.

*Example 1: Attitudes Toward Working Mothers*. Long and Freese (2003) present data from the 1977/1989 General Social Survey. Respondents are asked to evaluate the following statement: “A working mother can establish just as warm and secure a relationship with her child as a mother who does not work.” Responses were coded as 1 = Strongly Disagree (SD), 2 = Disagree (D), 3 = Agree (A), and 4 = Strongly Agree (SA). Explanatory variables are yr89 (survey year; 0 = 1977, 1 = 1989), male (0 = female, 1 = male), white (0 = nonwhite, 1 = white), age (measured in years) ed (years of education) and prst (occupational prestige scale). Based on their analysis (reproduced below), Long and Freese conclude that the parallel lines assumption is violated with these data and suggest that an alternative ordinal regression model or a multinomial logit model may be called for (the `brant` command requires that the `spost` routines be installed):

```
. use http://www.nd.edu/~rwilliam/stata/ordwarm2, clear
(77 & 89 General Social Survey)

. ologit  warm yr89 male white age ed prst

Iteration 0:   log likelihood = -2995.7704
Iteration 1:   log likelihood = -2846.4532
Iteration 2:   log likelihood = -2844.9142
Iteration 3:   log likelihood = -2844.9123

Ordered logit estimates                           Number of obs   =       2293
                                                  LR chi2(6)      =     301.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -2844.9123                       Pseudo R2       =     0.0504

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yr89 |   .5239025   .0798988     6.56   0.000     .3673037    .6805013
        male |  -.7332997   .0784827    -9.34   0.000    -.8871229   -.5794766
       white |  -.3911595   .1183808    -3.30   0.001    -.6231815   -.1591374
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
-------------+----------------------------------------------------------------
       _cut1 |  -2.465362   .2389126          (Ancillary parameters)
       _cut2 |   -.630904   .2333155
       _cut3 |   1.261854   .2340179
------------------------------------------------------------------------------

. brant

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     49.18    0.000    12
-------------+--------------------------
        yr89 |     13.01    0.001     2
        male |     22.24    0.000     2
       white |      1.27    0.531     2
         age |      7.38    0.025     2
          ed |      4.31    0.116     2
        prst |      4.33    0.115     2
----------------------------------------

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

The Brant test suggests that yr89 and male are especially problematic with regards to the parallel lines assumption, age is borderline, but the other variables do not appear to violate the assumption.

Using `gologit2`, we can (a) reproduce `ologit’s` estimates by using the `pl` parameter, i.e. estimate a model in which all variables are constrained to meet the proportional odds/ parallel regressions/ parallel lines assumption, (b) estimate a model (`gologit2`’s default) in which no variables have to meet the parallel lines assumption (c) do a global likelihood ratio chi-square test of the parallel lines assumption, and (d) use `autofit` to estimate a model in which some variables are constrained to meet the parallel lines assumption while others are not.

```
. * Part a.  Replicate ologit's results by using the pl and lrforce parameters.
. gologit2  warm yr89 male white age ed prst, pl lrforce store(constrained)

Generalized Ordered Logit Estimates               Number of obs   =       2293
                                                  LR chi2(6)      =     301.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -2844.9123                       Pseudo R2       =     0.0504

 ( 1)  [SD]yr89 - [D]yr89 = 0
 ( 2)  [SD]male - [D]male = 0
 ( 3)  [SD]white - [D]white = 0
 ( 4)  [SD]age - [D]age = 0
 ( 5)  [SD]ed - [D]ed = 0
 ( 6)  [SD]prst - [D]prst = 0
 ( 7)  [D]yr89 - [A]yr89 = 0
 ( 8)  [D]male - [A]male = 0
 ( 9)  [D]white - [A]white = 0
 (10)  [D]age - [A]age = 0
 (11)  [D]ed - [A]ed = 0
 (12)  [D]prst - [A]prst = 0
------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
SD           |
        yr89 |   .5239025   .0798989     6.56   0.000     .3673036    .6805014
        male |  -.7332998   .0784827    -9.34   0.000     -.887123   -.5794765
       white |  -.3911595   .1183808    -3.30   0.001    -.6231816   -.1591373
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
       _cons |   2.465362   .2389128    10.32   0.000     1.997102    2.933622
-------------+----------------------------------------------------------------
D            |
        yr89 |   .5239025   .0798989     6.56   0.000     .3673036    .6805014
        male |  -.7332998   .0784827    -9.34   0.000     -.887123   -.5794765
       white |  -.3911595   .1183808    -3.30   0.001    -.6231816   -.1591373
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
       _cons |    .630904   .2333156     2.70   0.007     .1736138    1.088194
-------------+----------------------------------------------------------------
A            |
        yr89 |   .5239025   .0798989     6.56   0.000     .3673036    .6805014
        male |  -.7332998   .0784827    -9.34   0.000     -.887123   -.5794765
       white |  -.3911595   .1183808    -3.30   0.001    -.6231816   -.1591373
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
       _cons |  -1.261854    .234018    -5.39   0.000    -1.720521   -.8031871
------------------------------------------------------------------------------
```

Notice that, by imposing the parallel lines assumption, the same parameter estimates appear multiple times, because the effects are constrained to be equal for each cut-point (the constraints applied are explicitly stated in the top part of the printout). Also, the cut-points in `ologit` are constants (with signs reversed) in `gologit2`. The LR chi2 statistics are the same as in the `ologit` model (the `lrforce` parameter told `gologit2` to report a lr chi-square rather than the Wald chi-square that Stata would report by default). In short, even though the `gologit2` and `ologit` output looks a little different, when the `gologit2 pl` parameter is used the exact same model is estimated by both. The use of the `store` option caused the results to be saved under the name "constrained" so we can use them for future hypothesis testing.

Now, we'll estimate a model in which no variables have to meet the parallel lines assumption (the `npl` parameter is explicitly specified here but it would have been used by default otherwise):

```
* Part b.  No variables constrained to meet the pl assumption.
. gologit2  warm yr89 male white age ed prst, npl lrforce store(unconstrained)

Generalized Ordered Logit Estimates               Number of obs   =       2293
                                                  LR chi2(18)     =     350.92
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2820.311                       Pseudo R2       =     0.0586

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
SD           |
        yr89 |     .95575   .1547185     6.18   0.000     .6525074    1.258993
        male |  -.3009776   .1287712    -2.34   0.019    -.5533645   -.0485906
       white |  -.5287268   .2278446    -2.32   0.020    -.9752941   -.0821595
         age |  -.0163486   .0039508    -4.14   0.000    -.0240921   -.0086051
          ed |   .1032469   .0247377     4.17   0.000     .0547619     .151732
        prst |  -.0016912   .0055997    -0.30   0.763    -.0126665     .009284
       _cons |   1.856951   .3872576     4.80   0.000      1.09794    2.615962
-------------+----------------------------------------------------------------
D            |
        yr89 |   .5363707   .0919074     5.84   0.000     .3562355     .716506
        male |   -.717995   .0894852    -8.02   0.000    -.8933827   -.5426072
       white |   -.349234   .1391882    -2.51   0.012    -.6220379     -.07643
         age |  -.0249764   .0028053    -8.90   0.000    -.0304747   -.0194782
          ed |   .0558691   .0183654     3.04   0.002     .0198737    .0918646
        prst |   .0098476   .0038216     2.58   0.010     .0023575    .0173377
       _cons |   .7198119    .265235     2.71   0.007     .1999609    1.239663
-------------+----------------------------------------------------------------
A            |
        yr89 |   .3312184   .1127882     2.94   0.003     .1101577    .5522792
        male |  -1.085618   .1217755    -8.91   0.000    -1.324294   -.8469423
       white |  -.3775375   .1568429    -2.41   0.016     -.684944    -.070131
         age |  -.0186902   .0037291    -5.01   0.000     -.025999   -.0113814
          ed |   .0566852   .0251836     2.25   0.024     .0073263    .1060441
        prst |   .0049225   .0048543     1.01   0.311    -.0045918    .0144368
       _cons |  -1.002225   .3446354    -2.91   0.004    -1.677698   -.3267523
------------------------------------------------------------------------------
```

We see that the estimates for yr89 and male (the variables which the Brant test said were most problematic) differ substantially across equations, while differences in the effects of other variables are fairly small. We can now do a global test of the proportional odds assumption by contrasting the two models we have just estimated:

```
. * Part c.  Do a global test of the parallel lines assumption
. lrtest constrained unconstrained

likelihood-ratio test                                  LR chi2(12) =     49.20
(Assumption: constrained nested in unconstrained)      Prob > chi2 =    0.0000
```

The chi-square statistic (which is similar to the Brant statistic reported earlier, but should be more accurate because it is a likelihood ratio test rather than Wald) shows that *at least one variable* does not meet the parallel lines assumption. But, *it need not mean that all fail to meet the assumption.* Hence, we can now use the `autofit` option to see whether a partial proportional odds model can fit the data. In a partial proportional odds model, some variables meet the proportional odds assumption while others do not.

```
. * Part d.  Use autofit to identify/estimate a partial proportional odds model that
fits the data
. gologit2 warm yr89 male white age ed prst, autofit lrf

------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of significance...

Step  1:  white meets the pl assumption (P Value = 0.7136)
Step  2:  ed meets the pl assumption (P Value = 0.1589)
Step  3:  prst meets the pl assumption (P Value = 0.2046)
Step  4:  age meets the pl assumption (P Value = 0.0743)
Step  5:  The following variables do not meet the pl assumption:
          yr89 (P Value = 0.00093)
          male (P Value = 0.00002)

If you re-estimate this exact same model with gologit2, instead
of autofit you can save time by using the parameter

pl(white ed prst age)

------------------------------------------------------------------------------

Generalized Ordered Logit Estimates               Number of obs   =       2293
                                                  LR chi2(10)     =     338.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -2826.6182                       Pseudo R2       =     0.0565

 ( 1)  [SD]white - [D]white = 0
 ( 2)  [SD]ed - [D]ed = 0
 ( 3)  [SD]prst - [D]prst = 0
 ( 4)  [SD]age - [D]age = 0
 ( 5)  [D]white - [A]white = 0
 ( 6)  [D]ed - [A]ed = 0
 ( 7)  [D]prst - [A]prst = 0
 ( 8)  [D]age - [A]age = 0
------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
SD           |
        yr89 |     .98368   .1530091     6.43   0.000     .6837876    1.283572
        male |  -.3328209   .1275129    -2.61   0.009    -.5827417   -.0829002
       white |  -.3832583   .1184635    -3.24   0.001    -.6154424   -.1510742
         age |  -.0216325   .0024751    -8.74   0.000    -.0264835   -.0167814
          ed |   .0670703   .0161311     4.16   0.000     .0354539    .0986866
        prst |   .0059146   .0033158     1.78   0.074    -.0005843    .0124135
       _cons |    2.12173   .2467146     8.60   0.000     1.638178    2.605282
-------------+----------------------------------------------------------------
```

```
D            |
        yr89 |    .534369   .0913937     5.85   0.000     .3552406    .7134974
        male |  -.6932772   .0885898    -7.83   0.000    -.8669099   -.5196444
       white |  -.3832583   .1184635    -3.24   0.001    -.6154424   -.1510742
         age |  -.0216325   .0024751    -8.74   0.000    -.0264835   -.0167814
          ed |   .0670703   .0161311     4.16   0.000     .0354539    .0986866
        prst |   .0059146   .0033158     1.78   0.074    -.0005843    .0124135
       _cons |   .6021625   .2358361     2.55   0.011     .1399323    1.064393
-------------+----------------------------------------------------------------
A            |
        yr89 |   .3258098   .1125481     2.89   0.004     .1052197       .5464
        male |  -1.097615   .1214597    -9.04   0.000    -1.335671   -.8595579
       white |  -.3832583   .1184635    -3.24   0.001    -.6154424   -.1510742
         age |  -.0216325   .0024751    -8.74   0.000    -.0264835   -.0167814
          ed |   .0670703   .0161311     4.16   0.000     .0354539    .0986866
        prst |   .0059146   .0033158     1.78   0.074    -.0005843    .0124135
       _cons |  -1.048137   .2393568    -4.38   0.000    -1.517268   -.5790061
------------------------------------------------------------------------------
```

The results show that 4 of the 6 variables (white, age, ed, prst) meet the parallel lines assumption. Only yr89 and male do not. This model is less restrictive than the model estimated by `ologit` (whose assumptions are violated in this case) but much more parsimonious than a non-ordinal alternative such as `mlogit`.

*Interpetation*. Of course, now that you've got these parameters, how do you interpret them? In general, you can interpret `gologit2` coefficients as coefficients from binary logit models where you have collapsed the categories of your outcome variable into two categories. Suppose your categories are numbered 1, 2, and 3. The first panel of coefficients can be interpreted as those from a binary logit regression where your dependent variable is recoded as 1 vs. 2+3. The second panel of coefficients can be interpreted as those from a binary logit regression where your dependent variable is recoded 1+2 vs. 3. Positive coefficients mean that higher values on the covariates make higher values on the dependent variable more likely.

Interpretation is particularly straightforward for those variables that meet the parallel lines assumption. From the above we can see that whites and older people tend to be less supportive of working mothers, while those who are better educated and have greater occupation prestige tend to be more supportive. The coefficients for yr89 are consistently positive but decline across cut-points. This means that respondents in 1989 were more supportive of working mothers than respondents in 1977, with the greatest differences being that 1989 respondents were less likely to put themselves in the strongly disagree and disagree categories. Conversely, the male effect is negative but gets larger across cutpoints. Hence, males tend to be less supportive of working mothers than are females, with the greatest differences being that males are less likely to place themselves in the Strongly Agree and Agree categories.

Hence, through the partial proportional odds model estimated by `gologit2`, the effects of the variables that meet the parallel lines assumption are easily interpretable (you interpret them the same way as you do in `ologit`). For other variables, an examination of the pattern of coefficients reveals insights that would be obscured or distorted if a proportional odds model were estimated instead. Conversely, an `mlogit` model might lead to similar conclusions as `gologit2` but there would be many more parameters to look at, and the increased number of parameters could cause some effects to become statistically insignificant.

Example 2: Using the `gamma` option. Here is an example from Lall and colleagues (2002). The dependent variable, hstatus, is measured on a 4 point scale with categories 4 = poor, 3 = fair, 2 = good, 1 = excellent. The independent variables are heart (0 = did not suffer from heart attack, 1 = did suffer from heart attack) and smoke (0 = does not smoke, 1 = does smoke).

**Table 5** Log odds ratios for unconstrained partial proportional odds model

| Variable | | | (Good, fair, poor) vs excellent | | (Fair, poor) vs (excellent, good) | | Poor vs (excellent, good, fair) | |
|---|---|---|---|---|---|---|---|---|
| | ln(O.R.) | s.e. ln(O.R.) | ln(O.R.) | s.e. ln(O.R.) | ln(O.R.) | s.e. ln(O.R.) | ln(O.R.) | s.e. ln(O.R.) |
| | *Constant component of log odds ratio across cut-off points* | | *Increment at cut-off points* | | | | | |
| Suffered from a heart attack (yes/no)? | 1.023 | 0.0554 | — | — | — | — | — | — |
| Do you smoke (yes/no)? | 0.1218 | 0.059 | 0 | | 0.00822 | (0.0628) | 0.3382 | (0.1006) |
| | | | *Log odds ratios at cut-off points* | | | | | |
| Do you smoke (yes/no)? | — | — | 0.1218 | (0.059) | 0.1300 | (0.0991) | 0.4600 | (0.1281) |

In the parameterization of the partial proportional odds model used in their paper, each X has a beta coefficient associated with it (called the "constant component" in the above table). In addition, each X can have M -2 Gamma coefficients (labeled above as the "Increment at cut-off points"), where M = the # of categories for Y and the Gammas represent deviations from proportionality. If the Gammas for a variable are all 0, the variable meets the proportional odds assumption. In the above, there are gammas for smoke but not heart; this means that heart is constrained to meet the proportional odds assumption but smoking is not. *In effect, then, a test of the parallel lines assumption for a variable is a test of whether its gammas equal zero.*

The parameterization used by Lall can be produced by using `gologit2`'s `gamma` option (with minor differences probably reflecting differences in the software used). Further, by using the `autofit` option, we can see whether we come up with the same final model that they do.

```
. use http://www.nd.edu/~rwilliam/stata/lall, clear
(Lall et al, 2002, Statistical Methods in Medical Research, p. 58)

. gologit2 hstatus  heart smoke, auto gamma lrf

------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of significance...

Step  1:  heart meets the pl assumption (P Value = 0.7444)
Step  2:  The following variables do not meet the pl assumption:
          smoke (P Value = 0.00044)

If you re-estimate this exact same model with gologit2, instead
of autofit you can save time by using the parameter

pl(heart)

------------------------------------------------------------------------------
```

```
Generalized Ordered Logit Estimates               Number of obs   =      12535
                                                  LR chi2(4)      =     373.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -14664.661                       Pseudo R2       =     0.0126

 ( 1)  [Excellent]heart - [Good]heart = 0
 ( 2)  [Good]heart - [Fair]heart = 0
------------------------------------------------------------------------------
     hstatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Excellent    |
       heart |   1.025339   .0551397    18.60   0.000     .9172672    1.133411
       smoke |    .127191   .0590098     2.16   0.031     .0115339    .2428482
       _cons |   1.303032   .0251244    51.86   0.000     1.253789    1.352275
-------------+----------------------------------------------------------------
Good         |
       heart |   1.025339   .0551397    18.60   0.000     .9172672    1.133411
       smoke |   .1283844   .0488556     2.63   0.009     .0326292    .2241396
       _cons |  -.8967713   .0226262   -39.63   0.000    -.9411177   -.8524248
-------------+----------------------------------------------------------------
Fair         |
       heart |   1.025339   .0551397    18.60   0.000     .9172672    1.133411
       smoke |   .4581369   .0894379     5.12   0.000     .2828418     .633432
       _cons |  -3.082652   .0463864   -66.46   0.000    -3.173568   -2.991737
------------------------------------------------------------------------------


Alternative parameterization: Gammas are deviations from proportionality
------------------------------------------------------------------------------
     hstatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Beta         |
       heart |   1.025339   .0551397    18.60   0.000     .9172672    1.133411
       smoke |    .127191   .0590098     2.16   0.031     .0115339    .2428482
-------------+----------------------------------------------------------------
Gamma_2      |
       smoke |   .0011933   .0629692     0.02   0.985    -.1222239    .1246106
-------------+----------------------------------------------------------------
Gamma_3      |
       smoke |   .3309459    .100827     3.28   0.001     .1333287    .5285631
-------------+----------------------------------------------------------------
Alpha        |
     _cons_1 |   1.303032   .0251244    51.86   0.000     1.253789    1.352275
     _cons_2 |  -.8967713   .0226262   -39.63   0.000    -.9411177   -.8524248
     _cons_3 |  -3.082652   .0463864   -66.46   0.000    -3.173568   -2.991737
------------------------------------------------------------------------------
```

The relationship between these two parameterizations is fairly straightforward. The coefficients for the first equation in the default parameterization correspond to the betas in the alternate parameterization. Gamma_2 parameters = Equation 2 – Equation 1 parameters and Gamma_3 parameters = Equation 3 – Equation 1 parameters. E.g. in equation 3 the coefficient for smoke is .458, and in equation 1 it is .127. Gamma_3 for smoke therefore equals .458 - .127 = .331. You only get gammas for variables that are NOT constrained to meet the proportional odds assumption.

The use of the `autofit` parameter confirms that Lall got it right, i.e. `autofit` produces the same partial proportional odds model that he got. But, if we wanted to just trust him, we could have estimated the same model by using the `pl` or `npl` parameters. The following two commands will each produce the same results in this case:

```
. gologit2 hstatus  heart smoke, pl(heart) gamma lrf
. gologit2 hstatus  heart smoke, npl(smoke) gamma lrf
```

Using either parameterization, the results suggest that those who have had heart attacks tend to report worse health. The same is true for smokers, but smokers are especially likely to report themselves as being in poor health as opposed to fair, good or excellent health.

A researcher might want to use the alternative gamma parameterization simply because it is standard practice in their field. But, even if you don't want to report things that way, there are several advantages to at least looking at it.

- You can see at a glance which variables are constrained to have proportional effects and which ones aren't. If there isn't a Gamma parameter, the variable is constrained to meet the proportional odds assumption.
- The printout is more parsimonious. The default parameterization will report the same values multiple times whenever a variable's effect has been constrained to be proportional. The alternate format only reports each parameter once. Note that the Model D.F. corresponds to the number of Betas and Gammas that are reported (unless additional constraints have been applied.)
- By starting with an unconstrained model, the alternate parameterization helps you to see at a glance where the potential problems in a model are. If the gammas for a variable are all statistically insignificant, it is probably safe to impose the proportionality constraint; but if one or more gammas are significant then you probably don't want to impose constraints. This could also lead to models that are even more parsimonious than those estimated by `autofit`. For example, with the Lall data,

```
. gologit2  hstatus heart smoke, lrf npl gamma

Generalized Ordered Logit Estimates               Number of obs   =      12535
                                                  LR chi2(6)      =     373.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -14664.362                       Pseudo R2       =     0.0126

[default parameterization delete]

Alternative parameterization: Gammas are deviations from proportionality
------------------------------------------------------------------------------
     hstatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Beta         |
       heart |   1.046722   .1023646    10.23   0.000     .8460913    1.247353
       smoke |   .1274032   .0590163     2.16   0.031     .0117334    .2430729
-------------+----------------------------------------------------------------
Gamma_2      |
       heart |  -.0109007    .100116    -0.11   0.913    -.2071244     .185323
       smoke |   .0012914   .0629834     0.02   0.984    -.1221537    .1247365
-------------+----------------------------------------------------------------
Gamma_3      |
       heart |  -.0821184   .1328688    -0.62   0.537    -.3425365    .1782996
       smoke |   .3305576   .1007839     3.28   0.001     .1330249    .5280903
-------------+----------------------------------------------------------------
Alpha        |
     _cons_1 |   1.302031   .0254276    51.21   0.000     1.252194    1.351868
     _cons_2 |  -.8973008   .0228198   -39.32   0.000    -.9420269   -.8525748
     _cons_3 |  -3.069089   .0494071   -62.12   0.000    -3.165925   -2.972252
------------------------------------------------------------------------------
```

We see that only gamma_3 for smoke significantly differs from 0. Ergo, we could use the `constraints` option to come up with an even more parsimonious model:

```
. constraint 1 [#1=#2]:smoke

. gologit2  hstatus heart smoke, lrf gamma pl(heart) constraint(1)

Generalized Ordered Logit Estimates               Number of obs   =      12535
                                                  LR chi2(3)      =     373.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -14664.661                       Pseudo R2       =     0.0126

 ( 1)  [Excellent]smoke - [Good]smoke = 0
 ( 2)  [Excellent]heart - [Good]heart = 0
 ( 3)  [Good]heart - [Fair]heart = 0
------------------------------------------------------------------------------
     hstatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Excellent    |
       heart |   1.025334    .055139    18.60   0.000     .9172638    1.133405
       smoke |   .1279526   .0432192     2.96   0.003     .0432446    .2126606
       _cons |     1.3029    .024137    53.98   0.000     1.255592    1.350208
-------------+----------------------------------------------------------------
Good         |
       heart |   1.025334    .055139    18.60   0.000     .9172638    1.133405
       smoke |   .1279526   .0432192     2.96   0.003     .0432446    .2126606
       _cons |  -.8966838   .0221497   -40.48   0.000    -.9400964   -.8532712
-------------+----------------------------------------------------------------
Fair         |
       heart |   1.025334    .055139    18.60   0.000     .9172638    1.133405
       smoke |   .4578386   .0880417     5.20   0.000       .28528    .6303971
       _cons |  -3.082591    .046273   -66.62   0.000    -3.173284   -2.991898
------------------------------------------------------------------------------


Alternative parameterization: Gammas are deviations from proportionality
------------------------------------------------------------------------------
     hstatus |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Beta         |
       heart |   1.025334    .055139    18.60   0.000     .9172638    1.133405
       smoke |   .1279526   .0432192     2.96   0.003     .0432446    .2126606
-------------+----------------------------------------------------------------
Gamma_2      |
       smoke |  -3.05e-16   6.59e-10    -0.00   1.000    -1.29e-09    1.29e-09
-------------+----------------------------------------------------------------
Gamma_3      |
       smoke |    .329886   .0838936     3.93   0.000     .1654577    .4943144
-------------+----------------------------------------------------------------
Alpha        |
     _cons_1 |     1.3029    .024137    53.98   0.000     1.255592    1.350208
     _cons_2 |  -.8966838   .0221497   -40.48   0.000    -.9400964   -.8532712
     _cons_3 |  -3.082591    .046273   -66.62   0.000    -3.173284   -2.991898
------------------------------------------------------------------------------
```

Note that `gologit2` is not smart enough to know that Gamma_2 should not be in there (it knows to omit it when either `pl` or `npl` have forced the parameter to be 0, but not when the `constraint` option has been used) but this is just a matter of aesthetics; everything is being done correctly. The fit for this model is virtual identical to the fit of the model that included gamma_2 (LR chi2 = 373.10 in both), so we conclude that this more parsimonious parameterizations is justified. Hence, while the assumptions of the 2-parameter proportional odds model estimated by `ologit` are violated by these data, we can get a model that fits whose assumptions are not violated simply by allowing one gamma parameter to differ from 0.

Example 3: svy estimation. The Stata 8 survey data manual presents an example where `svyologit` is used for an analysis of the NHANES II dataset. The variable health contains self-reported health status, where 1 = poor, 2 = fair, 3 = average, 4 = good, and 5 = excellent. `gologit2` can analyze survey data by including the `svy` parameter. Data must be `svyset` first. The original example includes variables for age and age^2. To make the results a little more interpretable, I have created centered age (c_age) and centered age^2 (c_age2) (you need to install Ben Jann's `center` command to do things the same way I did). This does not change the model selected or the model fit. Note that the `lrforce` option has no effect when doing svy estimation since likelihood ratio chi-squares are not appropriate in such cases.

```
. use http://www.stata-press.com/data/r8/nhanes2f.dta
. center age
. gen c_age2=c_age^2
. gologit2 health female black c_age c_age2, svy auto

------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of significance...

Step  1:  black meets the pl assumption (P Value = 0.2310)
Step  2:  The following variables do not meet the pl assumption:
          female (P Value = 0.00280)
          c_age (P Value = 0.00000)
          c_age2 (P Value = 0.00004)

If you re-estimate this exact same model with gologit2, instead
of autofit you can save time by using the parameter

pl(black)

------------------------------------------------------------------------------

Generalized Ordered Logit Estimates

pweight:  finalwgt                                  Number of obs    =     10335
Strata:   stratid                                   Number of strata =        31
PSU:      psuid                                     Number of PSUs   =        62
                                                    Population size  = 1.170e+08
                                                    F(  13,     19)  =     52.24
                                                    Prob > F         =    0.0000

 ( 1)  [poor]black - [fair]black = 0
 ( 2)  [fair]black - [average]black = 0
 ( 3)  [average]black - [good]black = 0
```

```
------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
poor         |
      female |   .1681817   .1034177     1.63   0.114    -.0427401    .3791034
       black |  -1.008808   .0836513   -12.06   0.000    -1.179416   -.8382006
       c_age |  -.0617038    .003537   -17.45   0.000    -.0689175     -.05449
      c_age2 |   .0006893   .0003049     2.26   0.031     .0000674    .0013111
       _cons |   2.962162   .1373065    21.57   0.000     2.682124    3.242201
-------------+----------------------------------------------------------------
fair         |
      female |  -.1545385   .0680284    -2.27   0.030    -.2932834   -.0157937
       black |  -1.008808   .0836513   -12.06   0.000    -1.179416   -.8382006
       c_age |  -.0525504    .002082   -25.24   0.000    -.0567966   -.0483042
      c_age2 |   -.000028   .0001237    -0.23   0.822    -.0002802    .0002242
       _cons |   1.718909   .0765319    22.46   0.000     1.562821    1.874997
-------------+----------------------------------------------------------------
average      |
      female |  -.1576817   .0596012    -2.65   0.013     -.279239   -.0361243
       black |  -1.008808   .0836513   -12.06   0.000    -1.179416   -.8382006
       c_age |  -.0409575   .0017576   -23.30   0.000    -.0445422   -.0373728
      c_age2 |   8.91e-06   .0000882     0.10   0.920     -.000171    .0001889
       _cons |   .1705633   .0534477     3.19   0.003     .0615559    .2795707
-------------+----------------------------------------------------------------
good         |
      female |  -.2133394   .0636419    -3.35   0.002    -.3431379   -.0835408
       black |  -1.008808   .0836513   -12.06   0.000    -1.179416   -.8382006
       c_age |  -.0356466   .0020002   -17.82   0.000     -.039726   -.0315672
      c_age2 |  -.0004546   .0001311    -3.47   0.002    -.0007221   -.0001872
       _cons |  -.9136691   .0574078   -15.92   0.000    -1.030753   -.7965851
------------------------------------------------------------------------------
```

In this example, only one variable, black, meets the parallel lines assumption. Blacks tend to report worse health than do whites. For females, the pattern is more complicated. They are less likely to report poor health than are males (see the positive female coefficient in the poor panel), but they are also less likely to report higher levels of health (see the negative female coefficients in the other panels), i.e. women tend to be less at the extremes of health than men are. Such a pattern would be obscured in a straight proportional odds model. The effect of age is more extreme on lower levels of health.

Example 4. gologit 1.0 compatibility. Some post-estimation commands – specifically, the `spost` routines of Long and Freese – currently work with the original `gologit` but not `gologit2`. That should change in the future. For now, you can use the `v1` parameter to make the stored results from `gologit2` compatible with `gologit` 1.0. (Note, however, that this may make the results non-compatible with post-estimation routines written for `gologit2`, including `predict`.) Using the working mother's data again,

```
. use http://www.nd.edu/~rwilliam/stata/ordwarm2, clear
(77 & 89 General Social Survey)

. * Use the v1 option to save internally stored results in gologit 1.0 format
. quietly gologit2  warm yr89 male white age ed prst, pl(yr89 male) lrf v1
```

```
. * Use spost routines.  Get predicted probability for a 30 year old average white
woman in 1989
. prvalue, x(male=0 yr89=1 age=30) rest(mean)

gologit: Predictions for warm

Predicted probabilities for each category:
  Pr(y=SD|x):         0.0473
  Pr(y=D|x):          0.1699
  Pr(y=A|x):          0.4487
  Pr(y=SA|x):         0.3340

         yr89       male      white        age          ed        prst
x=          1          0   .8765809         30   12.218055   39.585259

. * Now do 70 year old average black male in 1977
. prvalue, x(male=1 yr89=0 age=70) rest(mean)

gologit: Predictions for warm

Predicted probabilities for each category:
  Pr(y=SD|x):         0.2565
  Pr(y=D|x):          0.4699
  Pr(y=A|x):          0.2093
  Pr(y=SA|x):         0.0644

         yr89       male      white        age          ed        prst
x=          0          1   .8765809         70   12.218055   39.585259
```

These "representative" cases show us that a 30 year old average white woman in 1989 was much more supportive of working mothers than a 70 year old average black male in 1977. Various other `spost` routines that work with the original `gologit` (not all do) can also be used, e.g. `prtab`.

Example 5: The predict command. In addition to the standard options (`xb`, `stdp`, `stddp`) the `predict` command supports the `pr` option (abbreviated p) for predicted probabilities; `pr` is the default option if nothing else is specified. For example,

```
. quietly gologit2  warm yr89 male white age ed prst, pl(yr89 male) lrf

. predict p1 p2 p3 p4
(option p assumed; predicted probabilities)

. list  p1 p2 p3 p4 in 1/10

     +-------------------------------------------+
     |       p1         p2         p3         p4 |
     |-------------------------------------------|
  1. | .1083968   .2843347   .4195861   .1876824 |
  2. | .2057451   .4859219    .236662   .0716709 |
  3. | .1120911   .3004282   .4181407     .16934 |
  4. | .2099544   .4283575   .2636952   .0979929 |
  5. | .1407257   .3221328   .3887267   .1484148 |
     |-------------------------------------------|
  6. | .2279584   .3338488   .3237104   .1144824 |
  7. | .1652819   .3070716   .3804251   .1472214 |
  8. | .1100771   .3058248   .4105159   .1735823 |
  9. | .0930135   .2593877   .4754793   .1721194 |
 10. | .1997068   .3816947   .3235006    .095098 |
     +-------------------------------------------+
```

## Author

Richard Williams
Notre Dame Department of Sociology
Richard.A.Williams.5@ND.Edu
http://www.nd.edu/~rwilliam

## Acknowledgements

Vincent Kang Fu of the Utah Department of Sociology wrote `gologit` 1.0 and graciously allowed Richard Williams to incorporate parts of its source code and documentation in `gologit2`.

The documentation for Stata 8.2's `mlogit` command and the program `mlogit_p` were major aids in developing the `gologit2` documentation and in adding support for the predict command. Much of the code is adapted from *Maximum Likelihood Estimation with Stata*, Second Edition, by William Gould, Jeffrey Pitblado and William Sribney.

Sarah Mustillo, Dan Powers, J. Scott Long, and Nick Cox provided stimulating and helpful comments.

## References

Fu, Vincent. 1998. "Estimating Generalized Ordered Logit Models." *Stata Technical Bulletin* 8:160-164.

Lall, R., S.J. Walters, K. Morgan, and MRC CFAS Co-operative Institute of Public Health. 2002. "A Review of Ordinal Regression Models Applied on Health-Related Quality of Life Assessments." *Statistical Methods in Medical Research* 11:49-67.

Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, 2nd Edition.*

Peterson, Bercedis and Frank E. Harrell Jr. 1990. "Partial Proportional Odds Models for Ordinal Response Variables." *Applied Statistics* 39(2):205-217.

## Suggested citation if using gologit2 in published work (at least until something more formal comes along):

Williams, Richard. 2005. "Gologit2: A Program for Generalized Logistic Regression/ Partial Proportional Odds Models for Ordinal Variables." Retrieved May 12, 2005 (http://www.nd.edu/~rwilliam/stata/gologit2.pdf ).