Estimating Heterogeneous Choice Models with Stata

Richard Williams, Notre Dame Sociology, rwilliam@ND.Edu
oglm support page: http://www.nd.edu/~rwilliam/oglm/index.html

West Coast Stata Users Group Meeting, October 25, 2007

See the accompanying PowerPoint presentation for a discussion of the materials in this handout. Thanks to J. Scott Long, Robert Hauser and Megan Andrew for sharing the data sets used in these analyses.

Overview

When a binary or ordinal regression model incorrectly assumes that error variances are the same for all cases, the standard errors are wrong and (unlike OLS regression) the parameter estimates are biased. Heterogeneous choice/ location-scale models explicitly specify the determinants of heteroskedasticity in an attempt to correct for it. These models are also useful when the variability of underlying attitudes is itself of substantive interest.

This paper illustrates how Williams' user-written routine `oglm` (Ordinal Generalized Linear Models) can be used to estimate heterogeneous choice and related models. It further shows how two other models that have appeared in the literature – Allison's (1999) model for comparing logit and probit coefficients across groups, and Hauser and Andrew's (2006) logistic response model with partial proportionality constraints (LRPPC) – are special cases of the heterogeneous choice model and/or algebraically equivalent to it, and can also be estimated with `oglm`.

The Heterogeneous Choice (aka Location-Scale) Model

With heterogeneous choice models, the dependent variable can be ordinal or binary. For a binary dependent variable, the model (Keele & Park, 2006) can be written as

$$\Pr(y_i = 1) = g\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) = g\left(\frac{x_i\beta}{\exp(\ln(\sigma_i))}\right) = g\left(\frac{x_i\beta}{\sigma_i}\right)$$

In the above formula,

- g stands for the link function (in this case logit; probit is also commonly used, and other options are possible, such as the complementary log-log, log-log and cauchit).
- x is a vector of values for the ith observation. The x's are the explanatory variables and are said to be the determinants of the choice, or outcome.
- z is a vector of values for the ith observation. The z's define groups with different error variances in the underlying latent variable. The z's and x's need not include any of the same variables, although they can.
- β and γ are vectors of coefficients. They show how the x's affect the choice and the z's affect the variance (or more specifically, the log of σ).
- The numerator in the above formula is referred to as the choice equation, while the denominator is the variance equation. These are also referred to as the location and scale equations. Also, the choice equation includes a constant term but the variance equation does not.
- The conventional logit and probit models, which do not have variance equations, are special cases of the above.

In Stata, heterogeneous choice models can be estimated via the user-written routine `oglm`.

Example 1: Using heterogeneous choice models when the assumptions of the ordered logit model are violated.

Long and Freese (2006) 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 (1SD), 2 = Disagree (2D), 3 = Agree (3A), and 4 = Strongly Agree (4SA).
- 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).

```
. use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
. ologit  warm yr89 male white age ed prst, nolog

Ordered logistic regression                     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                     -2.933622   -1.997102
       /cut2 |   -.630904   .2333155                     -1.088194    -.173614
       /cut3 |   1.261854   .2340179                      .8031873    1.720521
------------------------------------------------------------------------------

. estimates store ologit

. 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.
```

```
. oglm  warm yr89 male white age ed prst, het(yr89 male) store(oglm) hc

Heteroskedastic Ordered Logistic Regression       Number of obs   =       2293
                                                  LR chi2(8)      =     331.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -2830.2563                       Pseudo R2       =     0.0552

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
choice       |
        yr89 |   .4531574   .0686839     6.60   0.000     .3185394    .5877755
        male |  -.6345402   .0697638    -9.10   0.000    -.7712748   -.4978057
       white |  -.3087676    .102739    -3.01   0.003    -.5101323   -.1074029
         age |  -.0186098   .0021728    -8.56   0.000    -.0228684   -.0143512
          ed |   .0535685   .0135944     3.94   0.000     .0269239     .080213
        prst |   .0052866     .00278     1.90   0.057    -.0001622    .0107353
-------------+----------------------------------------------------------------
variance     |
        yr89 |  -.1486188   .0458169    -3.24   0.001    -.2384183   -.0588192
        male |  -.1909211    .044807    -4.26   0.000    -.2787412   -.1031011
-------------+----------------------------------------------------------------
       /cut1 |  -2.151122   .2114069   -10.18   0.000    -2.565472   -1.736772
       /cut2 |  -.5696264   .1992724    -2.86   0.004    -.9601932   -.1790596
       /cut3 |   1.066508   .2022099     5.27   0.000     .6701839    1.462832
------------------------------------------------------------------------------

. lrtest ologit oglm, stats force

Likelihood-ratio test                                 LR chi2(2)  =     29.31
(Assumption: ologit nested in oglm)                   Prob > chi2 =    0.0000

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
      ologit |   2293    -2995.77   -2844.912      9     5707.825    5759.463
        oglm |   2293    -2995.77   -2830.256     11     5682.513    5745.626
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Example 2: Allison's (1999) model for group comparisons.

Using data originally collected by Long, Allison (Sociological Methods and Research, 1999) analyzes a data set of 301 male and 177 female biochemists. These scientists were assistant professors at graduate universities at some point in their careers. Allison uses logistic regressions to predict the probability of promotion to associate professor. The units of analysis are person-years rather than persons, with 1,741 person-years for men and 1,056 person-years for women. In his analysis,

- the dependent variable is coded 1 if the scientist was promoted to associate professor in that person-year, 0 otherwise. (After promotion no additional person-years are added for that case.)
- Duration is the number of years since the beginning of the assistant professorship
- undergraduate selectivity is a measure of the selectivity of the colleges where scientists received their bachelor's degrees
- number of articles is the cumulative number of articles published by the end of each person-year
- job prestige is a measure of prestige of the department in which scientists were employed.

TABLE 1: Results of Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists

	Men		*Women*		*Ratio of*	*Chi-Square*
Variable	*Coefficient*	SE	*Coefficient*	SE	*Coefficients*	*for Difference*
Intercept	−7.6802***	.6814	−5.8420***	.8659	.76	2.78
Duration	1.9089***	.2141	1.4078***	.2573	.74	2.24
Duration squared	−0.1432***	.0186	−0.0956***	.0219	.67	2.74
Undergraduate selectivity	0.2158***	.0614	0.0551	.0717	.25	2.90
Number of articles	0.0737***	.0116	0.0340**	.0126	.46	5.37*
Job prestige	−0.4312***	.1088	−0.3708*	.1560	.86	0.10
Log likelihood	−526.54		−306.19			

*$p < .05$. **$p < .01$. ***$p < .001$.

In Table 2, Allison adds a parameter to the model he calls delta. Delta adjusts for differences in residual variation across groups. His article includes Stata code for estimating his model, and Hoetker's `complogit` routine (available from SSC) will also estimate it.

TABLE 2: Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists, Disturbance Variances Unconstrained

	All Coefficients Equal		*Articles Coefficient Unconstrained*	
Variable	*Coefficient*	SE	*Coefficient*	SE
Intercept	7.4913***	.6845	−7.3655***	.6818
Female	−0.93918**	.3624	−0.37819	.4833
Duration	1.9097***	.2147	1.8384***	.2143
Duration squared	−0.13970***	.0173	−0.13429***	.01749
Undergraduate selectivity	0.18195**	.0615	0.16997***	.04959
Number of articles	0.06354***	.0117	0.07199***	.01079
Job prestige	−0.4460***	.1098	−0.42046***	.09007
$\hat{\delta}$	−0.26084*	.1116	−0.16262	.1505
Articles × Female			−0.03064	.0173
Log likelihood	−836.28		−835.13	

*$p < .05$. **$p < .01$. ***$p < .001$.

Allison's model with delta is actually a special case of a heterogeneous choice model, where the dependent variable is a dichotomy and the variance equation includes a single dichotomous variable. For example, here is the `oglm` replication of Allison's first model in his Table 2:

```
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/tenure01.dta", clear
(Gender differences in receipt of tenure (Scott Long 06Jul2006))
. * Allison limited the sample to the first 10 years untenured
. keep if pdasample
(148 observations deleted)

. oglm  tenure female year yearsq select  articles prestige , het(female)

Heteroskedastic Ordered Logistic Regression       Number of obs   =       2797
                                                  LR chi2(7)      =     413.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -836.28235                       Pseudo R2       =     0.1981

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tenure       |
      female |  -.9391907   .3705243    -2.53   0.011    -1.665405   -.2129763
        year |   1.909544   .1996935     9.56   0.000     1.518152    2.300936
      yearsq |  -.1396868   .0169425    -8.24   0.000    -.1728935   -.1064801
      select |   .1819201   .0526572     3.45   0.001     .0787139    .2851264
    articles |   .0635345    .010219     6.22   0.000     .0435055    .0835635
    prestige |  -.4462073    .096904    -4.60   0.000    -.6361356   -.2562791
-------------+----------------------------------------------------------------
lnsigma      |
      female |   .3022305    .146178     2.07   0.039     .0157268    .5887341
-------------+----------------------------------------------------------------
       /cut1 |   7.490506   .6596628    11.36   0.000      6.19759    8.783421
------------------------------------------------------------------------------

. * Compute Allison's delta
. display (1 - exp(.3022305))/ exp(.3022305)
-.26083233
```

Example 3. Hauser & Andrew's (2006) Logistic Response Model with Partial Proportionality Constraints.

Mare applied a logistic response model to school continuation, restricting the base population at risk for each successive transition to those who had completed the prior educational transition. Hauser & Andrew (Sociological Methodology, 2006) replicate & extend Mare's analysis using the same data he did, the 1973 Occupational Changes in a Generation (OCG) survey data. See their paper for a complete description of the data and variables.

Hauser and Andrew argue that the relative effects of some (but not all) background variables are the same at each transition, and that multiplicative scalars express proportional change in the effect of those variables across successive transitions. Specifically, Hauser & Andrew estimate two new types of models. The first is called the *logistic response model with proportionality constraints* (LRPC):

$$\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{j0} + \lambda_j \sum_k \beta_k X_{ijk}$$

Hauser and Andrew also propose a less restrictive model, which they call the *logistic response model with partial proportionality constraints* (LRPPC):

$$\log_e\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{j0} + \lambda_j \sum_{k=1}^{k'} \beta_k X_{ijk} + \sum_{k'+1}^{K} \beta_{jk} X_{ijk}$$

Even though the rationales behind the models are totally different, the heterogeneous choice models estimated by `oglm` produce identical fits to the LRPC and LRPPC models estimated by Hauser and Andrew. Hauser & Andrew summarize their models in Table 5 of their paper:

TABLE 5
Fit of Selected Models of Educational Transitions: 1973 Occupational Changes in a Generation Survey

Model	Description	Log-Likelihood	DF for Model	Model Chi-square	Contrast	Contrast Chi-square	Contrast BIC	Pseudo R-squared
1	Fit the grand mean	−46830.8	0	—		—		0
2	An intercept for each transition	−38674.3	5	16313.0	2 vs. 1	16313.0	16256.0	0.17
3	An intercept for each transition and constant social background effects	−34333.3	13	24995.0	3 vs. 2	8682.0	8590.8	0.27
4	An intercept for each transition and proportional social background effects	−33529.7	19	26602.2	4 vs. 3	1607.3	1538.9	0.28
5	An intercept for each transition, constant effects of socioeconomic variables, interactions of BROKEN, FARM, and SOUTH with transition	−34112.0	28	25437.6	5 vs. 3	442.6	271.7	0.27
6	An intercept for each transition, proportional effects of socioeconomic variables, interactions of BROKEN, FARM, and SOUTH with transition	−33399.7	34	26862.1	6 vs. 5	1424.6	1356.2	0.29
7	Saturated model: Intercepts for each transition and interactions of all social background variables with transition	−33332.2	53	26997.2	7 vs. 6	135.1	−81.4	0.29

Here are `oglm`'s algebraically-equivalent models. Note that the fits are identical to those reported by Hauser and Andrew.

```
------------------------------------------------------------------------------------------------
                    m1          m2          m3          m4          m5          m6          m7
------------------------------------------------------------------------------------------------
N                88768       88768       88768       88768       88768       88768       88768
ll           -46830.8    -38674.3    -34333.3    -33529.7    -34112.0    -33399.7    -33332.2
df_m                0           5          13          18          28          33          53
chi2          5.82e-11     16313.0     24995.0     26602.2     25437.6     26862.1     26997.2
r2_p          6.66e-16       0.174       0.267       0.284       0.272       0.287       0.288
------------------------------------------------------------------------------------------------
```

Five of the Hauser & Andrew models can be estimated via conventional logistic regression. Model 4 (LRPC) and Model 6 (LRPPC) can be estimated via Stata code they present in their paper.

Following is the `oglm` code for estimating models that are algebraically equivalent to m4 and m6. In both m4 and m6, dummy variables for transition are included in the variance equation. In m6, the non-ses variables are freed from constraints by including interaction terms for each non-ses variable with each transition.

```
*** Model 4: An intercept for each transition & proportional social background effects
* This is the first hetero choice model (equivalent to H & A's LRPC).
quietly oglm outcome trans2 trans3 trans4 trans5 trans6 dunc sibsttl9 ln_inc_trunc
edhifaom edhimoom broken farm16 south, het(trans2 trans3 trans4 trans5 trans6)
store(m4)

*** Model 6: An intercept for each transition, proportional effects of
* socioeconomic variables, interactions of broken, farm, and south with transition.
* This is the second hetero choice model (equivalent to H & A's LRPPC).
quietly oglm outcome trans2 trans3 trans4 trans5 trans6 broken farm16 south
trans2Xbroken trans2Xfarm16 trans2Xsouth trans3Xbroken trans3Xfarm16 trans3Xsouth
trans4Xbroken trans4Xfarm16 trans4Xsouth trans5Xbroken trans5Xfarm16 trans5Xsouth
trans6Xbroken trans6Xfarm16 trans6Xsouth dunc sibsttl9 ln_inc_trunc edhifaom edhimoom,
het(trans2 trans3 trans4 trans5 trans6) store(m6)
```

Example 4: Using Stepwise Selection as a Diagnostic/ Model Building Device

With `oglm`, stepwise selection can be used for either the choice or variance equation. If you want to do it for the variance equation, the `flip` option can be used to reverse the placement of the choice and variance equations in the command line. In the following, we use stepwise selection to build the variance equation with Allison's data.

```
. sw, pe(.01) lr: oglm  tenure female year yearsq select  articles prestige ,
eq2(female  year yearsq select  articles prestige) flip

LR test               begin with empty model
p = 0.0000 <  0.0100  adding  articles

Heteroskedastic Ordered Logistic Regression       Number of obs   =       2797
                                                  LR chi2(7)      =     428.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -828.81224                       Pseudo R2       =     0.2052

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tenure       |
      female |  -.4179259   .1742083    -2.40   0.016     -.759368   -.0764838
        year |   2.108752   .2486633     8.48   0.000     1.621381    2.596123
      yearsq |  -.1542213   .0208579    -7.39   0.000    -.1951019   -.1133406
      select |   .1744644   .0598623     2.91   0.004     .0571364    .2917924
    articles |   .0628407   .0157851     3.98   0.000     .0319026    .0937789
    prestige |  -.6118689   .1307262    -4.68   0.000    -.8680877   -.3556502
-------------+----------------------------------------------------------------
lnsigma      |
    articles |    .030149   .0091448     3.30   0.001     .0122256    .0480724
-------------+----------------------------------------------------------------
       /cut1 |   7.959556   .7637106    10.42   0.000      6.46271    9.456401
------------------------------------------------------------------------------
```

Example 5: Using Marginal Effects and mfx2 to Compare Models

While there are various ways of assessing whether the assumptions of the ordered logit model have been violated, it is more difficult to assess how worrisome violations are, i.e. how much harm is done if you do things the "wrong" way? One way of addressing these concerns is by comparing the marginal effects produced by different models. The `oglm`, `mfx2`, and `esttab` commands (all available from SSC) provide an easy way of doing this. Returning to the working mothers data,

```
. use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta"
(77 & 89 General Social Survey)

. * Baseline ordered logit model
. quietly oglm  warm yr89 male white age ed prst, store(ologit)
. quietly mfx2, stub(ologit)
. * Heterogeneous choice model with yr89 and male in the variance equation
. quietly oglm  warm yr89 male white age ed prst, store(oglm) het( yr89 male)
. quietly mfx2, stub(oglm)
. esttab ologit_mfx oglm_mfx, mtitle(ologit oglm) nonum not

--------------------------------------------
                   ologit            oglm
--------------------------------------------
1SD
yr89              -0.0499***      -0.0786***
male               0.0746***       0.0355**
white              0.0345***       0.0319***
age               0.00214***      0.00213***
ed                -0.00664***     -0.00613***
prst             -0.000600       -0.000605
--------------------------------------------
2D
yr89              -0.0775***      -0.0618***
male                0.105***        0.137***
white              0.0594**        0.0543**
age               0.00319***      0.00318***
ed                -0.00990***     -0.00916***
prst             -0.000895       -0.000904
--------------------------------------------
3A
yr89               0.0539***       0.0995***
male              -0.0814***      -0.0344*
white             -0.0356***      -0.0333***
age               -0.00241***     -0.00240***
ed                 0.00746***      0.00691***
prst              0.000675        0.000682
--------------------------------------------
4SA
yr89               0.0735***       0.0409**
male              -0.0979***       -0.138***
white             -0.0583**       -0.0529**
age               -0.00293***     -0.00291***
ed                 0.00908***      0.00839***
prst              0.000821        0.000828
--------------------------------------------
N                    2293            2293
--------------------------------------------
* p<0.05, ** p<0.01, *** p<0.001
```

Example 6: Other uses of oglm. Here are other examples of `oglm`'s capabilities.

* Basic models. By default, `oglm` will estimate the same models as `ologit`. The `store` option is convenient for saving results if you want to contrast different models.

```
use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
oglm warm yr89 male white age ed prst
oglm warm yr89 male white age ed prst, store(m1)
oglm warm yr89 male white age ed prst, robust
```

* The `predict` command.

```
use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
quietly oglm warm yr89 male white age ed prst
predict p1 p2 p3 p4
```

* Constrained logistic regression. `logit`, `ologit`, `probit` and `oprobit` provide other and generally faster means for estimating non-heteroskedastic models with logit and probit links; but none of these commands currently supports the use of linear constraints, such as two variables having equal effects. `oglm` can be used for this purpose. For example,

```
use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
recode warm (1 2 = 0)(3 4 = 1), gen(agree)
* Constrain the effects of male and white to be equal
constraint 1 male = white
oglm agree yr89 male white age ed prst, lrf store(constrained) c(1)
oglm agree yr89 male white age ed prst, store(unconstrained)
lrtest constrained unconstrained
```

* Other link functions. By default, `oglm` uses the logit link. If you prefer, however, you can specify probit, complementary log log, log log or log links. In the following example, the same model is estimated using each of the links supported by `oglm.`

```
use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
oglm warm yr89 male white age ed prst, link(l)
oglm warm yr89 male white age ed prst, link(p)
oglm warm yr89 male white age ed prst, link(c)
oglm warm yr89 male white age ed prst, link(ll)
oglm warm yr89 male white age ed prst, link(ca)
```

* Prefix commands. `oglm` supports many of Stata 9's prefix commands. For example,

```
use http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta, clear
sw, pe(.05): oglm warm yr89 male
xi: oglm warm yr89 i.male
nestreg: oglm warm (yr89 male white age) (ed prst)
use http://www.stata-press.com/data/r8/nhanes2f.dta, clear
svy: oglm health female black age age2
```