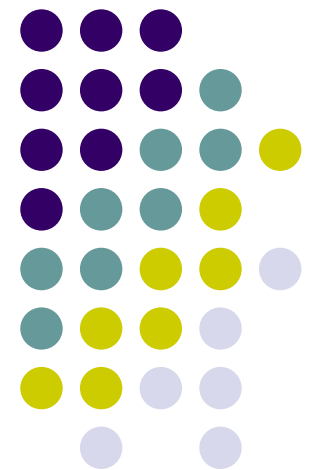# Estimating Heterogeneous Choice Models with Stata

Richard Williams
Notre Dame Sociology
rwilliam@ND.Edu
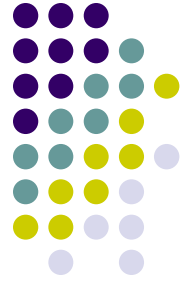West Coast Stata Users Group Meetings
October 25, 2007

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 presentation illustrates how Williams' user-written Stata 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 be estimated with oglm.

The Heterogeneous Choice (aka Location-Scale) Model

- Can be used for binary or ordinal models
- Two equations, choice & variance
- Binary case (see handout p. 1 for an explanation):

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

Example 1: Ordered logit assumptions 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 2 = Disagree, 3 = Agree, and 4 = Strongly Agree.

  - Explanatory variables are yr89 (survey year; 0 = 1977, 1 = 1989), male (0 = female, 1 = male), white (0 = nonwhite, 1 = white), age (in years), ed (years of education), and prst (occupational prestige scale).
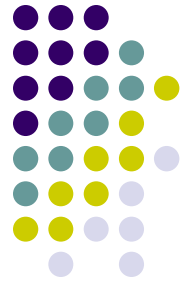
- See handout p. 2 for ologit results
- Results are easy to interpret
- But are they correct? Brant test suggests they may not be. yr89 and male are especially problematic
- Heterogeneous choice model fits much better (handout p. 3)
- The variance equation tells us there was less residual variability across time and that the residual variance was smaller for men than for women.

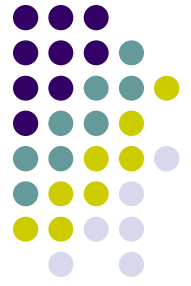# Example 2: Allison's (1999) model for group comparisons
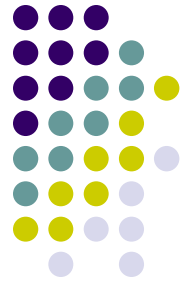
- Allison (Sociological Methods and Research, 1999) analyzes a data set of 301 male and 177 female biochemists.

- 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.

- As his Table 1 shows (p. 4 of handout), the effect of number of articles on promotion is about twice as great for males (.0737) as it is females (.0340).
- BUT, Allison warns, women may have more heterogeneous career patterns, and unmeasured variables affecting chances for promotion may be more important for women than for men.

- Comparing coefficients across populations using logistic regression has much the same problems as comparing standardized coefficients across populations using OLS regression.

  - In logistic regression, standardization is inherent. To identify coefficients, the variance of the residual is always fixed at 3.29.

  - Hence, unless the residual variability is identical across populations, the standardization of coefficients for each group will also differ.

- Ergo, in Table 2 (Handout p. 4), 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.

- The delta-hat coefficient value –.26 in Allison's Table 2 (first model) tells us that the standard deviation of the disturbance variance for men is 26 percent lower than the standard deviation for women.

  - This implies women have more variable career patterns than do men, which causes their coefficients to be lowered relative to men when differences in variability are not taken into account, as in the original logistic regressions.

- The interaction term for Articles x Female is NOT statistically significant
- Allison concludes "The apparent difference in the coefficients for article counts in Table 1 does not necessarily reflect a real difference in causal effects. It can be readily explained by differences in the degree of residual variation between men and women."

- See Williams (2007) for a detailed critique of Allison. For now, we focus on the Stata side of things.
- 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 that also appears in the choice equation.
- See handout p. 5 for the corresponding oglm code and output. Simple algebra converts oglm's sigma into Allison's delta

- As Williams (2007) notes, there are important advantages to turning to the broader class of heterogeneous choice models that can be estimated by oglm

- Dependent variables can be ordinal rather than binary. This is important, because ordinal vars have more information. Studies show that ordinal vars work better than binary vars when using hetero choice
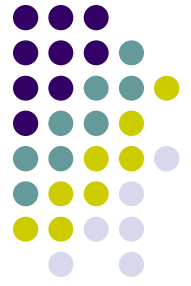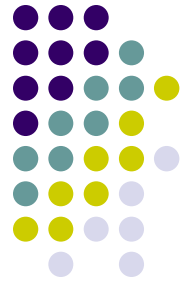
- The variance equation need not be limited to a single binary grouping variable. This is very important!!! It can be easily shown that a mis-specified variance equation can be worse than no variance equation at all!

# Example 3. Hauser & Andrew's (2006) LRPPC Model.
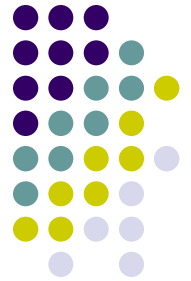
- Mare applied a logistic response model to school continuation

- Contrary to prior supposition, Mare's estimates suggested the effects of some socioeconomic background variables declined across six successive transitions including completion of elementary school through entry into graduate school.

- 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.
- Rather than analyzing each educational transition separately as Mare did, Hauser & Andrew estimate a single model across all educational transitions.
  - They take the original data set of 21,682 white men and restructure it into 88,768 person-transition records
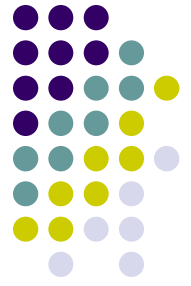
- Hauser and Andrew argue that the relative effects of some (but not all) background variables are the same at each transition, and that multiplicative scalars express proportional change in the effect of those variables across successive transitions.

- Specifically, Hauser & Andrew estimate two new types of models. The first is called the *logistic response model with proportionality constraints* (LRPC – see p. 5 of handout):

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

- The λj introduce proportional increases or decreases in the *βk* across transitions; thus the LRPC model implies proportional changes in main effects across transitions.
- Instead of having to estimate a different set of betas for each transition, you estimate a single set of betas, along with one λj proportionality factor for each transition (λ_1 is constrained to equal 1)
  - For example, if you have 10 independent variables and 6 transitions, you will have 60 coefficients and 6 intercepts if you estimate a separate model for each transition.
  - But, if the proportionality constraints hold, you only need to estimate 10 coefficients, 5 λs, and 6 intercepts.
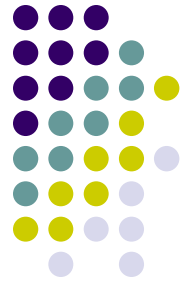
- The proportionality constraints would hold if, say, the coefficients for the 2nd transition were all 2/3 as large as the corresponding coefficients for the first transition, the coefficients for the 3rd transition were all half as large as for the first transition, etc.
- Put another way, if the model holds, you can think of the items as forming a composite scale
- If it holds, the model is both parsimonious and substantively interesting.

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

- This model maintains the proportionality constraints for some variables, while allowing the effects of other variables to freely differ across transitions

- For example, Hauser & Andrew say the LRPPC could apply to Mare's analysis where effects of socioeconomic variables appear to decline across transitions while those of farm origin, one-parent family, and Southern birth vary in other ways.

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

- Hauser & Andrew note, however, that "one cannot distinguish empirically between the hypothesis of uniform proportionality of effects across transitions and the hypothesis that group differences between parameters of binary regressions are artifacts of heterogeneity between groups in residual variation." (p. 8)
- Similarly, Mare (2006, p.32) notes that "the constants of proportionality, λj, are estimable, but their values incorporate both differences across equations in the effects of the regressors and also differences in the variances of the underlying dependent variables."
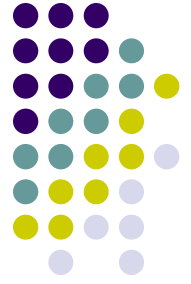
- Indeed, even though the rationales behind the models are totally different, the heterogeneous choice models estimated by oglm produce identical fits to the LRPC and LRPPC models estimated by Hauser and Andrew.

- See pp. 6-7 of the handout for Hauser and Andrew's original analysis and oglm's algebraically equivalent analysis

- The models are algebraically equivalent
  - The LRPC and LRPPC's lambda is the reciprocal of oglm's sigma
  - Hauser & Andrew actually report decrements to lambda across transitions. In the two transition case, these are identical to Allison's delta
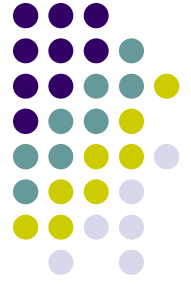
- HOWEVER, the substantive interpretations are very different
  - The LRPC says that effects differ across transitions by scale factors
  - The algebraically-equivalent heterogeneous choice model says that effects do not differ across transitions; they only appear to differ when you estimate separate models because the variances of residuals change across transitions
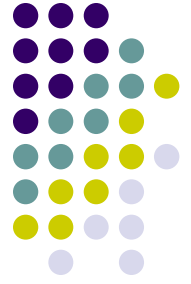
- Empirically, there is no way to distinguish between the two; but, you could make substantive arguments for the positions favored by Mare, Hauser & Andrew
  - As Hauser & Andrew's Table 2 shows, the observed variances of most of the SES variables tend to decline across transitions
  - BUT, according to the hetero choice model, the residual variances increase substantially across transitions. Indeed, if the model is to be believed, the residual standard deviation is about 11 times as large for the 6th transition as it is for the 1st.
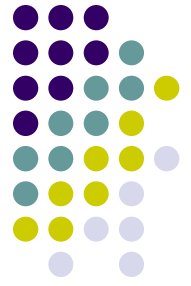
- So, what makes more sense?
  - Effects of SES vars decline across transitions?
  - Or, residual variances skyrocket while the variances of observed SES variables generally go down?
  - Effects declining seems more reasonable, although it could be a combination of the two.

- But, if the residual variances actually declined across transitions, like the observed variances generally did, the effects of SES during later transitions are actually being *over-estimated* by both Mare and Hauser & Andrew. That is, the decline in SES effects may be even greater than they claim.
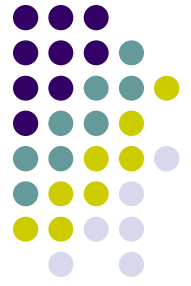
- In any event, there can be little arguing that the effects of SES relative to other influences decline across transitions.

  - The only question is whether this is because the effects of SES decline, or because the influence of other (omitted) variables go up.

Example 4: Using Stepwise Selection as a Diagnostic/ Model Building Device
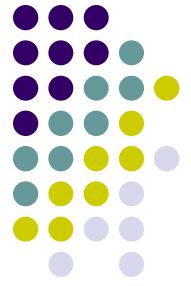
- Stepwise selection procedures have been heavily criticized, and rightfully so.

- However, they can be useful for exploratory purposes

- In the case of heterogeneous choice models, they can also help to identify those variables that cause the assumption of homoskedastic errors to be violated.

- 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.
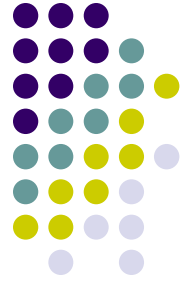
- As p. 7 of the handout shows, in Allison's Biochemist data, the only variable that enters into the variance equation using oglm's stepwise selection procedure is number of articles.
  - This is not surprising: there may be little residual variability among those with few articles (with most getting denied tenure) but there may be much more variability among those with more articles (having many articles may be a necessary but not sufficient condition for tenure).

- Hence, while heteroskedasticity may be a problem with these data, it may not be for the reasons first thought.
- HOWEVER, remember that heteroskedasticity problems often reflect other problems in a model. Variables could be missing, or variables may need to be transformed in some way, e.g. logged.
- So, even if you don't want to ultimately use a heterogeneous choice model, you may still wish to estimate one as a diagnostic check on whether or not there are problems with heteroskedasticity.
- When and if such problems are found, you can decide how best to handle them.

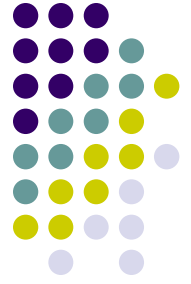# Example 5: Using Marginal Effects and mfx2 to Compare Models

- 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?

- People often go with the "wrong" way on the grounds that sign and significance of effects are the same across methods, and the "wrong" way is easier to interpret

- But, the "wrong" way may hide important substantive differences.

- 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.

- See p. 8 of the handout for an example of how this can aid in the analysis of the working mother's data.

- The analysis shows that the ordered logit approach creates a misleading impression of the effects of gender and year.
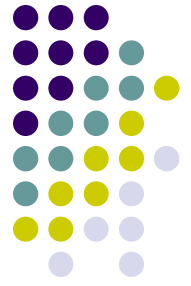
- The marginal effects for white, age, ed and prst are very similar in both models and for all outcomes. These are the four variables that were not included in the variance equation of the heterogeneous choice model.

- The story is very different for the variables yr89 and male. Both models agree that there was a shift toward more positive attitudes between 1977 and 1989, but they describe that shift differently.

- The heterogeneous choice model says that the main reason attitudes became more favorable across time was because people shifted from extremely negative positions to more moderate positions; there was only a fairly small increase in people strongly agreeing that women should work.

- The ordered logit model, on the other hand, understates how much people moved from an extremely negative position and overstates how much they became extremely positive.
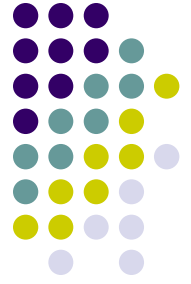
- The models also provide different pictures of the effect of gender on attitudes.
  - Again, the ordered logit model is creating a misleading image of why men were less supportive of working mothers
  - It isn't so much that men were extremely negative in their attitudes, it is more a matter of them being less likely than women to be extremely supportive.

Example 6: Other uses of oglm
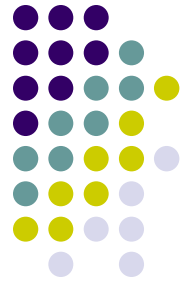
- See the oglm help and p. 9 of the handout for other capabilities of oglm. These include
  - Ability to estimate the same models as logit, ologit, probrit, oprobit, hetprob, cloglog, and others
  - Can compute predicted probabilities
  - Linear constraints, e.g. white = female, can be imposed and tested
  - Support for multiple link functions – logit, probit, loglog, cloglog, cauchit
  - Support for prefix commands, e.g. svy, nestreg, xi, sw

Selected References

- Allison, Paul. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods and Research* 28(2): 186-208.
- Hauser, Robert M. and Megan Andrew. 2006. "Another Look at the Stratification of Educational Transitions: The Logistic Response Model with Partial Proportionality Constraints." *Sociological Methodology* 36 (1), 1–26.
- Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata, Second Edition.* College Station, Texas: Stata Press.

- Mare, Robert D. 2006. "Response: Statistical Models of Educational Stratification - Hauser And Andrew's Models For School Transitions." *Sociological Methodology* 36 (1), 27–37.

- Williams, Richard. 2007. "Using Heterogeneous Choice Models To Compare Logit and Probit Coefficients Across Groups." Working Paper, last revised August 2007. Currently available at

  http://www.nd.edu/~rwilliam/oglm/RW_Hetero_Choice.pdf

- For more information on oglm and for related work on heterogeneous choice models, see

  http://www.nd.edu/~rwilliam/oglm/index.html