Ordinal regression models: Problems, solutions, and problems with the solutions

Richard Williams
Notre Dame Sociology
rwilliam@ND.Edu
German Stata User Group Meetings
June 27, 2008

Overview

- Ordered logit/probit models are among the most popular ordinal regression techniques
- The assumptions of these models, however, are often violated
  - Errors may not be homoskedastic – which can have far more serious consequences than is usually the case with OLS regression
  - The parallel lines/proportional odds assumption often does not hold

- This paper shows how heterogeneous choice/location scale models (estimated via oglm) and generalized ordered logit/probit models (estimated via gologit2) can often address these concerns in ways that are more parsimonious and easier to interpret than is the case with other suggested alternatives.

- At the same time, the paper cautions that these methods sometimes raise their own concerns that researchers need to be aware of and know how to deal with.

Problem 1: Heteroskedastic Error Variances

- 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.

Example: 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.

- As his Table 1 shows, 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.

Allison's solution for the problem

- Ergo, in his Table 2, Allison adds a parameter to the model he calls delta. Delta adjusts for differences in residual variation across groups.

- 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.”

A Broader Solution: Heterogeneous Choice Models

- 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.

The Heterogeneous Choice (aka Location-Scale) Model

- Can be used for binary or ordinal models
- Two equations, choice & variance
- Binary case :

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

- 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 for the corresponding oglm code and output. Simple algebra converts oglm's sigma into Allison's delta

- As Williams (forthcoming) 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.
- Further, heterogeneous choice methods can be used as a diagnostic device even if you don't want to ultimately use a heterogeneous choice model

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.

- As 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.

- For example, for the Allison problem, Maarten Buis suggested allowing for a nonlinear effect of # of articles.
  - Adding articles^2 significantly improves fit and makes the coefficient in the variance equation insignificant.

- 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.

- Also, a stepwise procedure can be used to see whether other plausible models (besides the one specified by your theory) are worth considering.

Problems with heterogeneous choice models

- There are several potential problems with heterogeneous choice models researchers should be aware of

Problem: Model Misspecification

- Buis: “The heterogeneous choice model seems to me a very fragile model: you estimate a model for both the effect of the observed variables and the errors, and you use your model for the errors to correct the effects of the observed variables. Any fault in your model will mean the errors are off, leading to faults in your model for those errors, which in turn will feed back into the estimates of all other parameters.”

- Keele & Park, and Williams (forthcoming) raise similar concerns
- The handout presents a series of simulations. In these simulations,
  - Errors were homoskedastic, but group membership was included in the variance equation anyway
  - Effects of variables differed across groups

- In the simulations,
  - Differences in coefficients were generally erroneously attributed to differences in residual variation
  - Differences in coefficients were generally mis-estimated, often leading to highly misleading substantive conclusions

- Keele and Park further warn that even a correctly specified model can suffer from “fragile” identification. Dichotomous DVs and multicollinearity across equations make the problem more likely.
- Oglm’s ability to use ordinal variables (which contain more information) and to specify multiple variables in the variance equation may help to reduce these concerns
- Still, the researcher needs to think through the model carefully, and consider whether alternative specifications lead to substantially different results

Problem: Radically different interpretations are possible

- Another issue to be aware of with heterogeneous choice models is that radically different interpretations of the results are possible

- Further, there is no straightforward empirical way of choosing between these interpretations, because the results are algebraically equivalent

Example: Hauser & Andrew's (2006) LRPC 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

- They argue that the *relative* effects of some background variables are the same at each transition

- Specifically, Hauser & Andrew estimate two new types of models. I'll focus on the first, the *logistic response model with proportionality constraints* (LRPC):

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

- 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)
  - 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

- 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)

- Indeed, even though the rationales behind the models are totally different, the heterogeneous choice models estimated by oglm produce identical fits to the models estimated by Hauser and Andrew.
  - The models are algebraically equivalent
  - The LRPC's lambda is the reciprocal of oglm's sigma
  - The handout illustrates these equivalencies and shows how to estimate the Hauser & Andrew models with oglm

- 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

- 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.

Problem II: Parallel lines/ proportional odds does not hold

- This problem is probably more widely understood – but nonetheless often ignored in practice
- An example will best illustrate the problem. Three models are presented:
  - Ordered logit
  - Unconstrained generalized ordered logit (gologit)
  - Constrained generalized ordered logit (gologit2), aka Partial Proportional Odds

Example: Proportional Odds Violated

- (Adapted from Long & Freese, 2003 – 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."
  - 1 = Strongly Disagree (SD)
  - 2 = Disagree (D)
  - 3 = Agree (A)
  - 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)
  - prst (occupational prestige scale).

Model 1: Ordered logit

- These results are relatively straightforward, intuitive and easy to interpret.
- But, while the results may be straightforward, intuitive, and easy to interpret, are they correct? Are the assumptions of the ologit model met?
- The following Brant test suggests they are not.

Brant test shows assumptions violated

```
. 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.
```

How are the assumptions violated?

- **brant, detail**

```
Estimated coefficients from j-1 binary regressions

                y>1         y>2         y>3
 yr89     .9647422   .56540626   .31907316
 male   -.30536425  -.69054232  -1.0837888
white   -.55265759  -.31427081  -.39299842
  age    -.0164704  -.02533448  -.01859051
   ed    .10479624   .05285265   .05755466
 prst   -.00141118   .00953216   .00553043
_cons    1.8584045   .73032873  -1.0245168
```

- This is a series of binary logistic regressions. First it is 1 versus 2,3,4; then 1 & 2 versus 3 & 4; then 1, 2, 3 versus 4
- If proportional odds/ parallel lines assumptions were not violated, all of these coefficients (except the intercepts) would be the same except for sampling variability.

Model 2: Unconstrained gologit

- Note that the gologit results are very similar to what we got with the series of binary logistic regressions and can be interpreted the same way.
- The gologit model can be written as

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta_j)}{1+[\exp(\alpha_j + X_i\beta_j)]}, \text{ j} = 1, 2, ..., \text{M} - 1$$

- The ologit model is a special case of the gologit model, where the betas are the same for each j (NOTE: ologit actually reports cut points, which equal the negatives of the alphas used here)

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta)}{1+[\exp(\alpha_j + X_i\beta)]}, \text{ j} = 1, 2, ..., \text{M} - 1$$

Model 3: Partial Proportional Odds

- A key enhancement of gologit2 is that it allows some of the beta coefficients to be the same for all values of j, while others can differ. i.e. it can estimate partial proportional odds models. For example, in the following the betas for X1 and X2 are constrained but the betas for X3 are not.

$$P(Y_i > j) = \frac{\exp(\alpha_j + X1_i\beta1 + X2_i\beta2 + X3_i\beta3_j)}{1 + [\exp(\alpha_j + X1_i\beta1 + X2_i\beta2 + X3_i\beta3_j)]}, \text{ j} = 1, 2, ..., \text{M} - 1$$

- Either mlogit or unconstrained gologit can be overkill – both generate many more parameters than ologit does.
  - All variables are freed from the proportional odds constraint, even though the assumption may only be violated by one or a few of them
- gologit2, with the *autofit* option, will only relax the parallel lines constraint for those variables where it is violated

Interpretation of the gologit2 results

- Effects of the constrained variables (white, age, ed, prst) can be interpreted pretty much the same as they were in the earlier ologit model. For yr89 and male, the differences from before are largely a matter of degree.
  - People became more supportive of working mothers across time, but the greatest effect of time was to push people away from the most extremely negative attitudes.
  - For gender, men were less supportive of working mothers than were women, but they were especially unlikely to have strongly favorable attitudes.

Concerns with the gologit model

- While the gologit model has many attractive features, there are many concerns researchers need to be aware of

Concern 1: Unconstrained model does not require ordinality

- As Clogg & Shihadeh (1994) point out, the totally unconstrained model arguably isn't even ordinal

- You can rearrange the categories, and fit can be hardly affected

Concern II: Estimated probabilities can go negative

- Unlike other categorical models, estimated probabilities can be negative.
- This was addressed by McCullaph & Nelder, Generalized Linear Models, 2nd edition, 1989, p. 155:

  "The usefulness of non-parallel regression models is limited to some extent by the fact that the lines must eventually intersect. Negative fitted values are then unavoidable for some values of x, though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the x-space, this flaw in the model need not be serious."

- Probabilities might go negative in unlikely or impossible X ranges, e.g. when years of education is negative
- Multiple tests with 10s of thousands of cases typically resulted in only 0 to 3 negative predicted probabilities.
- Seems most problematic with small samples, complicated models, analyses where the data are being spread very thin
  - they might be troublesome regardless - gologit2 could help expose problems that might otherwise be overlooked
- Can also get negative predicted probabilities when measurement of the outcome isn't actually ordinal

- gologit2 now checks to see if any in-sample predicted probabilities are negative.
  - It is still possible that plausible values not in-sample could produce negative predicted probabilities.
- You may want to use some other method if there are a non-trivial number of negative predicted probabilities and you are otherwise confident in your models and data.

Concern III: How do you interpret the results???

- One rationale for ordinal regression models is that there is an underlying, continuous y* that reflects the dependent variable we are interested in.
- y* is unobserved, however. Instead, we observe y, which is basically a collapsed/grouped version of the unobserved y*.
  - High Income, Moderate Income and Low Income are a collapsed version of a continuous Income variable
  - Some ranges of attitudes can be collapsed into a 5 category scale ranging from Strongly Disagree to Strongly Agree
- As individuals cross thresholds (aka cut-points) on y*, their value on the observed y changes

- Does the whole idea of an underlying y* go out the window once you allow a single non-proportional effect? If so, how do you interpret the model?
  - In an ordered logit (ologit) model, you only have one predicted value for y*
  - But in a gologit model, once you have a single non-parallel effect, you have M-1 linear predictions (similar to mlogit)

Interpretation 1: gologit as non-linear probability model

- As Long & Freese (2006, p. 187) point out "The ordinal regression model can also be developed as a nonlinear probability model without appealing to the idea of a latent variable."

- Ergo, the simplest thing may just be to interpret gologit as a non-linear probability model that lets you estimate the determinants & probability of each outcome occurring. Forget about the idea of a y*

- Other interpretations, however, can preserve or modify the idea of an underlying y*

Interpretation 2: State-dependent reporting bias - gologit as measurement model

- As noted, the idea behind y* is that there is an unobserved continuous variable that gets collapsed into the limited number of categories for the observed variable y.

- HOWEVER, respondents have to decide how that collapsing should be done, e.g. they have to decide whether their feelings cross the threshold between “agree” and “strongly agree,” whether their health is “good” or “very good,” etc.

- Respondents do NOT necessarily use the same frame of reference when answering, e.g. the elderly may use a different frame of reference than the young do when assessing their health

- Other factors can also cause respondents to employ different thresholds when describing things
  - Some groups may be more modest in describing their wealth, IQ or other characteristics

- In these cases the underlying latent variable may be the same for all groups; but the thresholds/cut points used may vary.
  - Example: an estimated gender effect could reflect differences in measurement across genders rather than a real gender effect on the outcome of interest.

- Lindeboom & Doorslaer (2004) note that this has been referred to as state-dependent reporting bias, scale of reference bias, response category cut-point shift, reporting heterogeneity & differential item functioning.

- If the difference in thresholds is constant (index shift), proportional odds will still hold
  - EX: Women's cutpoints are all a half point higher than the corresponding male cutpoints
  - ologit could be used in such cases
- If the difference is not constant (cut point shift), proportional odds will be violated
  - EX: Men and women might have the same thresholds at lower levels of pain but have different thresholds for higher levels
  - A gologit/ partial proportional odds model can capture this

- If you are confident that some apparent effects reflect differences in measurement rather than real differences in effects, then
  - Cutpoints (and their determinants) are substantively interesting, rather than just “nuisance” parameters
  - The idea of an underlying y* is preserved (Determinants of y* are the same for all, but cutpoints differ across individuals and groups)

- Key advantage: This could greatly improve cross-group comparisons, getting rid of artifactual differences caused by differences in measurement.

- Key Concern: Can you really be sure the coefficients reflect measurement and not real effects, or some combination of real & measurement effects?

- Theory may help – if your model strongly claims the effect of gender should be zero, then any observed effect of gender can be attributed to measurement differences.

- But regardless of what your theory says, you may at least want to acknowledge the possibility that apparent effects could be “real” or just measurement artifacts.

Interpretation 3: The outcome is multi-dimensional

- A variable that is ordinal in some respects may not be ordinal or else be differently-ordinal in others. E.g. variables could be ordered either by direction (Strongly disagree to Strongly Agree) or intensity (Indifferent to Feel Strongly)

- Suppose women tend to take less extreme political positions than men.
  - Using the first (directional) coding, an ordinal model might not work very well, whereas it could work well with the 2nd (intensity) coding.
  - But, suppose that for every other independent variable the directional coding works fine in an ordinal model.

- Our choices in the past have either been to (a) run ordered logit, with the model really not appropriate for the gender variable, or (b) run multinomial logit, ignoring the parsimony of the ordinal model just because one variable doesn't work with it.
- With gologit models, we have option (c) – constrain the vars where it works to meet the parallel lines assumption, while freeing up other vars (e.g. gender) from that constraint.

- This interpretation suggests that there may actually be multiple y*'s that give rise to a single observed y

- NOTE: This is very similar to the rationale for the multidimensional stereotype logit model estimated by slogit.

Interpretation 4: The effect of x on y depends on the value of y

- There are actually many situations where the effect of x on y is going to vary across the range of y.
  - EX: A 1-unit increase in x produces a 5% increase in y
  - So, if y = $10,000, the increase will be $500. But if y = $100,000, the increase will be $5,000.

- If we were using OLS, we might address this issue by transforming y, e.g. takes its log, so that the effect of x was linear and the same across all values of the transformed y.
- But with ordinal methods, we can't easily transform an unobserved latent variable; so with gologit we allow the effect of x to vary across values of y.
- This suggests that there is an underlying y*; but because we can't observe or transform it we have to allow the regression coefficients to vary across values of y instead.

- Hedeker and Mermelstein (1998) also raise the idea that the categories of the DV may represent stages, e.g. pre-contemplation, contemplation, and action.

- An intervention might be effective in moving people from pre-contemplation to contemplation, but be ineffective in moving people from contemplation to action.

- If so, the effects of an explanatory variable will not be the same across the K-1 cumulative logits of the model

- Substantive example: Boes & Winkelman, 2004:

  “Completely missing so far is any evidence whether the magnitude of the income effect depends on a person’s happiness: is it possible that the effect of income on happiness is different in different parts of the outcome distribution? Could it be that “money cannot buy happiness, but buy-off unhappiness” as a proverb says? And if so, how can such distributional effects be quantified?”

For more information, see:

http://www.nd.edu/~rwilliam/gologit2

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