Models for Ordered and Unordered Categorical Variables
Objectives

- Introduce models for multi-category outcomes
- Briefly discuss multinomial logit (probit) models
- Briefly discuss ordinal logit (probit) models
- Show examples in Stata
- Discuss practical issues, extensions, etc.
Models for Multi-Category Outcomes

- These models can be viewed as extensions of binary logit and binary probit regression.
- The dependent variable has three or more categories and is nominal or ordinal.
- Multinomial logit and ordered logit models are two of the most common models.
Multinomial Logit (Probit)

• Multinomial logit (probit) models
  ○ *Nominal* outcomes – no intrinsic order (qualitative)
  ○ Three or more unordered categories

• Examples:
  ○ Smoking status – never, current, former smoker
  ○ Marital status – married, divorced, widowed, never married
Multinomial Logit (Probit) Model

- Estimates a series of binary logit (probit) models
- One group is chosen to be the base (reference) category for the other groups (estimates equations for \( k - 1 \) groups)
- Example: If never smokers are the base category, then two models are estimated:
  - Current smokers vs. Never smokers
  - Former smokers vs. Never smokers
Stata Example: Multinomial Logit

- The data are from the NHIS Adult Sample Files (2009)
- **Outcome:** Smoking Status – Never Smoked (Base Category), Current Smoker, Former Smoker
- **Predictors:**
  - Education: <High School, High School, Some College, College (Ref.)
  - Race-Ethnicity: NH White (Ref.), NH Black, Hispanic,
  - Age in years
- **Stata Code:** “mlogit smk3 lths hs scol nhb hispanic age, base(o) rrr”
  - “base(o)” tells Stata that the comparison group is never smokers
  - “rrr” tells Stata to display relative risk ratios
### Stata Example: Multinomial Logit Output

```stata
.mlogit smk3 lths hs scol nhb hispanic age, base(0) rrr nolog cformat(%8.3f)
```

Multinomial logistic regression

|                | RRR  | Std. Err. |       z | P>|z| | [95% Conf. Interval] |
|----------------|------|-----------|---------|------|----------------------|
| **Never_Smoked** |      |           |         |      |                      |
| Current_Smoker  |      |           |         |      |                      |
| lths            | 4.861| 0.291     | 26.38   | 0.000| 4.322                | 5.468                |
| hs              | 4.105| 0.215     | 26.98   | 0.000| 3.705                | 4.549                |
| scol            | 2.732| 0.143     | 19.20   | 0.000| 2.465                | 3.027                |
| nhb             | 0.666| 0.029     | -9.21   | 0.000| 0.610                | 0.726                |
| hispanic        | 0.330| 0.016     | -22.58  | 0.000| 0.300                | 0.363                |
| age             | 0.988| 0.001     | -12.66  | 0.000| 0.986                | 0.990                |
| _cons           | 0.299| 0.019     | -19.50  | 0.000| 0.265                | 0.337                |
| Former_Smoker   |      |           |         |      |                      |
| lths            | 1.212| 0.066     | 3.54    | 0.000| 1.090                | 1.348                |
| hs              | 1.263| 0.057     | 5.22    | 0.000| 1.157                | 1.379                |
| scol            | 1.333| 0.057     | 6.67    | 0.000| 1.225                | 1.450                |
| nhb             | 0.551| 0.026     | -12.53  | 0.000| 0.502                | 0.605                |
| hispanic        | 0.554| 0.027     | -12.27  | 0.000| 0.504                | 0.609                |
| age             | 1.032| 0.001     | 33.10   | 0.000| 1.030                | 1.034                |
| _cons           | 0.081| 0.005     | -42.13  | 0.000| 0.072                | 0.091                |
```

Number of obs = 25578
LR chi2(12) = 3322.78
Prob > chi2 = 0.0000
Pseudo R2 = 0.0661
• The risk of being a current vs. never smoker is 4.86 times greater for persons without a high school diploma relative to college graduates net of race-ethnicity and age.

• The risk of being a former vs. never smoker is about 33% [(0.666 – 1)*100)] lower for blacks relative to whites when education and age are held constant.

• The risk of being a former vs. never smoker increases by about 3% (RRR = 1.03) with each additional year of age controlling for education and race-ethnicity.
Ordered Logit (Probit) Models

- Ordered logit (probit) models
  - *Ordinal* outcomes – inherently ordered categories
  - **Problem**: Distance between adjacent categories is unknown
  - **Solution**: Treat the ordinal scale as though it represents a latent interval/ratio scale

- Examples:
  - Self-Rated Health – poor, fair, good, very good, excellent
Ordered Logit (Probit) Models

- Estimates the cumulative probability of being in one category versus all lower or higher categories

- Proportionality Assumption – the distance between each category is equivalent (a.k.a., proportional odds assumption)
  - This assumption often is violated in practice
  - Need to test if this assumption holds (can use a “Brant test”)
  - Violating this assumption may or may not really “matter”
  - Refer to Long & Freese (2006) for more information
Stata Example: Ordered Logit Model

- The data are from the NHIS Adult Sample Files (2009)
- **Outcome**: Self-Rated Health, where 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor
- **Predictors**:
  - Education: <High School, High School, Some College, College (Ref.)
  - Race-Ethnicity: NH White (Ref.), NH Black, Hispanic,
  - Age in years
- **Stata Code**: “ologit health lths hs scol nhb hispanic age, or”
  - The model is predicting the log odds of reporting worse health
  - “or” tells Stata to display proportional odds ratios
. ologit health lths hs scol nbh hispanic age, cformat(%8.4f) or

Iteration 0:  log likelihood = -36758.193
Iteration 1:  log likelihood = -34707.594
Iteration 2:  log likelihood = -34684.358
Iteration 3:  log likelihood = -34684.323
Iteration 4:  log likelihood = -34684.323

Ordered logistic regression

Number of obs = 25578
LR chi2(6) = 4147.74
Prob > chi2 = 0.0000
Pseudo R2 = 0.0564

Log likelihood = -34684.323

| health     | Odds Ratio | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|------------|------------|-----------|-------|------|----------------------|
| lths       | 3.9714     | 0.1540    | 35.56 | 0.000| 3.6807 - 4.2850      |
| hs         | 2.4231     | 0.0779    | 27.53 | 0.000| 2.2751 - 2.5807      |
| scol       | 1.7792     | 0.0555    | 18.48 | 0.000| 1.6737 - 1.8912      |
| nbh        | 1.6333     | 0.0512    | 15.64 | 0.000| 1.5359 - 1.7369      |
| hispanic   | 1.1839     | 0.0380    | 5.26  | 0.000| 1.1118 - 1.2607      |
| age        | 1.0306     | 0.0007    | 44.54 | 0.000| 1.0292 - 1.0319      |
| /cut1      | 1.0515     | 0.0405    |       | 0.9720 |
| /cut2      | 2.5554     | 0.0430    |       | 2.4711 |
| /cut3      | 4.1457     | 0.0477    |       | 4.0523 |
| /cut4      | 5.7435     | 0.0566    |       | 5.6326 |
The odds of reporting poor vs. fair, good, very good, and excellent health are 3.97 times greater for persons who did not graduate high school in comparison to persons with a college degree net of race-ethnicity and age.

Each additional year of age is associated with 3.1% (OR = 1.036) increase in the odds of reporting poor vs. fair, good, very good, and excellent health when education and race-ethnicity are held constant.

The cut-points (or thresholds) Stata used to differentiate between the adjacent levels of self-rated health are at the bottom (cut1, cut2, etc.)
Testing for Proportionality

- Once again, the ordered logit (probit) model assumes that the distance between each category of the outcome is proportional.

- In practice, violating this assumption may or may not alter your substantive conclusions. You need to test whether this is the case.

- A Brant test can be used to test whether the proportional odds (i.e., parallel lines) assumption holds.
  - This is available as a user-added post-estimation command in Stata.
  - To download this command type “findit brant” in Stata.
  - Once downloaded, you can type “brant” immediately after you estimate a ordered logit model (“ologit”) to perform the test.
The Brant test indicates that the influence of education and race-ethnicity are not proportional across each category of self-rated health. Note, that the association between age and self-rated health is proportional though.

```
. brant

Brant Test of Parallel Regression Assumption

<table>
<thead>
<tr>
<th>Variable</th>
<th>chi2</th>
<th>p&gt;chi2</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>213.47</td>
<td>0.000</td>
<td>18</td>
</tr>
<tr>
<td>lths</td>
<td>106.74</td>
<td>0.000</td>
<td>3</td>
</tr>
<tr>
<td>hs</td>
<td>34.73</td>
<td>0.000</td>
<td>3</td>
</tr>
<tr>
<td>scol</td>
<td>17.12</td>
<td>0.001</td>
<td>3</td>
</tr>
<tr>
<td>nhb</td>
<td>35.55</td>
<td>0.000</td>
<td>3</td>
</tr>
<tr>
<td>hispanic</td>
<td>34.51</td>
<td>0.000</td>
<td>3</td>
</tr>
<tr>
<td>age</td>
<td>4.48</td>
<td>0.214</td>
<td>3</td>
</tr>
</tbody>
</table>
```

A significant test statistic provides evidence that the parallel regression assumption has been violated.
When the Proportionality Assumption is Violated...

- **Option 1:** Do nothing. Use ordered logistic regression because the practical implications of violating this assumption are minimal.

- **Option 2:** Use a multinomial logit model. This frees you of the proportionality assumption, but it is less parsimonious and often dubious on substantive grounds.

- **Option 3:** Dichotomize the outcome and use binary logistic regression. This is common, but you lose information and it could alter your substantive conclusions.

- **Option 4:** Use a model that does not assume proportionality. Increasingly, this is common. Two user-submitted Stata commands fit these kinds of models:
  - “gologit2” – generalized ordered logit models (see Williams 2007, *Stata Jn.*)
  - “oglm” – heterogeneous choice models (see Williams 2010, *Stata Jn.*)

- **Recommendation:** Try all the above and decide what to do based on your results.