MLE: Categorical and Limited Dependent Variables
Unit 2: Ordered, Multinomial, Count, and Limited Dependent Variables
2c. Utility Models and Multinomial Probit

PS2730-2021

Week 7

Professor Steven Finkel

- Assumption of Independence of Irrelevant Alternatives underlies both MNL and CL models we've considered. What to do when it is violated? In those instances some outcome are "close substitutes" for each other and inclusion/exclusion of one outcome would affect the relative odds of others
- Two possibilities (among others):
  – Nested Logit – assume that some outcomes cluster together in different "nests", and that decisions take place both across and within nests. Specify the nest structure in advance and estimate an extension of the MNL model, whereby the outcomes are assumed IIA **within** but not **across** nests. (Interesting but not frequently used in political science).
  – Multinomial Probit – specify utility from different choice outcomes, allow errors from different outcomes to correlate in order to capture degree to which outcomes are "closer" to each other
- Both models derive from extensions of the more general Random Utility Model (RUM) of decision-making, which can be/was also used to motivate earlier logit/probit models

- In the RUM, assume individual *i* can choose from *J* alternatives, with the utility of choice *m* being:

$$U_{im} = X_i\beta_m + \gamma Z_i + \varepsilon_{im}$$

where X are individual-specific variables and Z are choice-specific. The U_{im} is analogous to the latent Y* we used to motivate the probit model earlier.

- Assume that individuals choose the *m* alternative with the largest overall utility, so that

$$\Pr(\text{m}) = \Pr\ (U_{im} > U_{ij}) \qquad \forall\ \text{m} \neq \text{j}$$

$$\Pr(\text{m}) = \Pr\ ((X_i\beta_m + \gamma Z_{im} + \varepsilon_{im}) > (X_i\beta_j + \gamma Z_{ij} + \varepsilon_{ij}))$$

$$\Pr(\text{m}) = \Pr((\varepsilon_{ij} - \varepsilon_{im}) < ((X_i\beta_m + \gamma Z_{im}) - (X_i\beta_j + \gamma Z_{ij})))$$

- Individuals choose *m* over other alternatives to the extent that the differences in the unobservables ε of the alternatives are smaller than the differences in the systematic portions of the alternatives' utilities

- RUM models typically start with assumption that the ε follow a "Type 1 extreme value" distribution (Google this if you want!), where $f(\varepsilon_{im}) = exp^{\varepsilon_{im}} exp^{-exp^{\varepsilon_{im}}}$

- If IIA holds, then the model reduces to the CL/MNL model:

$$\frac{\Pr(y=m)}{\Pr(y=j)} = \frac{\dfrac{\exp^{(Z_m\gamma + X_i\beta_m)}}{\sum_{j=1}^{J}\exp^{(Z_j\gamma + X_i\beta_m)}}}{\dfrac{\exp^{(Z_j\gamma + X_i\beta j)}}{\sum_{j=1}^{J}\exp^{(Z_j\gamma + X_i\beta_j)}}} = \frac{\exp^{(Z_m\gamma + X_i\beta_m)}}{\exp^{(Z_j\gamma + X_i\beta_j)}} = \exp^{(Z_m - Z_j)\gamma + X_i(\beta_m - \beta_j)}$$

- But if IIA violated, then the errors are not independent. Different models proceed by making different assumptions about the errors

Nested Logit

- One way to think about violations of IIA is to conceive of a choice as *sequential*, e.g., individuals in a multimember, multiparty electoral system first choosing the party they support, and then which of the party's candidate's they wish to vote for. So there would a "tree-like" decision process, first among the parties representing different "nests", and second among the candidates within the parties/nests. So choices within the nests would be "closer" than choices across nests, thus violating IIA

- Or, the decision need not be sequential but nevertheless represent a choice between alternatives which can be grouped theoretically into different "nests".

- For example, the Bush/Clinton/Perot choice might be modeled as either:

1) as two candidates in a “major party” nest (Bush/Clinton) and the other in a “minor party” nest (Perot); the choice proceeds “as if” it was major/minor, and then, if major, the particular candidate within major, even if the actual decision for given individuals is not sequential

2) as one “incumbent party nest” (Bush) and the other two as a “non-incumbent nest” (Clinton/Perot); the choice proceeds accordingly

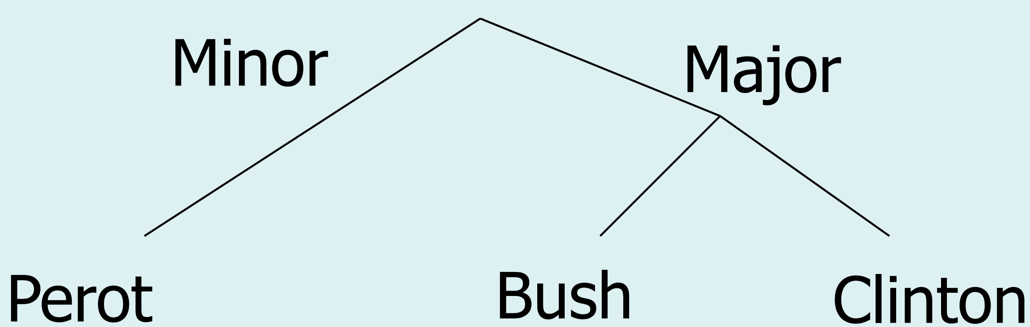
3) a “more conservative nest” (Bush/Perot) and a “more liberal nest” (Clinton); the choice proceeds accordingly

- Nested logit: **assume IIA within nests but not across nests**

- Briefly (we won’t have time to develop this model further):
- In NL, the probability of choosing an alternative *m* from a nest A is

$$\Pr(Y_i = m) = \Pr(A_d) * \Pr(Y_i = m \mid m \in A_d)$$

- or the probability of choosing the nest A which contains *m* multiplied by the probability of choosing *m* from among the choices in nest A_d
- We first model with multinomial logit which nest is selected, and then, within the nest, we have a multinomial logit model for selecting a particular outcome
- For example: following possibility (2) above, we could specify the upper nest as “MAJOR-MINOR” candidate, with Bush/Clinton as MAJOR and Perot as MINOR, and then work out the sequential MNLs (here 2 bivariates).



Multinomial Probit

- Alternative model to NL also relaxes IIA but doesn't depend on possible sequential decisions or nests
- If assume that error terms in RUM are multivariate normally distributed, arrive at the **multinomial probit** model that allow errors to correlate across choices
- Can estimate the degree to which unobservables affecting the choices are related to one another, and estimate effects of independent variables taking those error correlations into account
- Very difficult computationally to estimate and depends on having choice-specific variable(s) in the model along with some identifying assumptions
- Error interpretations are also not straightforward
- Use this model with caution!

- As before, start with utility for choice *m* and selection of choice with greatest net utility

$$U_{im} = X_i\beta_m + \gamma Z_i + \varepsilon_{im}$$

$$\Pr(\mathrm{m}) = \Pr\ (U_{im} > U_{ij}) \qquad \forall\ \mathrm{m} \square \mathrm{j}$$

$$\Pr(\mathrm{m}) = \Pr\ ((X_i\beta_m + \gamma Z_{im} + \varepsilon_{im}) > (X_i\beta_j + \gamma Z_{ij} + \varepsilon_{ij}))$$

$$\Pr(\mathrm{m}) = \Pr((\varepsilon_{ij} - \varepsilon_{im}) < ((X_i\beta_m + \gamma Z_{im}) - (X_i\beta_j + \gamma Z_{ij})))$$

- To identify the scale of utility, we’ll set one U_i to 0 and take differences in each choice’s utility relative to the baseline.
- Take a four outcome model as an example

$$U_{i2} - U_{i1} = 0$$
$$U_{i2} - U_{i1} = (X_{i2} - X_{i1})\beta + (\varepsilon_{i2} - \varepsilon_{i1})$$
$$U_{i3} - U_{i1} = (X_{i3} - X_{i1})\beta + (\varepsilon_{i3} - \varepsilon_{i1})$$
$$U_{i4} - U_{i1} = (X_{i4} - X_{i1})\beta + (\varepsilon_{i4} - \varepsilon_{i1})$$

- Then define $\varepsilon_{im}{}^{*} = \varepsilon_{im} - \varepsilon_{i1}$ and $U_{im}{}^{*} = U_{im} - U_{1}$

$$U_{i1}^{*} = 0$$
$$U_{i2}^{*} = (X_{i2} - X_{i1})\beta + \varepsilon_{i2}^{*}$$
$$U_{i3}^{*} = (X_{i3} - X_{i1})\beta + \varepsilon_{i3}^{*}$$
$$U_{i4}^{*} = (X_{i4} - X_{i1})\beta + \varepsilon_{i4}^{*}$$

- And allow those errors to freely correlate in the estimation

- In principle we have a 4x4 covariance matrix of the errors
- But setting the utility of the baseline category to 0 eliminated the covariance between that category's error and the other 3, so we're left with 6 covariances
- Need to set one variance for identification: Stata sets 1 to the value of 2 (following the standard normal distribution, where the variance of a difference would equal the sum of the two variances)
- So we have 5 free error covariances for a 4 outcome model; for a 3 outcome model we have 2 covariances to estimate after setting baseline and identification of one remaining variance
- If all off-diaganol covariances=0, reduces to MNL
- Maximizing the (log-)likelihood is very difficult and sometimes empirically underidentified

```
Multinomial probit choice model                  Number of obs      =      4,659
Case ID variable: id                             Number of cases    =      1,553

Alternatives variable: alt                       Alts per case: min =          3
                                                                avg =        3.0
                                                                max =          3

Integration sequence:      Hammersley
Integration points:               718               Wald chi2(11)   =     456.45
Log simulated-likelihood = -1076.7077               Prob > chi2     =     0.0000
```

choice	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
alt						
ideodist	.1085986	.0076027	14.28	0.000	.0936977	.1234996
1						
age	-.0069987	.0040554	-1.73	0.084	-.0149471	.0009497
male	-.1046073	.1400967	-0.75	0.455	-.3791917	.1699771
dem	-2.776547	.1870855	-14.84	0.000	-3.143228	-2.409866
indep	-1.376596	.1772305	-7.77	0.000	-1.723962	-1.029231
distrust	.020345	.0795168	0.26	0.798	-.135505	.176195
_cons	1.443199	.3854256	3.74	0.000	.6877784	2.198619
2	(base alternative)					
3						
age	-.0347949	.0089882	-3.87	0.000	-.0524116	-.0171783
male	.7489599	.2735987	2.74	0.006	.2127162	1.285204
dem	-2.450167	.5111575	-4.79	0.000	-3.452017	-1.448316
indep	-.3021459	.3644408	-0.83	0.407	-1.016437	.4121451
distrust	.5810333	.1786166	3.25	0.001	.2309511	.9311154
_cons	-2.260638	1.015619	-2.23	0.026	-4.251213	-.2700618
/lnl2_2	1.046452	.1876424	5.58	0.000	.6786792	1.414224
/l2_1	1.092274	.7290712	1.50	0.134	-.3366795	2.521227

(alt=2 is the alternative normalizing location)
(alt=1 is the alternative normalizing scale)

```
. estat correlation
```

	1	3
1	1.0000	
3	0.3581	1.0000

```
Note: Correlations are for alternatives differenced with 2.

.
```

Differenced error correlations

Controlling for ideological distance and individual-specific factors (partisanship, age, trust, sex), utilities for Bush and Perot, relative to Clinton, is positively correlated

So Bush-Perot are "nearer" choices