

Unordered Multichotomous Data

Nathaniel Beck

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

QII - Week 5

1

Random Utility

- Before moving to a multichotomous situation, let us do one more dichotomous choice situation
- For individual i

$$U_0 = x_0\beta + \epsilon_0 \quad (1)$$

$$U_1 = x_1\beta + \epsilon_1 \quad (2)$$

- This is called a Random Utility Model
- since utility is the sum of a systematic component ($x\beta$) and a random component.
- Note that the covariate varies between the choices, but the β is the same.
- Thus the x represents attributes of the choices and the β are the weights put on those choices.
- In all that follows I will suppress the i so everything is for a generic case
- An individual chooses choice 1 iff $U_1 > U_0$, that is,

$$(x_1 - x_0)\beta > \epsilon_0 - \epsilon_1. \quad (3)_2$$

- If both random terms are normal, so is their difference, leading to another probit model.
- If the random terms are INDEPENDENT Gumbels (type I extreme value - note that the log of a Weibull is a Gumbel),
- their difference is logistic and the cumulative of a logistic is the logit,
- leading back to the logit model.
- That is, $\epsilon_1 - \epsilon_0$ is logistic

3

Unconditional Multiple Logit

- Now suppose y takes on values $0, 1, \dots, k$ (so $k + 1$ choices).
- Note that y is JUST A NOMINAL VARIABLE, so we are dealing with UNORDERED MODEL.
- π_j be the probability of choosing choice j (for a generic unit)
- As before we need to parametrize π_j as a function of independent variables, x .
- As before, only constraint is that $0 \leq \pi_j \leq 1$.
- Can get unconditional logit by taking

$$\pi_j = \frac{e^{x\beta_j}}{\sum_{l=0}^k e^{x\beta_l}} \quad (4)$$

4

- We must normalize by assuming $\beta_0 = 0$.
- Thus the β_l measure how much the “weights” on the different sociological variables differs from the weight on choice 0.
- Why “sociological?”
- Note that the \mathbf{x} 's are measurements of the individual but do not vary by choice (that is, they are sociological)
- but the β 's are choice specific.
- This is the opposite of what we had in the random utility model.
- Thus the β_j give the weighting of the sociological variables involved in the choice of j (relative to the reference category)
- We can arbitrarily choose a different base category and get estimates that are related to the original estimates in the obvious way.
- (Note different programs use different categories (either 0 or highest value) as the base. Life is hard. Pay attention!)

5

- The way to think about this is that the logic is the same as breaking a $k + 1$ category variable into k dummies and the interpretation of those in multiple regression.
- The β 's give the log odds of choice, in that

$$\ln \left[\frac{\pi_j}{\pi_0} \right] = \mathbf{x}\beta_j \quad (5)$$

$$\ln \left[\frac{\pi_j}{\pi_k} \right] = \mathbf{x}(\beta_j - \beta_k) \quad (6)$$

- Thus the odds of someone voting one choice relative to another is a function of attributes of that someone, including demographics and attitudes,
- but not a function of how the individual perceives the two parties.
- (Thus a left-right scale is fine, but not how far the individual is from each party in terms of the left-right scale.)

6

Conditional (MNL) model

- Return to the random utility model, and let x_j refer to individual i 's perception of choice j (on one or a vector of attributes)
- Thus, for individual i (again, suppress the superscript)

$$U_j = (\mathbf{x}_j)\boldsymbol{\beta} + \epsilon_j \quad (7)$$

- Note that now the independent variables are attributes of the choice, and the $\boldsymbol{\beta}$ are the same for all choices.
- Choice j is chosen if it provides more utility than any other choice, so

$$\pi_j = P(U_j > U_k, \forall k \neq j). \quad (8)$$

- If ϵ is Gumbel, the max of a bunch of INDEPENDENT Gumbels is Gumbel
- and so we end up with the conditional (multinomial) logit

$$\pi_j = \frac{(e^{x_j})^\beta}{\sum_{l=0}^c (e^{x_l})^\beta} \quad (9)$$

7

- Note that if there is some sociological x (the same over all x_j) it would just cancel out of Equation 9.
- We can handle this with the dummy variable type of trick, that is make $c-1$ sets of variables, $(x_i, 0, \dots, 0)$, the second $(0, x_i, \dots, 0), \dots, (0, \dots, 0, x_i, 0)$
- This also holds for the constant term.
- If we have alternative specific constant terms the claim is that $U_j =$ something plus a specific constant, so there is something that makes j special that we do not understand.
- We cannot estimate an overall constant since it would just cancel out of the utility functions.
- Note that mathematically unconditional and conditional logit are identical, but the interpretation is different.

8

INDEPENDENCE OF IRRELEVANT ALTERNATIVES - IIA

- If you use mnl, you note that the odds of choosing A or B are a function of the attributes of choices A and B but do not depend on what other choices are available.
- This property is called Independence of Irrelevant Alternatives (IIA) though it really should be called Independence of Relevant Alternatives.
- It is similar to the Axiom of Revealed Preference in micro-theory which says that the preference between A and B should not depend on what other choices are available.
- One way to think about is (this is the political equivalent of the famous “red bus—blue bus” example):

9

- Say your choices are Not Voting, Voting for Candidate A or Voting for Candidate B (N,A,B). You choose N if

$$U(N) > \max\{U(A), U(B)\} \quad (10)$$

- Now suppose we add another candidate, C. Suppose C is identical to B.
- Then this should not affect P(N) or P(A) but should split the B choosers identically between B and C.
- IIA says that $\frac{P(N)}{P(B)}$ is independent of whether or not C is on the ballot.
- This makes no sense, since without C all the BC voters would vote B while with C half these voters will vote B, half vote C, changing the relative probability of N or B.
- (IIA assumes that if C were initially on the ballot but withdrew, the C voters would split between N, A and B, whereas in reality they would all go to B.)

- The mathematics that makes mnl work is that the errors in the random utilities are independent (else the max of the Gumbel errors would not be Gumbel and all hell would break loose).
- Another way to think of this is that the (incorrect) assumption of independent random error terms causes us to overestimate $\max\{U(B), U(C)\}$.
- To see this, suppose that B and C are identical to the analyst, so that $x_B = x_C$. An individual chooses A over B and C if

$$U(A) > \max[U(B), U(C)] \quad (11)$$

$$= \max[\mathbf{x}_B\boldsymbol{\beta} + \epsilon_B, \mathbf{x}_C\boldsymbol{\beta} + \epsilon_C] \quad (12)$$

$$= \mathbf{x}_B\boldsymbol{\beta} + \max[\epsilon_B, \epsilon_C] \quad (13)$$

- If we assume that ϵ_B and ϵ_C are independent Gumbels,
- the assumption we make to allow us to use mnl, then their maximum is also Gumbel and, in expectation, larger than the expectation of either error alone.

11

- But if B and C are identical, the errors in the utility function should also be similar, and so the mnl assumptions overstates the maximum of ϵ_B and ϵ_C , causing us to overestimate $\pi_B + \pi_C$.
- One solution, of course, is to just collapse B and C into one single choice (as we might do with some minor parties in a large multi-party system).
- But how do we know which parties to collapse. What if the choices are not identical but only, as in reality, similar?
- Another way to think of IIA is that the estimate of β (the weighting function) should be invariant to which choices are available.
- Thus we can see IIA as an estimation assumption, which is testable, and if false has the usual types of consequences, rather than as an abstract property of decision makers.
- Note that even if in the abstract IIA is correct, if we omit a variable that is common to two choices, we will appear to have violations IIA.
- The common omitted variable is in the error and is a problem in many analyses term. This is why candidates B and C appear identical to the analyst.

12

- Thus IIA is a property of a specific set of variables and choices, not a properties of abstract human beings.
- As always with assumptions they buy something and cost something.
- There is no reason to believe that slight violations of IIA have serious consequences.
- We can test for IIA and we can design estimation strategies to get around it.
- The problem also arises with unconditional logit, but is much less severe since we are already estimating choice specific β s.
- These will be inaccurately estimated if IIA is false (since it is using some of the information about other choices), but it is likely that the problems will not be immense.
- We do not know that this will always be the case so you have to worry.

13

Testing for IIA

- McFadden has designed a test for IIA.
- It is an example of what is known as a Hausman test. Hausman's idea is as follows
- . Suppose we have two estimators which are both consistent under a null hypothesis, but one is inconsistent under the alternate.
- Suppose that estimator is also more efficient under the null (else why would we even consider it).
- Hausman then showed that the difference between the two estimators is χ^2 , (at least when multiplied by the appropriate variance-covariance matrix).
- If the difference is large in a χ^2 table then we are in the position where the second estimate is likely to be inconsistent, that is, the alternative is correct.
- If the difference is small then the null is likely to be correct.

14

- So let $\hat{\beta}_r$ indicate parameter estimates if “r” restrict to only use data on those who chose 1 or 2 (throwing out all individuals who chose 3) and let $\hat{\beta}_f$ indicate the parameter estimates based on the “f” full data set. Let \hat{V}_r and \hat{V}_f represent the corresponding variance covariance matrices, then

$$(\hat{\beta}_r - \hat{\beta}_f)[\hat{V}_r - \hat{V}_f]^{-1}(\hat{\beta}_r - \hat{\beta}_f) \sim \chi_k^2 \quad (14)$$

where k is the number of elements in the β vector.

- Note: If the possibly inconsistent “f” estimate were not more efficient than the “r” estimate, then under the null nothing would guarantee that the difference in the variances is invertible.
- Make sure to clean up any choice specific alternatives so that you are not estimating an unidentified restricted model. (No problem in the conditional model.)

15

Solutions

- One solution is Multinomial (unordered) Probit
- another, which is less general but easier to estimate, is Nested Multinomial Logit.
- begin with the latter.

16

- Suppose we think of our choice problem as having two levels (e.g. vote-not vote, if vote, vote for A or B, or vote Party A or Party B, if Party A vote candidate w or x, Party B vote for candidate y or z).
- Note that the choice of party is implicit in the choice over all lower level attributes (that is the person will vote for Party A if she chooses w or x).
- Consider the case where some attributes of candidates don't vary over parties, but some do. The decision between candidates of the same party is not affected by attributes which are common to the party.
- For notation, suppose we have two levels of choice, and attributes of choices (for a given individual fixed who is fixed and omitted from the notation) are denoted x_{jk} for those which vary at both levels and z_j for those which only vary at the top level.

17

- x_{Aw} might be how close you are to candidate w in Party A and z_A might be whether you think that Party A will generally do a good job.
- If you think of this as a choice of both what town to live in and what house to buy in the town, then the x refers to attributes of houses and the z to attributes of towns.
- So far this is straightforward conditional log, so

$$P_{jk} = \frac{e^{x_{jk}\beta + z_j\gamma}}{\sum_{j^*} \sum_{k^*} e^{x_{j^*k^*}\beta + z_{j^*}\gamma}} \quad (15)$$

- This is straightforward, except that we are estimating over lots of choices, and γ only depends on top level choices.

18

- We can make the problem easier by noting that some things (the z) do not vary over the lower levels.

$$P_{jk} = P_{k|j} P_j \quad (16)$$

$$P_{k|j} = \frac{e^{x_{jk}\beta}}{\sum_{j^*} e^{x_{jk^*}\beta}} \quad (17)$$

$$P_j = \sum_{k^*} P_{jk^*} \quad (18)$$

$$= \frac{e^{z_j\gamma + I_j}}{\sum_{j^*} e^{z_{j^*}\gamma + I_{j^*}}} \quad (19)$$

$$I_j = \ln\left(\sum_{k^*} e^{x_{jk^*}\beta}\right) \quad (20)$$

I_j is called the inclusive value, and is a measure of the systematic component of the maximum utility of all the choices that branch from j .

19

- Note the trick here.

$$e^{z_k\gamma + I_j} = e^{I_j} e^{z_k\gamma} \quad (21)$$

but

$$e^{I_j} = e^{\ln(\sum_{k^*} e^{x_{jk^*}\beta})} \quad (22)$$

$$= \sum_{k^*} e^{x_{jk^*}\beta} \quad (23)$$

which is what makes this look like a standard MNL model.

- (Remember that e to the log of something is just that something.)
- NOTHING IS DIFFERENT ABOUT THIS THAN ORDINARY MNL, EXCEPT IT IS COMPUTATIONALLY SIMPLER AND IT GIVES US THE NOTION OF INCLUSIVE VALUE.
- In applied work there may be thousands of choice objects, and so this saving is important.
- In the social sciences even ten choices would set a record, so computations are not the issue.
- What is good about the above formulation is that it paves the way for Nested Multinomial Logit.

20

- If the choices in the lower level branches are similar (within each branch), then the inclusive value will overestimate the utility of the sub-branch.
- McFadden has fixed this by assuming that the random components of the utility come from a General Extreme Value distribution instead of an Extreme Value (Gumbel) distribution.
- This GEV has a second parameter, σ which measures the similarities of the choices.
- This leads to Nested Multinomial Logit (NMNL)
- While the statistics are quite complex, the idea is easy enough (though it did help to win Dan McFadden the Nobel Prize).
- The basic difference is that the equation for the top (second) level is

$$P_j = \frac{e^{z_j\gamma+(1-\sigma)I_j}}{\sum_{k^*} e^{z_{k^*}\gamma+(1-\sigma)I_{k^*}}} \quad (24)$$

where σ measures the similarity between choices in the lower level.

21

- If $\sigma = 0$ then the errors are independent (IIA holds).
- Thus σ just deflates the inclusive value to account for correlation (similarity) of the error terms.
- Note that $0 \leq \sigma \leq 1$. If you get estimates outside that range this indicates the model is misspecified or there is some other problem. Make sure to check this.
- This is easy to generalize to three or more levels of choice. It is also possible to have different choices in different branches, and different branches going down different numbers of levels.
- Note that it makes a difference how you order the choices and branching, since the assumption is that the random components of choices in different sub-branches are uncorrelated.
- That is, make sure to remember that you have to think about whether the stochastic portion of the random utilities are correlated. NMNL allows for correlated stochastic portions (“errors”) in the same branch, but not across branches.

22

- IIA is introduced by the assumption that the random components are independent.
- Suppose they are dependent in a manner where MNML doesn't work, that is, you have no idea what errors are correlated or uncorrelated with what other errors, or perhaps, as is likely, all of them may be correlated.
- We assume that the errors are drawn from a multivariate normal distribution with zero mean but arbitrary correlation matrix (which is to be estimated).

$$\pi_j = P \left[(x^j - x^k)\beta > \epsilon_k - \epsilon_j \forall k \neq j \right] \quad (25)$$

- But the difference of two normals is normal so this probability is an n-1 fold multiple integral of the multivariate normal (with 0 mean and variance covariance matrix Σ to be estimated)

$$\pi_j = \int_{(x^c - x^j)\beta}^{\infty} \cdots \int_{(x^1 - x^j)\beta}^{\infty} \phi_{c-1}(z_1, z_2, \dots, z_c) dz_1 dz_2 \dots dz_c \quad (26)$$

23

- This is conceptually straightforward but numerically difficult
- Direct evaluation of the n-1 fold integrals is impossible for reasonable sized n (say greater than 4).
- Relatively new work, by McFadden and others, estimates these integrals by simulation. It appears that we can now estimate choice models with as many as 10 choices by multinomial (unordered) probit.
- Alvarez and Nagler estimate a three choice problem (Clinton, Bush, Perot) using direct evaluation of the integral. Question to ask is whether there appears to be enough dependence in the errors to make this stuff interesting, or whether we can give some political interpretation to this dependence.

24

- It is suggested (Greene, Glasgow in Political Analysis) that the random parameter logit is a better alternative than mprobit - easier to estimate, more flexible.
- Most apps right now to transport, only Glasgow has used in ps, so hard to know.
- Basic idea is pretty simple. Just take the MNL model, but make the coefficients random, that is:

$$\beta_{jk} = \beta_k + \mathbf{z}_i \theta_k + \sigma_k \mu_{ik} \quad (27)$$

where μ is normal and the z move the mean β around between individuals in a deterministic way.

- This idea, do mostly to Train, allows one to estimate a variety of models and substitution patterns. For those with such data (conditional mnl), this may be something worth looking into.
- Since the notation gets fierce, and the data demands are very high
- I leave this topic to those who really need it