

# Limited Dependent Variables

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

January 2004

## Binary Dependent Variable

(Note: Since we do not care much about whether we have one or more IV's, I have not tried to bold vectors here, but note that all the  $x_i$  and  $\beta$  are really vectors, though nothing depends on this. Where it matters that we have a vector of IV's, boldface denotes this.)

Suppose we have  $N$  individuals for whom we observe  $y_i = 0$  or  $1$ . (Can code in any other way, these are just two states of the world. This coding is the most convenient.) Each observation has one or more covariates, denote  $x_i$ . (Thus will either have a vector of parameters  $\vec{\beta}$  or a single  $\beta$ .)

The likelihood of the sample is just

$$L = P(y_1)P(y_2) \dots P(y_N). \quad (1)$$

We now need to put this in terms of the covariates and parameters.  $P$  is a probability. Thus any function which returns values between 0 and 1 would be plausible. We also think that, in general, for a single covariate, as  $x$  increases  $P$  should increase (or decrease), that is, there is a monotonic relationship between  $x$  and  $P$ . While this is probably reasonable in general, it clearly isn't always true. Can deal with this later.

Given this, and cumulative distribution function will do. One that was commonly used is the normal, which leads to probit. We will do logit, which is hard to tell from probit.

## Logit

Suppose

$$P(y_i = 1|x_i) = \frac{1}{1 + e^{-x_i\beta}} = \Lambda(x_i\beta). \quad (2)$$

(The  $\Lambda$  just simplifies my typing!) We will usually denote the probability that observation  $i$  is one by  $\pi_i$ .

Note that this ties the observed outcomes to the covariates and the model parameters. Note that the logit function works well, in that when  $x$  is very negative  $P \rightarrow 0$ , when  $x = 0$  we get  $P = .5$  and when  $x$  is large and positive,  $P \rightarrow 1$  (all assuming  $\beta > 0$ , reverse if  $\beta < 0$ ).

Work with  $x_i\beta$  for a vector of covariates. In this case the interpretation is on the linear form  $\beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_k x_{ik}$ . It is the interpretation of this that is interesting. Note: if you don't like the monotonic assumption, you can use squares and such, or multiplicative interactions, or anything else you would use in a regression. We will have a neater way of handling this on Friday.

The easiest (not the most common) way to write the likelihood is to put all the zero observations first, followed by all the one observations. Also remember that for this case  $P(0) = 1 - P(1)$ .

The likelihood is then

$$L = (1 - \Lambda(\beta x_1))(1 - \Lambda(\beta x_2)) \dots (1 - \Lambda(\beta x_{\text{last zero}})) \Lambda(\beta x_{\text{first one}}) \Lambda(\beta x_{\text{next one}}) \dots \Lambda(\beta x_{\text{last one}}). \quad (3)$$

This can then be maximized. We get coefficients and standard errors and the log likelihood for testing.

Simple calculus gives that the first order conditions for the loglike to be maximized as

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^n (y_i - \Lambda_i) \mathbf{x}_i = \mathbf{0} \quad (4)$$

Note this means that for the constant term ( $\mathbf{x}_i = 1$ ) we have

$$\sum_{i=1}^n (y_i - \Lambda_i) = 0 \quad (5)$$

so that the average predicted probability (that  $y_i = 1$ ) is just the proportion of  $y$ 's that are one.

Thus, if we have rare events (say war), then the constant term will just make the average  $\pi$  small; if we drop a lot of rare events (say we go to Politically Relevant Dyads) we just increase the value of the constant term.

The Hessian also takes a very simple form

$$\mathbf{H} = - \sum_{i=1}^n \Lambda_i(1 - \Lambda_i)\mathbf{x}_i\mathbf{x}_i \quad (6)$$

Note that  $\Lambda_i$  and  $1 - \Lambda_i$  are always between 0 and 1 (strictly), and  $\mathbf{x}_i\mathbf{x}_i$  is the  $k \times k$  matrix of the squares and cross-products of the  $k$  independent variables. This is positive definite, so the sum (each multiplied by a positive number) is pd, so the negation is negative definite everywhere (for any admissible value of  $\Lambda_i$  so the log likelihood function is globally concave which is why maximization is so easy).

You should note that in the likelihood, we only have  $\beta$ 's, there is no variance term. Since it is easier to understand this in terms of normals, just remember that what is true of probit and variance term is also true of logit.

## Probit

It should be noted that the likelihood for any binary dependent variable has the same format, that is

$$P(y_1 = 0)P(y_2 = 0) \dots P(y_{\text{last zero}} = 0)P(y_{\text{first one}} = 0) \dots P(y_{\text{last one}} = 1) \quad (7)$$

with appropriate replacement for the probabilities.

Any admissible model for the probabilities (that is,  $p_i = f(\mathbf{x}_i, \beta)$ ) that returns values strictly in the unit interval, is a perfectly reasonable model.

Thus, for probit, we would use  $P(y = 1) = \Phi(x_i\beta)$  where  $\Phi$  is the standard cumulative normal distribution. OTHER THAN THAT THE SETUP IS IDENTICAL. Empirically the two setups are very similar, and so we can't really tell which better fits the data (and have no tests to do so). Logit is generally used these days because it is numerically simpler.

Note that the coefficients of a probit will be different than those of the corresponding logit, since the transformation from probit to probability is different than the transformation from logit to probability. For historical reasons some things (e.g. ordered probit) are always done in a probit context, while other things (e.g. multi-choice unordered logit) are always done in a logit context. There is no reason why we can't substitute logit or probit in these contexts. In general, whatever we say about logit holds for probit, *mutatis mutandis*.

Note that logit and probit, while giving different values for the  $\beta$ , cannot really give very different values for the  $\hat{\pi}_i$ . Why??

Because ml basically is choosing the parms so that for the observations where  $y_i = 1$  we get  $\pi_i$  as close to one as we can, and for  $y_i = 0$  we get  $\pi_i$  as close to zero as we can. Of course we cannot hit this perfectly, but any reasonable model should generate similar  $\hat{\pi}$ 's.

Note also: Since logit and probit are not nested (define!), there is no nice lr or Wald test that will discriminate between them.

Note that we are assuming that the  $\pi$ 's are generated by a standard normal CDF (with varying mean, but  $\sigma^2 = 1$ ). We cannot estimate a separate variance term. This may be clearer next week when we go to latent variable notation. But for now, think of the following.

If the variance of the normal were not one, but something else (and it would by assumption be the same something else for everyone), we could accomplish the same thing by multiplying all the  $\beta$ 's by some number, which would have the same effect on the variance. So we normalize by allowing the  $\beta$ 's to be free, but the variance to be one.

We can think of probit (and logit) as a defective regression, where we only know the sign of the dep var. So while we use (0,1) for the two values of  $y$ , it could equally well be (-10,23).

In regression, the values of  $y$  are meaningful, and it is these that tie down the variance term.

To see this more clearly, we need to think in terms of latents.

## Latent variable interpretation

Suppose that there is some unobserved ("latent")

$$y_i^* = \mathbf{x}_i\beta + \epsilon_i, \quad \epsilon \sim N(0, 1). \quad (8)$$

Suppose that

$$y_i = 1 \text{ if } y_i^* > 0 \quad (9)$$

$$y_i = 0 \text{ otherwise} \quad (10)$$

$$\pi_i = P(y_i = 1) \quad (11)$$

$$\pi_i = P(y + i^* > 0) \quad (12)$$

$$= P(x_i\beta > -\epsilon) \quad (13)$$

$$= 1 - \Phi(-x_i\beta) \quad (14)$$

which gives a probit model. Can do similar for logit, though notation is not quite so pretty.

Note this shows that logit/probit are just latent variable models, where we have a linear model for the latent and then a measurement model which ties the latent to observables.

This means that probit/logit is just regression with less information, that is, we know all about the covariates but only the sign of the dependent variable.

(Why is threshold fixed at 0 - because there is a constant term in  $x$ .) (Why is variance of epsilon fixed - because can arbitrarily change the variance of  $y^*$  by multiplying  $\beta$  by a constant.)

## Interpretation

We like linear regression because it is easy to compute the impact of a change in an independent variable. I often hear the coefficients of logit or probit are uninterpretable. This is nonsense. They are just a bit harder to interpret.

For discrete covariates (say dummy variables), all you need to produce is  $P(y = 1 | \text{dummy} = 0, \text{other covariates})$  and  $P(y = 1 | \text{dummy} = 1, \text{other covariates})$ . Only question is what values to set those other covariates to. Either choose interesting combinations or set at the mean.

For continuous covariates, want  $\frac{\partial P(y_i=1)}{\partial x_k}$ . For a linear model,  $\frac{\partial y_i}{\partial x_k} = \beta_k$ . The lovely thing about the logit is that

$$\frac{\partial P(y_i = 1)}{\partial x_k} = \beta_k \Lambda(x_i \beta) (1 - \Lambda(x_i \beta)) \tag{15}$$

$$= \beta_k P(y_i = 1) P(y_i = 0) \tag{16}$$

More generally, for any CDF we use to define the binary choice model, say  $F$ , we would have by the definition of a cdf,

$$\frac{\partial P(y_i = 1)}{\partial x_k} = F'(\mathbf{x}_i \beta) \beta_k \tag{17}$$

## Using CLARIFY for interpretation

Easy enough to compute  $P(Y = 1)$  by hand for interesting values.

But can get more (confidence intervals) by simulation.

The estimate,  $\hat{\beta}$  is uncertain, though from ML it is  $AN(\beta, V)$  where  $V$  is the estimated VCV matrix (the inverse of the negative expected value of the Hessian).

So we can repeatedly draw different values of  $b$  from this distribution, use them to calculate  $P(y = 1)$  and then look at the mean to see what this is on average, but we can also look at the distribution (top and bottom 2.5%) to get a confidence interval.

This can be extended to any “QOI” (“quantity of interest.”) Refer to King, Tomz and Wittenberg, AJPS, 2000.

In Stata, commands are

```
estsimp logit returned checks age surplus marginal  
setx checks median age median surplus median marginal 1  
simqi, fd(pr) changex(checks median p75)
```

See help in stata to get all the subcommands and options

## Scobit

Nagler has proposed that logit (and probit) may be too inflexible in that they assume that variables have their maximal impact when  $P=.5$ . (Just look at where the maximum of Equation 15 is.) Nagler proposes that the world may look logit except that the maximal impact of variables may occur at any value of  $P$ .

He proposes the “Scobit” model. Remember that we can define  $\pi_i$  by any probability distribution function. Nagler proposes the “Burr-10”, which has an additional parameter  $\alpha$ , so

$$\pi_i = (1 + e^{x_i\beta})^{-\alpha} \quad (18)$$

Note that  $\alpha = 1$  yields the logit model, so the logit is nested inside the scobit.

The likelihood for scobit is then given in the usual manner, substituting this new function for  $\pi_i$ .

$$\ln L = \sum_{i=1}^N (1 - y_i) \ln[F(-X_i\beta)] + y_i \ln[1 - F(-X_i\beta)] \quad (19)$$

where  $F(-X_i\beta) = (1 + e^{x_i\beta})^{-\alpha}$ .

We can then test the null hypothesis that  $\alpha = 1$  in the usual manner. If we do not reject  $\alpha = 1$  we can then use logit, which is easier to interpret. If we stay with scobit we need to carefully interpret results.

In practice I find that the standard error on  $\alpha$  is large and with reasonable sized samples it is hard to find  $\alpha$  significantly different from 1. But even if scobit is not helpful, it does sensitize you to the assumption of logit/probit that the maximal impact is for people with  $P = .5$ . This is an assumption, not an empirical conclusion! Note Nagler’s discussion of Wolfinger and Rosenstone. The stata command is “scobit” Note that the scobit only works if the  $P$  where the impact is maximal is less than some value (.6x), one can get nonsense if in your data set that  $P$  corresponds to a value greater than this. However, we can of course freely reverse the 0’s and 1’s, and hence if the  $P$  where the impact is maximal is over .6x, in the reversed model it is under .4. Thus one wants to investigate for both codings.

## Heteroskedastic Logit/Probit

(Dubin and Zeng)

The logit model assumes that all individuals have the same variance (homoskedasticity). Can we deal with heteroskedasticity?

Dubin and Zeng propose a model which is like the standard conditional logit (Equation 57) except for the parameter,  $\theta_i$ , which measures heteroskedasticity.

$$\pi_i = \frac{1}{1 - e^{x_i\beta\theta_i}} \quad (20)$$

The effect of the  $\theta$  is to “spread out” the logistic curve for some people. For example, people with higher knowledge may be more able to discern their self interest (Gerber and Lupia), so that their probability of voting for something is high if in their interest, but low if not; for those with low knowledge, it takes a bigger movement in the iv to induce the same change in probabilities.

As with any model with heteroskedasticity, we cannot estimate a separate  $\theta$  for each individual. As usual we attempt to parameterize it.

Note that one can just as easily do heteroskedastic probit. In fact, Stata only does the the heteroskedastic probit form, using `hetprobit`.

Just go back to latent form

$$y_i^* = \mathbf{x}_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2). \quad (21)$$

where

$$\sigma_i^2 = e^{\mathbf{z}_i\gamma} \quad (22)$$

Why exponentiate - easiest way to ensure that  $\sigma^2$  stays positive.

Note: there is still no scale on  $\sigma^2$ , that is, no constant term in the equation, so assuming that, on average, variance of  $y^*$  is one for usual reason.

## Multichotomous DV

Categorize DV

- Ordered DV
- Example Approval of President or Party ID
- Count DV

- How many debates one watched
- How many wars a nation fights in
- DV takes on integer values, 0, 1, /dots, n, /dots
- Is there a top? Is zero at bottom; makes a difference
- Unordered DV
- Which of three parties did you vote for?
- Whether DV is ordered or not is a matter of theory - in coding, is 1 between 0 and 2 are is coding just nominal

## Ordered Probit

Suppose that the categories for  $y$  are ordered (apathetic, somewhat interested, very interested). While this looks similar to what we have done, the ordered and unordered models are VERY different. DO NOT CONFUSE ONE WITH THE OTHER.

Suppose we use a threshold model with latent dependent variable, so

$$y^* = x\beta + \epsilon, \epsilon \sim N(0, 1). \quad (23)$$

$$y = c \text{ if } y^* \geq \tau_{c-1} \quad (24)$$

$$y = c - 1 \text{ if } \tau_{c-1} > y^* \geq \tau_{c-2} \quad (25)$$

$$\vdots \quad (26)$$

$$y = 1 \text{ if } \tau_1 > y^* \geq 0 \quad (27)$$

$$y = 0 \text{ if } y^* < 0 \quad (28)$$

where the  $\tau$  are thresholds to be estimated.

The likelihood is

$$L = \prod_{i=1}^N (\pi_0^{y_i=0} \dots (\pi_c^{y_i=c}) \quad (29)$$

where the exponent is 0 or 1 depending on the value of  $y$ . For individual  $i$ ,

$$\pi_0 = \Phi(-x\beta) \quad (30)$$

$$\pi_1 = \Phi(\tau_1 - x\beta) - \Phi(-x\beta) \quad (31)$$

$$\vdots \quad (32)$$

$$\pi_c = 1 - \Phi(\tau_{c-1} - x\beta) \quad (33)$$

The first threshold is zero and the normal has variance 1 for the same reason as in the probit. The estimated thresholds must have the appropriate order.

## Interpretation

For ordered probit or logit you want to compute the probability of being in the various categories and then compare those probabilities for interesting combinations of the independent variables.

You also need to look at the thresholds. Are adjacent pairs significantly different from each other? Do they seem to imply a linear pattern or are they bunched in some interesting way? Do not treat the thresholds as a nuisance - they are an interesting part of the model.

Note: you can also do ordered logit, but because notation is simpler, almost everyone does ordered probit (is this a good reason? fortunately does not matter)

## Event Counts

Suppose  $y_i$  counts the number of times that  $i$  does something

- Number of nations joining a sanction

So  $y$  takes on values  $0, 1, \dots$  (may in practice be finite, but should be large). If only takes on values 0 and 1, use binary models.

## Poisson

Simplest model is that  $y_i$  is Poisson with mean  $\lambda_i$ . This follows from the assumptions:

- Events occur over time

- P(event) is constant over time
- P(event) is independent of how many prior events occurred
- P(2 events in small unit of time) = 0

Under these conditions, the number of events that we observe for a unit follows a Poisson distribution characterized by the single parameter density

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}. \quad (34)$$

Note that the variance of  $y_i$  is also  $\lambda$ , so the model is restrictive (as is any one parameter model).

As always, need to parameterize  $\lambda_i$  as  $f(x_i, \beta)$ .

$\lambda$  must be positive.

Usually assume

$$\lambda_i = e^{x_i \beta} \quad (35)$$

Why? Look at

$$\frac{\partial \lambda_i}{\partial x_j} = \lambda_i \beta_j. \quad (36)$$

Why is this reasonable? Harder to move from 0 to 1 than from 100 to 101.

Likelihood is then

$$L = \prod_{i=1}^N \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad (37)$$

$$= \prod_{i=1}^N \frac{e^{-e^{x_i \beta}} (e^{x_i \beta})^{y_i}}{y_i!} \quad (38)$$

or the log likelihood is

$$\ln L = \sum_{i=1}^N [-\lambda_i + y_i x_i \beta - \ln y_i!] \quad (39)$$

$$= \sum_{i=1}^N [-e^{x_i \beta} + y_i x_i \beta - \ln y_i!], \quad (40)$$

This is a very well behaved likelihood which is easy to maximize.

## Negative Binomial

The assumption that the variance equals the mean is very strong. Thus the Poisson may be too restrictive. In practice the variance is usually larger than the assumed by the Poisson. Such a model is called over dispersed.

Let the mean be as before

$$\lambda_i = e^{x_i\beta} \quad (41)$$

but assume that the variance is

$$\sigma^2 = \lambda_i(1 + \alpha\lambda_i) \quad (42)$$

One model that can handle this and still allow  $y$  to take on only discrete values (DATA ADMISSIBLE) is the NEGATIVE BINOMIAL. Thus over-dispersion is

$$\frac{\text{Var}(y_i)}{E(y_i)} = 1 + \alpha E(y_i), \quad \alpha > 0. \quad (43)$$

## Other parameterizations

Some people (e.g. King) assume that the over-dispersion (43) is  $1 + \alpha$  which leads to a slightly different model. (Would you expect the degree of over dispersion to increase as the mean increases?)

If you go to a probability text, you will find the usual parameterization of the negative binomial:

$$f(n; r, p) = \binom{n-1}{r-1} p^r (1-p)^{n-r} \quad (44)$$

where the mean is  $\frac{r}{p}$  and the variance is  $\frac{r(1-p)}{p^2}$ .

The 'King' parameterization is

$$f(n; \lambda, \gamma) = \binom{n-1}{\lambda(2+\gamma)-1} \left(\frac{1}{2+\gamma}\right)^{\lambda(2+\gamma)} \left(\frac{1+\gamma}{2+\gamma}\right)^{n-\lambda(2+\gamma)} \quad (45)$$

This difference causes no problems so long as you realize what parameterization you are using! Note that this works for over dispersion. The Poisson is not a special case of the negative binomial, which is not defined for  $\alpha = 0$ . Of course you can decide to go with Poisson if  $\alpha$  is small or insignificant.

Don't have a good way of dealing with under-dispersion, but this appears to be rare. King has a "general event count" distribution which may work.

## Censoring

Sometime the highest event count is  $N$  or more (e.g. you might have a few cases with a huge number of events, but only recorded as  $N$  or more). Then the last entry in the likelihood is not  $P(y_i = N)$  but  $P(y_i \geq N)$ . It is easy to take this into account.

## Truncation at zero

Sometimes because of our design we only observe  $y_i \geq 1$ . This may be because of choice based sampling. That is, instead of doing a survey and asking people how many times they ride the bus, we get on a bus and ask people how many times they ride. The former is obviously the only way to get information on the decision of whether to ride at all, but the large number of no rides makes it hard to study how many times rides ride.

If we fit a Poisson (or Neg. Bin) we are falsely allowing probability of zero to be positive. The simple solution is to change each of the probability entries in the likelihood as follows (with  $P_{tr}$  being the truncated (correct) probability and  $P$  being the incorrect probability):

$$P_{tr}(y_i = n) = \frac{P(y_i = n)}{1 - P(y_i = 0)}. \quad (46)$$

Note the assumption we are making is that the process of going from zero to one is the same Poisson process we observe, we just happen not to have chosen to observe zeros. THIS IS A VERY STRONG ASSUMPTION.

## Zero Inflated Probability (ZIP)

Perhaps a better assumption is that the process which govern zero-not zero is different than the process which determines how many times something is done, given it is done. Obviously to estimate this model you need info on zeros and non-zeros, though presumably you could get from different samples.

Think of some binary random variable,  $z_i$  which measures if you would ever do something or not. (Some people get counts of zero because they didn't do something but might have, others would never +have even considered doing the thing.)

If  $z_i = 1$  then  $y_i^*$  measures how many times you do something.  $y_i$  is the observed event count. Then we change our probabilities in the likelihood function so (with  $P_P$  being the Poisson (or negative binomial),  $P_B$  being the probit and  $P_Z$  being the ZIP probabilities):

$$P_Z(y_i = 0) = P_B(z_i = 0) + P_B(z_i = 1)P_P(y_i^* = 0) \quad (47)$$

$$P_Z(y_i, y_i > 0) = P(z_i = 1)P_P(y_i) \quad (48)$$

Note that the probit/logit and Poisson/negative binomial parts may be estimated with the same or different covariates. Relationship to hurdle model

Hurdle - going from 0 to 1, measured by a probit

Then if we get over the hurdle, event count (with truncation)

Makes sense - first we ask if someone watches debates at all, then a second and different model for how many

### Aside: Bootstrapping

In the homework, you will see a small N example (26 or so cases). Remember that all se's in ml are asymptotic. So how off are they in small samples?

There are a variety of ways to deal with this, including MCMC. But one way is bootstrapping.

Idea is very simple. We have 26 obs. Sample 26 obs from these (obviously WITH replacement, since only one sample without!). Then compute the relevant statistic. Do this again and again (100 or 1000 times) and look at the 95% CI (just take the bottom and top 2.5% points, as in Clarify). This may well give a more accurate assessment of the variability of the parameter estimates. There are lots of complications and tweaks on this, but the basic move as described is all you need.

Stata command:

```
bootstrap "logit dv ivlist" _b _se, reps(100)
```

The stuff in quotes is the command you want to bootstrap, `_b` and `_se` tell stata to compute the bootstrap ci's for  $\beta$  and its se, `reps(100)` says use 100 bootstrapped samples (not a lot, but won't waste a lot of time).

Just interpret the simplest of the bootstrapped ci's. The (N) CI is just based on computing the se of 100 estimates and using a normal approximation, the (P) CI computes the bottom and top 2.5%tile. Either are fine, hard to tell which is better. Ignore the other output.

### Unordered Logit: Unconditional and Conditional

There are two types of extensions. One has all the covariates being attributes of individuals (unconditional), the other has covariates being attributes by choices as perceived by individuals (conditional). The first I often call sociological since the covariates are sociological types of covariates (demographics, attitudes). The second I often call economic, since the covariates are often things like the cost of a choice

to an individual. conditional approach fits well into RANDOM UTILITY MODELS.

Before doing this, let us get in the habit of looking at some stories. Let us start with a binary DV

## Random Utility

For individual  $i$

$$U_0 = x_0\beta + \epsilon_0 \quad (49)$$

$$U_1 = x_1\beta + \epsilon_1 \quad (50)$$

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  $\beta$  is the same. Thus the  $x$  represents attributes of the choices and the  $\beta$  are the weights put on those choices.

An individual chooses choice 1 iff  $U_1 > U_0$ , that is,

$$(x_1 - x_0)\beta > \epsilon_0 - \epsilon_1. \quad (51)$$

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 and we get the probability of choosing 1 as the integral up to where this difference is just equal to the difference in the systematic parts of the utility. The Gumbel is a very useful distribution in discrete choice and we shall see it often.

Lots of stories, same statistical model. But the different interpretations can be helpful, as we shall see.

## Unconditional Multiple Logit

Now suppose  $y_i$  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.

For individual  $i$ , let  $\pi_j^i$  be the probability of choosing choice  $j$ . For simplicity, we try to avoid the use of the  $i$  superscript, though in a few places it is useful to be explicit. Thus  $\pi_j$  refers to the probability of choosing choice  $j$  for some generic individual.

As before we need to parameterize  $\pi_j$  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 (for individual  $i$ )

$$\pi_j^i = \frac{e^{x_i \beta_j}}{\sum_{l=0}^k e^{x_i \beta_l}} \quad (52)$$

## Interpretation

We must normalize by assuming  $\beta_0 = 0$ . Thus the  $\beta_l$  measure how much the ‘‘weights’’ on the different sociological variables differs from the weight on choice 0.

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!) 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.

Note that the  $x_i$ 's are measurements of the individual but do not vary by choice (that is, they are sociological) but the  $\beta$ 's are choice specific. This is the opposite of what we had in the random utility model. Thus the  $\beta_j$  give the weighting of the sociological variables involved in the choice of  $j$  (relative to the reference category).

The  $\beta$ 's give the log odds of choice, in that

$$\ln \left[ \frac{\pi_j^i}{\pi_0^i} \right] = x_i \beta_j \quad (53)$$

$$\ln \left[ \frac{\pi_j^i}{\pi_k^i} \right] = x_i (\beta_j - \beta_k) \quad (54)$$

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.)

## Conditional (MNL) model

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

$$U_j = (x_j)\beta + \epsilon_j \quad (55)$$

(Note that now the independent variables are attributes of the choice, and the  $\beta$  are the same for all choices.)

Choice  $j$  is chosen if it provides more utility than any other choice, so

$$\pi_j = P(U_j > U_k, \forall k \neq j). \quad (56)$$

If  $\epsilon$  is Gumbel, the max of a bunch of INDEPENDENT Gumbels is Gumbel and so we end up with the conditional (multi-nomial) logit

$$\pi_j^i = \frac{(e^{x_j^i})^\beta}{\sum_{l=0}^c e^{(x_l^i)^\beta}} \quad (57)$$

Note that if there is some sociological  $x$  (the same over all  $x_j$ ) it would just cancel out of Equation 57. 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.

## 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):

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

$$U(N) > \max\{U(A), U(B)\} \tag{58}$$

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

$$U(A) > \max[U(B), U(C)] \tag{59}$$

$$= \max[x_B\beta + \epsilon_B, x_C\beta + \epsilon_C] \tag{60}$$

$$= x_B\beta + \max[\epsilon_B, \epsilon_C] \tag{61}$$

. If we assume that  $\epsilon_B$  and  $\epsilon_C$  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. 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  $\epsilon_B$  and  $\epsilon_C$ , 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  $\beta$  (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 term. This is why candidates B and C appear identical to the analyst.) This is probably a very common situation, so we might expect to see many violations. 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  $\beta$ 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.

## 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  $\chi^2$ , (at least when multiplied by the appropriate variance-covariance matrix). If the difference is large in a  $\chi^2$  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.

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

$$(\hat{\beta}_r - \hat{\beta}_f)[\hat{V}_r - \hat{V}_f]^{-1}(\hat{\beta}_r - \hat{\beta}_f) \sim \chi_k^2 \quad (62)$$

where  $k$  is the number of elements in the  $\beta$  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.)

## Solutions

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

### NESTED MULTINOMIAL LOGIT - NMNL

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  $x_{ij}$  for those which vary at both levels and  $z_i$  for those which only vary at the top level.

( $x_{Aw}$  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

$$P_{ij} = \frac{e^{x_{ij}\beta + z_i\gamma}}{\sum_{i^*} \sum_{j^*} e^{x_{i^*j^*}\beta + z_{i^*}\gamma}} \quad (63)$$

This is straightforward, except that we are estimating over lots of choices, and  $\gamma$  only depends on top level choices. We can make the problem easier by noting that some things (the  $z$ ) do not vary over the lower levels.

$$P_{ij} = P_{j|i} P_i \quad (64)$$

$$P_{j|i} = \frac{e^{x_{ij}\beta}}{\sum_{j^*} e^{x_{ij^*}\beta}} \quad (65)$$

$$P_i = \sum_{j^*} P_{ij^*} \quad (66)$$

$$= \frac{e^{z_i\gamma + I_i}}{\sum_{i^*} e^{z_{i^*}\gamma + I_{i^*}}} \quad (67)$$

$$I_i = \ln\left(\sum_{j^*} e^{x_{ij^*}\beta}\right) \quad (68)$$

$I_i$  is called the inclusive value, and is measure of the systematic component of the maximum utility of all the choices that branch from  $i$ .

Note the trick here.

$$e^{z_i\gamma + I_i} = e^{I_i} e^{z_i\gamma} \quad (69)$$

but

$$e^{I_i} = e^{\ln(\sum_{j^*} e^{x_{ij^*}\beta})} \quad (70)$$

$$= \sum_{j^*} e^{x_{ij^*}\beta} \quad (71)$$

which is what makes this look like a standard MNL model. (Remember that  $e$  to the log of something is just that something. You should have paid attention when your teacher taught you about logarithms!)

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.

## NMNL

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,  $\sigma$  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

$$P_i = \frac{e^{z_i\gamma + (1-\sigma)I_i}}{\sum_{i^*} e^{z_{i^*}\gamma + (1-\sigma)I_{i^*}}} \quad (72)$$

where  $\sigma$  measures the similarity between choices in the lower level. If  $\sigma = 0$  then the errors are independent (IIA holds). Thus  $\sigma$  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. MNML allows for correlated stochastic portions ('errors') in the same branch, but not across branches.

## Unordered probit

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).

$$\pi_j = P \left[ (x^j - x^k)\beta > \epsilon_k - \epsilon_j \forall k \neq j \right] \quad (73)$$

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  $\Sigma$  to be estimated

$$\pi_j = \int_{(x^c - x^j)\beta}^{\infty} \cdots \int_{(x^1 - x^j)\beta}^{\infty} \phi_{c-1}(z_1, z_2, \dots, z_c) dz_1 dz_2 \dots dz_c \quad (74)$$

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). 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.

### Random parameter logit as alternative

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. Also need to do in Limdep.

Basic idea is pretty simple. Just take the MNL model, but make the coefficients random, that is:

$$\beta_{ik} = \beta_k + \mathbf{z}_i \theta_k + \sigma_k \mu_{ik} \quad (75)$$

where  $\mu$  is normal and the  $z$  move the mean  $\beta$  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.