

# ML and Binary DVS - complications

Nathaniel Beck

Department of Politics, NYU, New York, NY 10012, [nathaniel.beck@nyu.edu](mailto:nathaniel.beck@nyu.edu)

QII - Week 3

1

## Latent variable interpretation

- Suppose that there is some unobserved (“latent”)

$$y^* = \mathbf{x}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N(0, 1). \quad (1)$$

Suppose that

$$y = 1 \text{ if } y^* > 0 \quad (2)$$

$$y = 0 \text{ otherwise} \quad (3)$$

$$\pi = P(y = 1) \quad (4)$$

$$\pi = P(y_i^* > 0) \quad (5)$$

$$= P(\mathbf{x}\boldsymbol{\beta} > -\epsilon) \quad (6)$$

$$= 1 - \Phi(-\mathbf{x}\boldsymbol{\beta}) \quad (7)$$

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

2

- 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.)

3

## 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.

4

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

$$\frac{\partial P(y = 1)}{\partial x_k} = \beta_k \Lambda(\mathbf{x}\boldsymbol{\beta})(1 - \Lambda(\mathbf{x}\boldsymbol{\beta})) \quad (8)$$

$$= \beta_k P(y = 1)P(y = 0) \quad (9)$$

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

$$\frac{\partial P(y = 1)}{\partial x_k} = F'(\mathbf{x}\boldsymbol{\beta})\beta_k \quad (10)$$

5

## Simulations: Clarify

- Can get confidence intervals of QOI 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).
- This implies any function of  $\hat{\beta}$ , and all QOIs are such, is uncertain
- 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.

6

- Make sure deal correctly with interactions or polynomials (if  $y = a + bx + dx^2$  neither b nor d are of interest per se, and  $\frac{dy}{dx} = b + 2d$ )
- For variables of little interest stick at typical values (mean, median)
- For discrete variables could set to mode (eg White, college) or might want to do a full comparison setting, say, M and then F.
- For continuous variables graphs with ci's are nice (but don't get them too cluttered)
- If typical CI is about k percent, leave off graph and just say figures accurate to that CI
- The bigger the change in x, the bigger the first difference. So do not go overboard. First difference over reasonable regions

7

## Analytic methods: Delta method

- Suppose  $\hat{\beta}$  is  $AN(\beta, V)$
- Suppose QOI is  $g(\beta)$
- From properties of ML, ML estimate of this QOI is  $g(\hat{\beta})$
- From simple calculus, the VCV of  $g(\hat{\beta})$  is  $\hat{R}V\hat{R}'$
- Where  $R = \frac{\partial g}{\partial \beta'}$
- And the hats are at the estimated ml of  $\beta$
- Since  $g$  is known, the derivatives are easy if messy
- Political scientists seem to prefer the ease of clarify
- Stata gives marginal effects in the mfx command
- Both are similar, but not identical

8

- 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 8 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  $\pi$  by any probability distribution function. Nagler proposes the “Burr-10”, which has an additional parameter  $\alpha$ , so

$$\pi = (1 + e^{x\beta})^{-\alpha} \quad (11)$$

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

9

## Scobit: Likelihood

$$\ln L = \sum_{i=1}^N (1 - y_i) \ln[F(-\mathbf{x}_i\beta)] + y_i \ln[1 - F(-\mathbf{x}_i\beta)] \quad (12)$$

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

- We can then test the null hypothesis that  $\alpha = 1$  in the usual manner. If not reject  $\alpha = 1$  can then use logit
- 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  $P = .5$ .
- Note that the scobit only works if the  $P$  where the impact is maximal is less than some value ( $.6x$ ). 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.

- 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 except for the parameter,  $\theta$ , which measures heteroskedasticity.

- $$\pi = \frac{1}{1 - e^{x\beta\theta}} \quad (13)$$

11

- 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 parametrize it.
- Note that one can just as easily do heteroskedastic probit.
- Just go back to latent form

$$y^* = \mathbf{x}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \quad (14)$$

where

$$\sigma^2 = e^{z\gamma} \quad (15)$$

- 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.

12

- Ordered DV
- Example Approval of President or Party ID
- Count DV
- Examples How many debates one watched
- Examples 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
- Example 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

13

## 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

$$y^* = \mathbf{x}\boldsymbol{\beta} + \epsilon, \epsilon \sim N(0, 1). \quad (16)$$

$$y = c \text{ if } y^* \geq \tau_{c-1} \quad (17)$$

$$y = c - 1 \text{ if } \tau_{c-1} > y^* \geq \tau_{c-2} \quad (18)$$

$$\vdots \quad (19)$$

$$y = 1 \text{ if } \tau_1 > y^* \geq 0 \quad (20)$$

$$y = 0 \text{ if } y^* < 0 \quad (21)$$

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

14

- 

$$L = \prod_{i=1}^N (\pi_0^i)^{y=0} \dots (\pi_c^i)^{y=c} \quad (22)$$

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

$$\pi_0 = \Phi(-\mathbf{x}\beta) \quad (23)$$

$$\pi_1 = \Phi(\tau_1 - \mathbf{x}\beta) - \Phi(-\mathbf{x}\beta) \quad (24)$$

$$\vdots \quad (25)$$

$$\pi_c = 1 - \Phi(\tau_{c-1} - \mathbf{x}\beta) \quad (26)$$

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.

15

## 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)
- Assumption of “parallel regressions” (Long). Since the latent  $y^*$  is always just  $\mathbf{x}\beta$ , the process moving from 0 to 1, 1 to 2, etc., are all identical except for the “constant term” (the threshold, as we have seen, is essentially the same as a constant, particularly if we have only 2 categories).

16

- Imagine a model of the number of whether a nation is pacific, realistic or aggressive which is an ordered variable, 0, 1, 2, which is measured somehow. The parallel regressions assumption means that the process of going from being totally pacific to realistic (some war) is same as from realistic to aggressive (lots of wars), except for constant term. But it may be that the process is totally different, and we should think of one process as moving a nation from pacific to realistic and a second moving a nation from realistic to aggressive.
- Another example would be apathetic, minor participation (voting), heavy involvement (at least donation).
- The likelihood  $f(w)$  would then be more complicated: the probability of getting a zero is given by one probit, while the probability of reaching the top tier, conditional on being in the middle tier, is another probit.
- This is a simple “hurdle mode.” We will see more about hurdle models in the event count context.