

ML and Binary DVS - intro

Nathaniel Beck

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

QII - Week 3

1

Binary Dependent Variable

- 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 \mathbf{x}_i . (Thus will either have a vector of parameters β or a scalar β .)
- Where we have iid obs, we can suppress the subscript i as needed
- For a vector of variables, we use k to generate a single variable (scalar x_k or β_k)

2

- 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.

3

Single index model

- Let $y = f(\mathbf{x}\boldsymbol{\beta}) = f(z)$ where $z = \mathbf{x}\boldsymbol{\beta}$
- Note that z is scalar, and we can take advantage of

$$\frac{\partial y}{\partial x_k} = \frac{\partial y}{\partial z} \frac{\partial z}{\partial x_k} = \frac{\partial y}{\partial z} \beta_k \quad (2)$$

- So for single index model interpretation is easy, just β_k times the partial of y wrt z
- Particularly nice since we can compare the ratio of marginal effects for two variables as just $\frac{\beta_k}{\beta_{k'}}$ which does not vary with \mathbf{x} .

4

- Suppose

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}\beta}} = \Lambda(\mathbf{x}\beta). \quad (3)$$

(The Λ just simplifies my typing!) We will usually denote the probability that observation i is one by π_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 $\mathbf{x}\beta$ for a vector of covariates. In this case the interpretation is on the linear form $\beta_1 x_{i1} + \beta_2 x_{i2} \dots + \beta_k x_{ik}$. 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.

5

- 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

$$L = (1 - \Lambda(\beta x_1))(1 - \Lambda(\beta x_2)) \dots (1 - \Lambda(\beta x_{\text{last zero}})) \Lambda(\beta x_{\text{first one}}) \Lambda(\beta x_{\text{next one}}) \dots \Lambda(\beta x_{\text{last one}}). \quad (4)$$

- Then much easier to maximize the log of this which is the sum of the individual logs
- 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

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^n (y_i - \Lambda_i) \mathbf{x}_i = \mathbf{0} \quad (5)$$

6

- For the constant term we have

$$\sum_{i=1}^n (y_i - \Lambda_i) = 0 \quad (6)$$

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 π 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

$$\mathbf{H} = - \sum_{i=1}^n \Lambda_i (1 - \Lambda_i) \mathbf{x}_i \mathbf{x}_i \quad (7)$$

7

- Λ_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 Λ_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 β '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.

8

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

$$P(y_1 = 0)P(y_2 = 0) \dots P(y_{\text{last zero}} = 0)P(y_{\text{first one}} = 0) \dots P(y_{\text{last one}} = 1) \quad (8)$$

with appropriate replacement for the probabilities.

- Any admissible model for the probabilities (that is, $p_{ij} = f(\mathbf{x}_i, \beta)$) that returns values strictly in the unit interval, is a perfectly reasonable model.
- Thus, for probit, we would use $P(y = 1) = \Phi(\mathbf{x}\beta)$ where Φ 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.

9

- 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 β , 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 π_i as close to one as we can, and for $y_i = 0$ we get π_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 π '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 β 's by some number, which would have the same effect on the variance. So we normalize by allowing the β '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.