Part 4: Binary Choice

Chris Conlon

Microeconometrics

May 30, 2017

Binary Choice: Overview

Many problems we are interested in look at discrete rather than continuous outcomes.

- We are familiar with limitations of the linear probability model (LPM)
  - Predictions outside of $[0, 1]$
  - Estimates of marginal effects need not be consistent.
- What about the case where Y is binary and a regressor X is endogenous?
  - The usual 2SLS estimator is NOT consistent.
  - Or we can ignore the fact that Y is binary...
  - Neither seems like a good option
- Suppose we have panel data on repeated binary choices
  - Adding FE to the probit model produces biased estimates.

Problem #1: Endogeneity

Four possible solutions (maybe there are more?)

1. Run the LPM with instruments (Suggested by MHE).
2. Specify the distribution of errors in first and second stage and do MLE (`ivprobit` in STATA).
3. Control Function Estimation
4. ‘Special Regressor’ Methods

Problem #1: Endogeneity

Setup:

- Binary variable D: the outcome of interest
- X is a vector of observed regressors with coefficient β
  - (Can think about X^e: endogenous and X^0: exogenous).
  - In an treatment model we might have that T is a binary treatment indicator within X
- ϵ is unobserved error. Specifying $f(e)$ can give logit/probit.
- Threshold Crossing / Latent Variable Model:

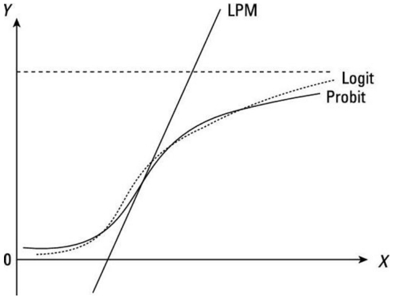
$$D = \mathbf{1}(X\beta + \epsilon \geq 0)$$

- Goal is not usually $\hat{\beta}$ or it's CI, but rather $P(D = 1|X)$ or $\frac{\partial P[D=1|x]}{\partial X}$ (marginal effects).

Linear Probability Model

Consider the LPM with a single continuous regressor

- LPM prediction departs greatly from CDF long before $[0,1]$ limits.
- We get probabilities that are too extreme even for $X\hat{\beta}$ "in bounds".
- Some (MHE) argue that though $\hat{Y}$ is flawed, constant marginal effects are still OK.



Some well known textbooks

(Baby) Wooldrige:

> *"Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample'.' (2009, p. 249)*

- Mentions heteroskedasticity of error (which is binomial) but does not address the violation of the first LSA.

Some well known textbooks

Angrist and Pischke (MHE)

- several examples where marginal effects of probit and LPM are "indistinguishable".

> *...while a nonlinear model may fit the CEF (conditional expectation function) for LDVs (limited dependent variable models) more closely than a linear model, when it comes to marginal effects, this probably matters little. This optimistic conclusion is not a theorem, but as in the empirical example here, it seems to be fairly robustly true.(2009, p. 107)*

and continue...

> *...extra complexity comes into the inference step as well, since we need standard errors for marginal effects. (ibid.)*

Linear Probability Model

How does the LPM work?

$$D = X\beta + \varepsilon$$

- Estimated $\hat{\beta}$ are the MFX.
- With exogenous X we have $E[D|X] = Pr[D = 1|X] = X\beta$.
- If some elements of X (including treatment indicators) are endogenous or mismeasured they will be correlated with E.
- In that case we can do IV via 2SLS or IV-GMM given some instruments Z.
- We need the usual $E[\varepsilon|X] = 0$ or $E[\varepsilon|Z] = 0$.
- An obvious flaw: Given any $\varepsilon|X$ must equal either $1 - X\beta$ or $-X\beta$ which are functions of X
- Only the trivial binary X with no other regressors satisfies this!

Alarming Example: Lewbel Dong and Yang (2012)

- LPM is not just about taste and convenience.
- Three treated observations, three untreated
- Assume that $f(\varepsilon) \sim N(0, \sigma^2)$

$$D = I(1 + Treated + R + \varepsilon \geq 0)$$

- Each individual treatment effect given by:

$$I(2 + R + \varepsilon \geq 0) - I(1 + R + \varepsilon \geq 0) = I(0 \leq 1 + R + \varepsilon \leq 1)$$

- All treatment effects are positive for all (R, ε).
- Construct a sample where true effect $= 1$ for 5th individual, 0 otherwise. $ATE = \frac{1}{6}$.

Alarming Example: Lewbel Dong and Yang (2012)

```
. list
     |     R   Treated   D |
  1. |  -1.8         0   0 |
  2. |   -.9         0   1 |
  3. |  -.92         0   1 |
  4. |  -2.1         1   0 |
  5. | -1.92         1   1 |
  6. |    10         1   1 |

. reg D Treated R, robust

Linear regression                                Number of obs     =          6
                                                 F(2, 3)           =       1.02
                                                 Prob > F          =     0.4604
                                                 R-squared         =     0.1704
                                                 Root MSE          =     .60723


---------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t     P>|t|     [95% Conf. Interval]
-------+------------------------------
     Treated |  -.1550841   .5844637    -0.27   0.808     -2.015108     1.70494
           R |   .0484638   .0419179     1.16   0.331     -.0849376    .1818651
       _cons |   .7251463   .3676811     1.97   0.143     -.4449791    1.895272
---------------------------------------
. nlcom _b[Treated]/_b[R]
       _nl_1:  _b[Treated]/_b[R]
---------------------------------------
           D |      Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------+------------------------------
       _nl_1 |       -3.2   10.23042    -0.31   0.754     -23.25125    16.85125
---------------------------------------
```

Alarming Example: Lewbel Dong and Yang (2012)

- That went well, except that:
  - we got the wrong sign of β_T
  - β_1/β_2 was the wrong sign and three times too big.
- this is not because of small sample size or $\beta_1 \approx 0$.
- As $n \to \infty$ we can get an arbitrarily precise wrong answer.
- We don't even get the sign right!
- This is still in OLS (not much hope for 2SLS).

```
. expand 30
(...)
. reg D Treated R, robust

Linear regression                                Number of obs      =        180
                                                 F(2, 177)          =      59.93
                                                 Prob > F           =     0.0000
                                                 R-squared          =     0.1704
                                                 Root MSE           =       .433

---------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
-------+--------------------------------
     Treated |  -.1550841   .0760907    -2.04   0.043     -.3052458   -.0049224
           R |   .0484638   .0054572     8.88   0.000      .0376941    .0592334
       _cons |   .7251463    .047868    15.15   0.000      .6306808    .8196117
```

Solution #0 : LPM

Advantages

- Just like 2SLS.
- Computationally easy (no numerical searches)
- Missing Z is about efficiency not consistency.
- X^e can be discrete or continuous (same estimator)
- allows general heteroskedasticity (random coefficients)

Disadvantages

- $\hat{F}_D$ is linear not S-shaped, approximation valid for small range of X.
- $\hat{F}_D$ can be outside $[0, 1]$.
- no element of X can have ∞ support (e.g. no normally distributed regressors).
- ε not independent of any regressors (even the exogenous ones). How do we also get $E[X^0\varepsilon] = 0$?

Solution #1 : MLE

$$D = I(X'\beta + \varepsilon \geq 0) \quad \text{and} \quad X^e = G(Z, \theta, e)$$

- Fully specified G (could be vector). Could be linear if X^e continuous or probit if X^e binary.
- Need to fully specify distribution of $(\varepsilon, e, |Z)$, parametrized.
- Implementation (see book), `biprobit` for joint normal in Stata.

Solution #1 : MLE

Advantages

- Nests logit, probit, etc. as special cases.
- Can have any kind of X^e
- Allows heteroskedasticity, random coefficients
- Asymptotically efficient (if correctly specified)

Disadvantages

- Need to parametrize everything G, $F_{\varepsilon,e|Z}$.
- Numerical optimization issues
- Many nusiance parameters, sometimes poorly identified, especially with discrete X^e, correlation between latent (ε, e).
- Need to know all required instruments Z. Omitting just one Z causes inconsistency in G. (Not sure if something is exogenous, too bad!).

Solution #2 : Control Functions

$$\begin{aligned} D &= I(X'\beta + \varepsilon \geq 0) \quad \text{and} \\ X^e &= G(Z) + e \quad \text{or} \quad X^e = G(Z, e) \quad \text{identified and invertible in } e \\ \varepsilon &= \lambda' e + U \quad \text{or} \quad \varepsilon = H(U, e) \quad \text{with conditions and } U \perp Z, e. \end{aligned}$$

Simple Case:

- Estimate a vector of functions G in the X^e models, get estimated errors $\hat{e}$.
- Estimate the D model including $\hat{e}$ as additional regressors in addition to X.
- This "cleans" the errors in U.

Solution #2 : Control Functions (Stata Version)

- ivprobit assumes that $G(Z, e)$ is linear, and (e, ε) jointly normal, independent of Z.
- not exactly the semi-parametric flexibility we were looking for...
- It is actually Control Function not IV

$$\begin{aligned} D &= I(X^e \beta_e + X^0 \beta_0 + \varepsilon \geq 0) \\ X^e &= \gamma Z + e \end{aligned}$$

Run first-stage OLS and get residuals $\hat{e}$. Then plug into

$$D = I(X^e \beta_e + X^0 \beta_0 + +\lambda \hat{e} + U \geq 0)$$

and do a conventional probit estimator.

- If you forget a Z_2 the resulting model isn't even probit!
- God help you if X^e isnt continuous.

Solution #2 : Control Functions

$$D \;=\; I(X'\beta + \varepsilon \geq 0), \quad X^e = G(Z, e), \varepsilon = H(U, e), \quad U \perp X, e.$$

Much stronger requirements that 2SLS

- Must be able to solve for errors e in X^e equations (not just orthogonality)
- Endogeneity must be caused only by ε relation to e so after conditioning on e must be that $f(\varepsilon|e, X^e) = f(\varepsilon|e)$.
- I need a consistent estimator for e which means nothing is omitted.

Not Quite MLE

- First stage can be semi/non-parametric .
- Don't need to fully specify joint distribution of (ε, e) (Stata does though!).

Solution #2 : Control Functions

Advantages

- Nests logit, probit, etc. as special cases
- Requires less parametric information than MLE
- Some versions are computationally easy without numerical optimization (Bootstrap!)
- Less efficient than MLE due to less restrictions,but can be semiparametrically efficient given information.

Solution #2 : Control Functions (Disadvantages: Not well known)

- Only allows limited heteroskedasticity
- Need to correctly specify vector $G(Z, e)$ including all Z. Omitting a Z or misspecified G causes inconsistency because we need to have joint conditions on (ε, e).
- Generally inconsistent for X^e that is discrete, censored, limited, or not continuous.
- If you cannot solve for a latent e in $G(Z, e)$ then you can't get $\hat{e}$ for the censored observations (e.g.: $X^e = \max(0, Z'\gamma + e)$.
- An observable e is $e = X^e - E[X^e|Z]$ but for discontinuous X^e that e violates assumptions (except in very strange cases)
  - Ex: $\varepsilon = [X^e - E[X^e|Z]]\lambda + U$ satisfies CF, but if X^e is discrete then e has some strange distribution that depends on regressors.
  - Hard to generate a model of behavior that justifies this!

Solution #2 : Control Functions Generalized Residuals

What if X^e isn't continuous? Technically possible...

- Given the probit estimate in first stage we could construct a generalized residual (see Imbens and Wooldridge notes)
- $e^g \propto E[\varepsilon|Z, e]$. An estimate $\hat{e}^g$ of e^g can be included as a regressor in the model to fix the endogeneity problem, just as $\hat{e}$ would have been used if the endogenous regressor were continuous.

Why would you ever want to do this..

- In the linear model we should just do IV with far fewer restrictions
- In the nonlinear model, $\hat{e}^g$ requires almost as many assumptions as MLE which is efficient!

Solution #3 : Special Regressor

This approach draws on these two papers:

- Lewbel (JoE 2002)
- Dong and Lewbel (Econometric Reviews 2015)

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

Special Regressor V has three properties

- Exogenous $E[\varepsilon|V] = 0$
- Additively separable in the model
- Continuous distribution and large support (such as Normal)
- Helpful to have thick-tails (kurtosis). Why? We want to trace $Pr(D|V)$ from $[0, 1]$.

Learn about SR

Binary, ordered, and multinomial choice, censored regression, selection, and treatment models (Lewbel 1998, 2000, 2007a), truncated regression models (Khan and Lewbel 2007), binary panel models with FE (Honore and Lewbel 2002), dynamic choice models (Heckman and Navarro 2007, Abbring and Heckman 2007), contingent valuation models (Lewbel, Linton, and McFadden 2008), market equilibrium models of multinomial choice (Berry and Haile 2009a, 2009b), models with (partly) nonseparable errors (Lewbel 2007b, Matzkin 2007, Briesch, Chintagunta, and Matzkin 2009).

Other empirical applications: Anton, Fernandez Sainz, and Rodriguez-Poo (2002), Cogneau and Maurin (2002), Goux and Maurin (2005), Stewart (2005), Lewbel and Schennach (2007), and Tiwari, Mohnen, Palm, and van der Loeff (2007).

Precursors: Matzkin (1992, 1994) and Lewbel (1997).

Recent theory: Magnac and Maurin (2007, 2008), Jacho-Ch·vez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

Solution #3 : Special Regressor

Requirements:

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

- Exogenous $E[\varepsilon|V] = 0$ (Strict Exogeneity) This is the key!
- Additively separable in the model
- Continuous distribution and large support (such as Normal)
- NOT interacted with other regressors.
- enters LINEARLY, e.g. V must be continuously distributed after conditioning on other regressors
- Can normalize its coefficient to 1
- 2SLS Assumptions: $E[\varepsilon|Z] = 0$ and $E[Z'X]$ is full rank.

Solution #3 : Special Regressor: How it works

1. Demean or center V at zero.
2. Assume that $f_v(V|Z, X^e) = f_v(V|Z)$ and let $\hat{f}_v(V|Z)$ be a nonparametric kernel estimator of $f_v(V|Z)$. Or just use a kernel of the whole thing $\hat{f}_v(V|Z, X^e)$
3. For each observation i, Construct $\hat{T}_i = I[D_i - I(V_i \geq 0)]/\hat{f}_v(V_i|Z_i)$
4. Linear 2SLS regression of $\hat{T}$ on X using instruments Z to get the estimated coefficients $\hat{\beta}$.

Here $\hat{f}_v(V|Z)$ is high dimensional. So will consider some simpler parametric or semi-parametric version of f_v.
By properly adjusting T_i we guarantee to stay in $[0, 1]$.

Solution #3 : Special Regressor Advantages

- Unlike LPM it stays “in bounds” and is consistent with threshold crossing models.
- Unlike MLE and CF, does not require correctly specified first stage model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike MLE, the SR method has a linear form, not requiring iterative search
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML there is a single estimation method, regardless of the characteristics of X^e
- Unlike MLE, the SR method permits unknown heteroskedasticity in the model errors.

Solution #3 : Special Regressor Disadvantages

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

- Because assumptions are weaker we give up a lot of potential efficiency (larger SEs).
- Of course this presumes the assumptions were valid and alternatives were consistent.
- SR Methods are generally valid under more general conditions.

The average index function (AIF)/ Propensity Score

In the original problem

$$D = I(X'\beta + \varepsilon \geq 0)$$

- V is part of X with coefficient $= 1$
- When $\varepsilon \perp x$ write the propensity score: $E[D|X] = E[D|X\beta] = F_{-\varepsilon}(X\beta) = Pr(-\varepsilon \leq X\beta)$.
- Under independence $X \perp \varepsilon$ these are the same, under endogeneity or even heteroskedasticity they are not.

The average index function (AIF)/ Propensity Score

- Blundell and Powell (ReStud 2004, this is actually the most important control function paper) use the average structural function (ASF) $= F_{-\varepsilon}(X\beta)$ to summarize choice probabilities. But when $\varepsilon \perp X$ is violated then they have to compute $F_{-\varepsilon|X}(X\beta)$ which is quite difficult (especially semiparametrically).
- Lewbel, Dong and Tang (CJE 2012) propose using the AIF estimator $E[D|X\beta]$ instead.
- Like ASF the AIF is based on the estimated index $X\beta$ and is equal to the propensity score if $\varepsilon \perp X$. However, when this is violated (endogeneity, heteroskedasticity) the AIF is easier to estimate, via unidimensional nonparametric regression of D on $X\beta$.

What is the difference

Propensity Score: Conditions on ALL covariates using $F_{-\varepsilon|X}$.

ASF: Conditions on no covariates using $F_{-\varepsilon}$.

AIF: Conditions on index only using $F_{-\varepsilon|X\beta}$.

- Unlike ASF, AIF is always identified and easy to estimate.
- Unlike Propensity score AIF uses β and isn't high dimensional
- ASF, AIF and propensity score all coincide under exogeneity.

Marginal Effects

- With exogenous X: MFX are $m(X) = p'(X) = \frac{\partial E[D|X\beta]}{\partial X}$.
- Let $f_{-\varepsilon}$ be marginal pdf of $-\varepsilon$. If $D = I(X'\beta + \varepsilon \geq 0)$ with $\varepsilon \perp X$ then:

$$m(X\beta)\beta = \frac{\partial E[D|X]}{\partial X} = \frac{\partial E[D|X\beta]}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$$

With endogenous X:

- Propensity Score marginal effects are $m(X) = p'(X) = \frac{\partial E[D|X]}{\partial X}$.
- ASF marginal effects are $m(X) = \frac{\partial ASF(X'\beta)}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$.
- AIF marginal effects are $m(X) = \frac{\partial ASF(X'\beta)}{\partial X'\beta}\beta = \frac{\partial E[D|X'\beta]}{\partial X'\beta}\beta$

Given $\hat{\beta}$ ASF and AIF mfx require just one dimensional index derivative.

Binary Choice with Endogenous Regressors

- Linear probability models, Maximum Likelihood, and Control functions (including `ivprobit` have more drawbacks and limitations than are usually recognized.
- Special Regressor estimators are a viable alternative (or at least they have completely different drawbacks and may be more generally applicable than has been recognized).
- In practice, best might be to try all estimators and check robustness of results. Can use marginal effects to normalize them the same when comparing.
- Average Index Functions can be used to construct estimated probabilities and comparable marginal effects across estimators, often simpler to calculate than Average Structural Functions.
- Implementation of special regressor in Stata is done in `sspecialreg`.

Empirical Example: Dong and Lewbel (2015)

- Binary dependent variable: does i migrate from one state to another.
- Special Regressor V_i: age. Human capital theory suggests it should appear linearly (or at least monotonic) in a threshold crossing model
- Migration is drive by maximizing expected lifetime income and potential gain from a permanent change in income declines linearly in age.
- V_i is defined as negative of age, demeaned so that coefficient is positive with mean zero.
- Other endogenous regressors: family income pre migration, home ownership.

Empirical Example: Dong and Lewbel (2015)

As a reminder, normally we would be in trouble here:

- MLE would be very complicated with multiple endogenous variables
- Control functions `ivprobit` won't work with 0/1 homeowner variable.

Empirical Example: Dong and Lewbel (2015)

1990 PSID

- male head of household (23-59 years), completed education and not-retired (key!)
- $D = 1$ indicates migration during 1991-1993.
- 4689 Individuals, 807 migrants.
- Exogenous regressors: years of education, # of children, white, disabled, married.
- Instruments: level of govt benefits in 1989-1990, state median residential tax rate.

Empirical Example: Dong and Lewbel (2015)

Specifications

- Special Regressors: kernel density vs. sorted data density.
- Special Regressor: homoskedastic vs. heteroskedastic errors.
- LPM vs 2SLS
- Probit (assuming exogeneity)
- Control Function (`ivprobit`) misspecified for `homeowner` endogenous binary variable.

Empirical Example

Table: Marginal effects: binary outcome, binary endogenous regressor

	kdens	sortdens	kdens_hetero	sortdens_hetero	IV-LPM	probit	ivprobit
age	0.0146 (0.003)***	0.0112 (0.003)***	0.0071 (0.003)*	0.0104 (0.003)***	-0.0010 (0.002)	0.0019 (0.001)**	-0.0005 (0.007)
log income	-0.0079 (0.028)	0.0024 (0.027)	0.0382 (0.024)	0.0176 (0.026)	0.0550 (0.080)	-0.0089 (0.007)	0.1406 (0.286)
homeowner	0.0485 (0.072)	-0.0104 (0.065)	-0.0627 (0.059)	-0.0111 (0.061)	-0.3506 (0.204)	-0.0855 (0.013)***	-1.0647 (0.708)
white	0.0095 (0.008)	0.0021 (0.010)	0.0021 (0.007)	0.0011 (0.008)	0.0086 (0.018)	-0.0099 (0.012)	0.0134 (0.065)
disabled	0.1106 (0.036)**	0.0730 (0.042)	0.0908 (0.026)***	0.0916 (0.037)*	0.0114 (0.055)	-0.0122 (0.033)	0.0104 (0.203)
education	-0.0043 (0.002)*	-0.0023 (0.003)	-0.0038 (0.002)*	-0.0036 (0.002)	0.0015 (0.004)	0.0004 (0.002)	0.0047 (0.015)
married	0.0628 (0.020)**	0.0437 (0.028)	0.0258 (0.013)	0.0303 (0.020)	0.0322 (0.031)	-0.0064 (0.017)	0.0749 (0.114)
nr. children	-0.0169 (0.005)***	-0.0117 (0.005)*	0.0006 (0.002)	-0.0021 (0.003)	0.0137 (0.006)*	0.0097 (0.005)*	0.0502 (0.023)*

Note: bootstrapped standard errors in parentheses (100 replications)

Empirical Example: Dong and Lewbel (2015)

- SEs of MFX are computed from 100 bootstrap replications
- MFX of special regressor (age) is estimated as positive and significant but LPM and ivprobit estimate negative effects!
- household income and home ownership status do not seem to play significant roles in migration decision.
- Kernel density estimator seems to give most significant results.