Advanced Binary Choice: LPM? Incidental Parameters

Chris Conlon

April 26, 2020

Panel Data Econometrics

Intro

Binary Choice: Overview

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

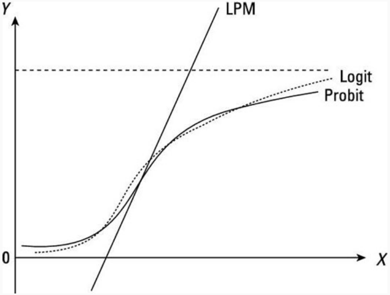
- We are familiar with limitations of the linear probability model (LPM)
  - Predictions outside of $[0, 1]$
  - Estimates of marginal effects need not be consistent.
- Suppose we have panel data on repeated binary choices
  - When N is large and T is small.
  - Adding FE to the logit/probit model produces biased (and inconsistent) estimates.
- 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

Problem #1:
Failures of the LPM

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 given X) 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 ε.
- In that case we can do IV via 2SLS or IV-GMM given some instruments Z.

Linear Probability Model

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

- 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!
- Should you believe $\widehat{\beta}_{LPM}$ if $E[\varepsilon|X] \neq 0$?

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)

For stable draws – I chose the outcome D:

```
draw_sample = function(){
  tibble(
  r = c(-1.8, -0.9, -0.92,-2.1, -1.92, 10),
  treated = c(0,0,0,1,1,1),
  true_te = c(0,0,0,0,1,0),
  error = rnorm(6),
  D = c(0,1,1,0,1,1)
  )  %>% mutate(y = (r + true_te * treated + error)>0)
  }
```

Alarming Example: Lewbel Dong and Yang (2012)

```
  lm(D~ treated+r, data=fake_data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.725146   0.008016   90.46   <2e-16 ***
treated     -0.155084   0.011938  -12.99   <2e-16 ***
r            0.048464   0.001381   35.09   <2e-16 ***
```

A Small Sample Issue ?

```
# Now with 1000x more data
for (n in 1:1000){data[[n]]=draw_sample()}
df<-bind_rows(data)
summary(lm(y~ treated+r, data=df))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2364993  0.0055112   42.91   <2e-16 ***
treated     0.0055820  0.0082069    0.68    0.496
r           0.0763806  0.0009495   80.44   <2e-16 ***
```

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).

Problem #2:
Incidental Parameters Problem

Problem #2: Fixed Effects and Incidental Parameters

Threshold Crossing / Latent Variable Model:

$$y_{it} = \mathbf{1}(X_{it}\beta + \epsilon_{it} \geq 0)$$

Individuals $i = 1, \ldots, N$, and periods $t = 1, \ldots, T$.

- Goal is not usually $\hat{\beta}$ or it's CI
- Instead $P(y_{it} = 1|x_{it})$ or $\frac{\partial P[y_{it}=1|x_{it}]}{\partial x_{it}}$ (marginal effects).

What about if we want to add FE α_i?

$$y_{it} = \mathbf{1}(X_{it}\beta + \alpha_i + \epsilon_{it} \geq 0)$$

Best case scenario $E[\epsilon_{it}|x_{i1}, \ldots, x_{iT}, \alpha_i] = 0$.

Fixed Effects and Incidental Parameters

Let's try and difference out (within transform) the FE:

$$E[y_{it} - y_{i,t-1}|X_i, \alpha_i] = E[F(X_{i,t}\beta + \alpha_i) - F(X_{i,t-1}\beta + \alpha_i)|X_i]$$

- We can't difference out the FE anymore!
- α_i doesn't have the interpretation as $\overline{y}_i$.
- The effect of α_i will now depend on $X_{i,t}$.
- We need to estimate N parameters $\alpha_1, \ldots, \alpha_N$.

Fixed Effects and Incidental Parameters

- FE requires maximizing LL over (α_i, β).
- FE model inconsistent as $N \to \infty$ as T fixed.
- Under OLS we have a consistent estimator of α_i, as N gets big, random noise in α_i washes out.
- Under Logit/Probit/etc. now we only have T observations.
- Under logit/probit bias in $\hat{\beta}$ can be pretty bad.
- Under probit, bias in marginal effects tends not to be as bad as bias in $\hat{\beta}$ (Hahn and Newey, 2004).

Fixed Effects and Incidental Parameters

$$\begin{aligned} Pr(y_{it} = 1|z_{it}) &= \frac{\exp(z_{it})}{1 + \exp(z_{it})} \\ z_{it} &= x_{it} + \alpha_i + u_{it} \\ x_{it} &= \alpha_i + v_{it} \end{aligned}$$

Generate data from a binary logit with some simple FE.

- $\alpha_i \sim N(0,1)$
- $v_{it} \sim N(0,1)$
- $u_{it} \sim N(0,2)$

How bad is it? An example

```
library(fabricatr)
fe.sd <- 1 # Specify the standard deviation of the fixed effed
x.sd  <- 1 # Specify the base standard deviation of x
nperson <- 5000 # Number of persons
nobs <- 5      # Number of observations per person
panels <- fabricate(
  individuals = add_level(N = nperson, id_fe = rnorm(N,0,fe.sd)),
  periods = add_level(N = nobs, nest = FALSE),
  obs = cross_levels(
    by = join(individuals, periods),
    # put the FE into X so there is something to de-mean
    x = id_fe + rnorm(N,0,x.sd),
    z = 1*id_fe  + 1*x, # + rnorm(N,0,1) -- adding this breaks bias correction
    logit_prob = exp(z)/(1+exp(z)),
    yl= logit_prob>runif(N)
  )
)
```

How bad is it? An example

```
summary(glm(yl ~ x, data = panels, family = "binomial"))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.0002284  0.0162293  -0.014    0.989
x            1.4135577  0.0181815  77.747   <2e-16 ***

summary(glm(yl ~ x+factor(id_fe), data = panels, family = "binomial"))
Estimates:
  Estimate Std. error z value Pr(> |z|)
x  1.36383    0.02852   47.82    <2e-16 ***
...
```

The latter takes 10+ minutes and often FE make things worse.

Two Solutions

Two possible solutions:

1. Work out an analytic expression for the bias and subtract it off: analytic bias correction.
2. Conditional Logit. Exploit a sufficient statistic formulation (Chamberlain)

Approach #1: Estimate the FE, Fix the Bias

What does `bife` do?

- Estimate the model with the FE in there as parameters (dummy variables).
- This model is biased and inconsistent.
- It is a big pain to estimate if N becomes large (lots of parameters, lots of derivatives, etc.)
    - Exploits the sparsity of the Hessian, much faster than `glm`
- Fix the bias on the back end by working out an analytic expression or jackknife: Hahn and Newey (2004) or Stammann, Heiss, and McFadden (2016).

bife First the inconsistent model

```
res<-bife(yl ~ x | id_fe, data = panels, 'logit')

binomial - logit link

yl ~ x | id_fe

Estimates:
  Estimate Std. error z value Pr(> |z|)
x  1.36383    0.02852   47.82    <2e-16 ***
--
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

residual deviance= 15933.43,
null deviance= 22929.24,
nT= 16540, N= 3308

( 8460 observation(s) deleted due to perfect classification )

...
```

bife Analytic Bias Correction (Ex-post)

```
> summary(bias_corr(res))
binomial - logit link

yl ~ x | id_fe

Estimates:
  Estimate Std. error z value Pr(> |z|)
x  1.02455    0.02483   41.26    <2e-16 ***
--
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

residual deviance= 16089.15,
null deviance= 22929.24,
nT= 16540, N= 3308

( 8460 observation(s) deleted due to perfect classification )

Number of Fisher Scoring Iterations: 5
```

Conditional MLE

Imagine a sufficient statistic S_i:

$$f\left(Y_i|X_i, S_i, \theta, \alpha_i\right) = f\left(Y_i|X_i, S_i, \theta\right)$$

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} f\left(Y_i|X_i, S_i, \theta\right)$$

These are hard to find, but examples:

- Binary Logit
- Gaussian linear model
- Poisson
- (and as we have seen) PH model for durations

Conditional Logit (Chamberlain 1984/1992)

Here $S_i = \sum_t y_{it}$. Assume that $T = 2$ to make things clear

$$E[y_{i1}|X_i, \alpha_i, y_{i1} + y_{i2} = 1]$$

Consider the $\sum_t y_{it} = 1$ case:

$$\Pr\left(\mathrm{y_{i1}} + \mathrm{y_{i2}} = 1\right) = \Pr\left(\mathrm{y_{i1}} = 0, \mathrm{y_{i2}} = 1\right) + \Pr\left(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0\right)$$

And then we can write:

$$\begin{aligned}
\Pr\left(\mathrm{y_{i1}} = 1\right) &= \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{it}}\beta\right) / \left[1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{it}}\beta\right)\right] \\
\Pr\left(\mathrm{y_{i1}} = 0\right) &= 1 - \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{it}}\beta\right) / \left[1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{it}}\beta\right)\right] \\
&= 1/\left[1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{it}}\beta\right)\right]
\end{aligned}$$

Conditional Logit (Chamberlain 1984/1992)

$$\Pr\left(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0\right) = \frac{\exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right)}{1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right)} \cdot \frac{1}{1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)}$$

$$\Pr\left(\mathrm{y_{i1}} = 0, \mathrm{y_{i2}} = 1\right) = \frac{1}{1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right)} \cdot \frac{\exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)}{1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)}$$

Putting it together

$$\begin{aligned} \Pr\left(\mathrm{y_{i1}} + \mathrm{y_{i2}} = 1\right) \quad &= \Pr\left(\mathrm{y_{i1}} = 0, \mathrm{y_{i2}} = 1\right) + \Pr\left(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0\right) \\ &= \frac{\exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right) + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)}{\left(1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{il}}\beta\right)\right)\left(1 + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)\right)} \end{aligned}$$

And the conditional:

$$\Pr\left(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0 | \mathrm{y_{i1}} + \mathrm{y_{i2}} = 1\right) \quad = \quad \frac{\Pr\left(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0\right)}{\Pr\left(\mathrm{y_{i1}} + \mathrm{y_{i2}} = 1\right)}$$

$$\frac{\exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right)}{\exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i1}}\beta\right) + \exp\left(\alpha_i + \mathrm{x}'_{\mathrm{i2}}\beta\right)} = \frac{\exp\left(\mathrm{x}'_{\mathrm{i1}}\beta\right)}{\exp\left(\mathrm{x}'_{\mathrm{i1}}\beta\right) + \exp\left(\mathrm{x}'_{\mathrm{i2}}\beta\right)}$$

Eliminating the FE (Chamberlain 1984/1992)

Notice that we've now eliminated the FE!

$$\Pr\left(\mathrm{y}_{\mathrm{i1}} = 1, \mathrm{y}_{\mathrm{i2}} = 0 | \mathrm{y}_{\mathrm{i1}} + \mathrm{y}_{\mathrm{i2}} = 1\right) = \frac{1}{1 + \exp\left(\mathrm{x}_{\mathrm{i2}} - \mathrm{x}_{\mathrm{i1}}\right)' \beta}$$

$$\Pr\left(\mathrm{y}_{\mathrm{i1}} = 1, \mathrm{y}_{\mathrm{i2}} = 0 | \mathrm{y}_{\mathrm{i1}} + \mathrm{y}_{\mathrm{i2}} = 1\right) = \frac{\exp\left(\mathrm{x}_{\mathrm{i2}} - \mathrm{x}_{\mathrm{i1}}\right)' \beta}{1 + \exp\left(\mathrm{x}_{\mathrm{i2}} - \mathrm{x}_{\mathrm{i1}}\right)' \beta}$$

- As per usual if x_{it} doesn't vary over t it drops out too.
- We can skip the $y_{i1} + y_{i2} = 2$ and $y_{i1} + y_{i2} = 0$ case, why?
  - The FE $\alpha_i \rightarrow \pm\infty$

In practice clogit in R

```
library(survival)
> summary(clogit(yl~ x +strata(id_fe), data=panels))
Call:
coxph(formula = Surv(rep(1, 25000L), yl) ~ x + strata(id_fe),
    data = panels, method = "exact")

  n= 25000, number of events= 12517

     coef exp(coef) se(coef)     z Pr(>|z|)
x 1.04162   2.83381  0.02405 43.31   <2e-16 ***
```

Thanks!