EC968

Panel Data Analysis

Steve Pudney

ISER

Lecture 4: Models for discrete variables

- Types of discreteness
- Linear regression
- Latent linear regression
- Binary models: conditional and random-effects logit
- Dynamic discrete models

Discreteness

Inherent discreteness involves transitions between states

(*e.g.* employment & unemployment, married, unmarried)

Observational discreteness is an artefact of the observation process

(*e.g.* income questions based on ranges, Likert attitudinal questions)

Forms of discreteness

Censoring/corner solutions generate variables which are mixed discrete/continuous

(*e.g.* hours of work are 0 for non-employed, any positive value for employees)

Truncation involves discarding part of the population

(*e.g.* low-income targeted samples, or earnings models for employees only)

Count variables are the outcome of some counting process

(*e.g.* the number of durables owned, or the number of employees of a firm)

Binary variables reflect a distinction between two states

(*e.g.* unemployed or not, married or not)

Ordinal variables are ordered variables, possibly taking more than two values

(*e.g.* happiness on a scale 1=miserable … 5=ecstatic)

Unordered variables reflect outcomes which are discrete but with no natural ordering

(*e.g.* choice of occupation)

Binary models

We concentrate on binary models, with a dependent variable

y_{it} = 0 or 1

This describes:

- situations of choice between 2 alternatives
- sequences of events defining durations

E.g. suppose:

- $\mathbf{y}_i$ = (0, 0, 0, 0, 1, 1, 1, 0, 1, 1) is a monthly panel observation
- 0 indicates unemployment, 1 indicates employment

Then $\mathbf{y}_i$ represents a history of 4 months' unemployment followed by 3 months' employment, followed by 1 month's unemployment then 2 months' employment.

An alternative to modelling the sequence $\mathbf{y}_i$ is to model the set of durations: (U4, E3, U1, E2) $\Rightarrow$ survival analysis

An important issue concerns dynamics – how does the length of time already spent out of work affect this month's probability of finding work: *duration dependence*

Why are special methods needed ?

Consider a binary variable, y_{it} = 0 or 1

Common practice is to use a linear probability model:

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it} \qquad (1)$$

With panel data methods (*e.g.* within-group or random-effects)

Linear model implies:

$$E(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) \equiv \Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)$$

Model (1) requires:

$$P(\mathbf{z}_i, \mathbf{x}_{it}, u_i) \approx \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i$$

But this may fall outside the admissible [0, 1] interval.

Moreover, $\text{var}(y_{it} \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)[1-P(\mathbf{z}_i, \mathbf{x}_{it}, u_i)]$ is not constant $\Rightarrow$ heteroskedasticity is a problem

Latent regression models: the binary case

Define a latent (unobservable) continuous counterpart, y_{it}^* (*e.g.* if y_{it}=1 defines employment, then:

y_{it}^* = offered wage – reservation wage).

Let y_{it}^* be generated by a linear regression structure:

$$y_{it}^* = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it} \qquad (1)$$

Then employment is chosen whenever (offered wage-reservation wage) is positive:

$$y_{it} = 1(y_{it}^* > 0)$$

$$\Rightarrow \quad \Pr(y_{it} = 1 \mid \mathbf{z}_i, \mathbf{x}_{it}, u_i) = \Pr(-\varepsilon_{it} < [\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i])$$
$$= F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i)$$

where $F(.)$ is the cdf of $-\varepsilon_{it}$

Probit model: $F(\,.\,) = \Phi(\,.\,) \Rightarrow$ cdf of the N(0,1) distribution

Logit model: $F(s) = e^s/[1+e^s] \Rightarrow$ cdf of the logistic distribution

Conditional logit

Subsume $\mathbf{z}_i$ in $\mathbf{x}_{it}$ for notational simplicity.

If we try to estimate the u_i using individual-specific dummy variables, there is no simplification analogous to within-group regression.

Moreover, the number of parameters $\rightarrow \infty$ with n, so the MLDV estimator is not consistent.

Log-likelihood for the logit model for individual i conditional on u_i:

$$L(\beta, u_1 \ldots u_n) = \sum_{t=1}^{T_i} y_{it} \ln\left(\frac{1}{1+e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}\right) + \sum_{t=1}^{T_i} (1-y_{it}) \ln\left(\frac{e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}{1+e^{\mathbf{x}_{it}\boldsymbol{\beta}+u_i}}\right)$$

The statistic $\sum_t y_{it}$ is a sufficient statistic for u_i: $\Pr(\mathbf{y}_i \mid \sum_t y_{it})$ does not depend on u_i.

Example $T_i = 2$; $\sum_t y_{it}$ can take values 0, 1, 2. Conditional on $\sum_t y_{it} = 0$, $y_{i1} = y_{i2} = 0$ and, conditional on $\sum_t y_{it} = 2$, $y_{i1} = y_{i2} = 1$ with prob 1. So only cases with $\sum_t y_{it} = 1$ are of interest.

Conditional logit

Probability of the conditioning event:

$$\Pr(\textstyle\sum_t y_{it} = 1) = \Pr(y_{i1} = 1, y_{i2} = 0) + \Pr(y_{i1} = 0, y_{i2} = 1)$$

$$= P_{i1}(1-P_{i2}) + (1-P_{i1})P_{i2}$$

$$= \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}}{\left(1+e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}\right)}\frac{1}{\left(1+e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}\right)} + \frac{1}{\left(1+e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}\right)}\frac{e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}}{\left(1+e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}\right)}$$

$$= \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i} + e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}}{\left(1+e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}\right)\left(1+e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}\right)}$$

E.g. conditional probability of observing 1 then 0:

$$\Pr(y_{i1}=1, y_{i2}=0 \mid y_{i1}+y_{i2}=1) = \frac{\Pr(y_{i1}=1, y_{i2}=0)}{\Pr(y_{i1}+y_{i2}=1)}$$

$$= \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}}{e^{\mathbf{x}_{i1}\boldsymbol{\beta}+u_i}+e^{\mathbf{x}_{i2}\boldsymbol{\beta}+u_i}} = \frac{e^{\mathbf{x}_{i1}\boldsymbol{\beta}}}{e^{\mathbf{x}_{i1}\boldsymbol{\beta}}+e^{\mathbf{x}_{i2}\boldsymbol{\beta}}} = \frac{e^{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}}{1+e^{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}}$$

$\Rightarrow$ u_i is eliminated by conditioning on $\sum_t y_{it}$

Conditional logit (continued)

With $T = 2$, the conditional log-likelihood is:

$$L(\boldsymbol{\beta}) = \sum_{i:\Sigma y=1} d_i(\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta} - \ln\left(1 + e^{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}\right)$$

where $d_i = 1$ if $y_{i1} = 1$, $y_{i2} = 0$ and 0 if $y_{i1} = 0$, $y_{i2} = 1$.

Note that, if $\mathbf{x}_{it}$ contains time-invariant covariates (*i.e.* $\mathbf{z}_i$), these disappear from $(\mathbf{x}_{i1}\text{-}\mathbf{x}_{i2}) \Rightarrow \boldsymbol{\alpha}$ cannot be estimated.

In general, conditional logit only uses data from individuals who experience change in y_{it} over time. This sacrifices sample variation.

A Hausman test can be used to compare conditional logit estimates with random-effects logit which assumes independence between u_i and $(\mathbf{z}_i , \mathbf{X}_i)$

- The same conditioning approach does not work with probit and other functional forms, nor with general dynamic models
- But it can be generalised to:
  - unordered multinomial logit models
  - ordered logit models with more than two outcomes.

Random effects logit/probit

If we want to:

- estimate the coefficients of $\mathbf{z}_i$
- use a non-logistic form
- allow for dynamic adjustment,

then conditional likelihood is not available. The random effects approach is a natural solution.

Consider a dynamic example - a simple model displaying ***state dependence***.

Latent regression:

$$y_{it}^{*} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it-1} + u_i + \varepsilon_{it}$$
$$y_{it} = 1(y_{it}^{*} > 0)$$

Note the initial condition problem: what are the properties of y_{i0}?

Make standard random effects assumptions (including independence of $(\mathbf{z}_i , \mathbf{x}_{it})$ and u_i).

Assume $\Pr(y_{i0} \mid \mathbf{z}_i , \mathbf{X}_i , u_i)$ has a known parametric form.

The random effects likelihood function

Construct a likelihood by sequential conditioning:

$$\Pr(y_{i0} \mid \mathbf{z}_i , \mathbf{X}_i , u_i) = P_{i0}(u_i)$$
$$\Pr(y_{i1} \mid y_{i0} , \mathbf{z}_i , \mathbf{x}_{i1} , u_i) = P_{i1}(y_{i0} , u_i)$$
$$\vdots$$
$$\Pr(y_{iT} \mid y_{iT\text{-}1} , \mathbf{z}_i , \mathbf{x}_{iT} , u_i) = P_{iT}(y_{iT\text{-}1} , u_i)$$

The probabilities P_{it} are of the form:

$$F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it\text{-}1} + u_i) \quad \text{for } y_{it} = 1$$

or $$1 - F(\alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \gamma y_{it\text{-}1} + u_i) \quad \text{for } y_{it} = 0.$$

Likelihood function for individual i, conditional on u_i:

$$L_i(u_i) = P_{i0}(u_i) \prod_{t=1}^{T_i} P_{it}(y_{it-1}, u_i)$$

N.B. alternative treatment of initial conditions: model $u_i \mid y_{i0} , \mathbf{z}_i , \mathbf{X}_i$ rather than $y_{i0} \mid \mathbf{z}_i , \mathbf{X}_i , u_i$ (Wooldridge 2005)

Integrating out the random effects

Marginalise with respect to u_i:

$$L_i = E\left(P_{i0}(u_i) \prod_{t=1}^{T_i} P_{it}(y_{it-1}, u_i) \right)$$

$$= \int_{-\infty}^{\infty} P_{i0}(u) \prod_{t=1}^{T_i} P_{it}(y_{it-1}, u) g(u) du \qquad (1)$$

where $g(u)$ is an assumed density for u (*e.g.* Gaussian: $g(u) = \sigma_u^{-1} \phi(u/\sigma_u)$)

Evaluation of the likelihood function requires the integral in (1) to be approximated numerically by a quadrature algorithm.

This is implemented in Stata, but computing run times can be quite long

Example: Conditional (fixed effects) logit

```
. gen lowpay=w_hr<5

. clogit lowpay age tenure postGCSE2 female cohort, group(pid)

note: multiple positive outcomes within groups encountered.
note: 5607 groups (24447 obs) dropped due to all positive or all negative
outcomes.
Conditional (fixed-effects) logistic regression   Number of obs   =      13957
                                                  LR chi2(2)      =     976.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -4828.6042                       Pseudo R2       =     0.0919
------------------------------------------------------------------------------
      lowpay |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1943886   .0071101   -27.34   0.000    -.2083241    -.180453
      tenure |  -.0526074   .0068704    -7.66   0.000    -.0660732   -.0391417
------------------------------------------------------------------------------
```

Static random effects logit
(NB no initial conditions problem)

```
. xtlogit lowpay age tenure postGCSE2 female cohort

Random-effects logistic regression              Number of obs      =     38404
Group variable (i): pid                         Number of groups   =      7700
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       5.0
                                                               max =        11
                                                Wald chi2(5)       =   1723.72
Log likelihood  = -14248.767                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      lowpay |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1722222   .0066515   -25.89   0.000    -.1852588   -.1591855
      tenure |  -.0601414    .005933   -10.14   0.000    -.0717698    -.048513
   postGCSE2 |  -2.548309   .0975108   -26.13   0.000    -2.739427   -2.357192
      female |    1.98682   .0918006    21.64   0.000     1.806894    2.166746
      cohort |  -.1454163   .0069869   -20.81   0.000    -.1591105   -.1317222
       _cons |   290.0056   13.88017    20.89   0.000      262.801    317.2102
-------------+----------------------------------------------------------------
    /lnsig2u |   2.220962   .0337331                      2.154846    2.287078
-------------+----------------------------------------------------------------
     sigma_u |   3.035818   .0512038                      2.937101    3.137853
         rho |    .736938   .0065395                      .7239217    .7495531
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  1.1e+04 Prob >= chibar2 = 0.000
```

Hausman test comparing fixed & random effects logit

```
. hausman clogit relogit, equations(1:1)

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     clogit      relogit       Difference          S.E.
-------------+----------------------------------------------------------------
         age |   -.1943886    -.1722222       -.0221664        .0025122
      tenure |   -.0526074    -.0601414         .007534        .0034645
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from clogit
         B = inconsistent under Ha, efficient under Ho; obtained from xtlogit

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       79.72
                Prob>chi2 =      0.0000
```

No huge coefficient differences, but highly significant test result