-redprob-
A Stata program for the Heckman estimator of the random effects dynamic probit model

Mark B. Stewart[*]
University of Warwick

January 2006

1 The model

The latent equation for the random effects dynamic probit model to be considered is specified as

$$y_{it}^{*} = \gamma y_{it-1} + x_{it}'\beta + \alpha_i + u_{it} \tag{1}$$

$(i = 1, \ldots, N; t = 2, \ldots, T)$, where y_{it}^{*} is the latent dependent variable, x_{it} is a vector of exogenous explanatory variables, α_i are (unobserved) individual-specific random effects, and the u_{it} are assumed to be distributed $N(0, \sigma_u^2)$. The observed binary outcome variable is defined as:

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^{*} \geq 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

The subscript i indexes individuals and the subscript t indexes time periods. N is taken to be large, but T is typically small and regarded as fixed. Even when the errors u_{it} are assumed serially independent, the composite error term, $v_{it} = \alpha_i + u_{it}$, will be correlated over time due to the individual–specific time–invariant α_i terms. The individual-specific random effects specification adopted implies equi-correlation between the v_{it} in any two (different) time periods:

$$\lambda = Corr(v_{it}, v_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2} \qquad t, s = 2, \ldots, T; t \neq s \tag{3}$$

The standard (uncorrelated) random effects model also assumes α_i uncorrelated with x_{it}. Alternatively, adopting the Mundlak–Chamberlain approach, correlation

[*] **Address for correspondence**: Mark Stewart, Economics Department, University of Warwick, Coventry CV4 7AL, UK. Tel: (44/0)-24-7653-3043. Fax: (44/0)-24-7652-3032.
E-mail: Mark.Stewart@warwick.ac.uk

between α_i and the observed characteristics in the model can be allowed for by assuming a relationship between α and either the time means of the x-variables or a combination of their lags and leads, e.g.: $\alpha_i = \overline{x}_i' a + \zeta_i$, where $\zeta_i \sim iid$ Normal and independent of x_{it} and u_{it} for all i, t. In terms of implementation, this simply has the effect of adding time means or lags and leads to the set of explanatory variables. To simplify notation the original form (1) will be used here with the understanding that these additional terms are subsumed into the x-vector in the case of the correlated random effects model.

Since y is a binary variable, a normalization is required. A convenient one is that $\sigma_u^2 = 1$. Since u_{it} is normally distributed, the transition probability for individual i at time t, given α_i, is given by

$$P\left[y_{it}|x_{it}, y_{it-1}, \alpha_i\right] = \Phi\left[(\gamma y_{it-1} + x_{it}'\beta + \alpha_i)(2y_{it} - 1)\right]. \tag{4}$$

where Φ is the cumulative distribution function of a standard normal.

Estimation of the model requires an assumption about the initial observations, y_{i1}, and in particular about their relationship with the α_i. The assumption giving rise to the simplest form of model for estimation would be to take the initial conditions, y_{i1}, to be exogenous. This would be appropriate if the start of the process coincided with the start of the observation period for each individual, but this is typically not the case. Under this assumption a standard Random Effects Probit program (such as **xtprobit**) can be used, since the likelihood can be decomposed into two independent factors and the joint probability for $t = 2, \ldots, T$ maximized without reference to that for $t = 1$. However, if the initial conditions are correlated with the α_i, as would be expected in most situations, this estimator will be inconsistent and will tend to overestimate γ and hence overstate the extent of state dependence.

The approach to the initial conditions problem proposed by Heckman (1981) involves specifying a linearized approximation to the reduced form equation for the initial value of the latent variable:

$$y_{i1}^* = z_{i1}'\pi + \eta_i \tag{5}$$

($i = 1, \ldots, N$), where z_{i1} is a vector of exogenous instruments (for example pre-sample variables) and includes x_{i1}, and η_i is correlated with α_i, but uncorrelated with u_{it} for $t \geq 2$. Using an orthogonal projection, it can be written as:

$$\eta_i = \theta\alpha_i + u_{i1} \tag{6}$$

($\theta > 0$), with α_i and u_{i1} independent of one another. It is also assumed that u_{i1} satisfies the same distributional assumptions as u_{it} for $t = 2, \ldots, T$. (Any change in error variance will also be captured in θ.) The linearized reduced form for the latent variable for the initial time period is therefore specified as

$$y_{i1}^* = z_{i1}'\pi + \theta\alpha_i + u_{i1} \tag{7}$$

($i = 1, \ldots, N$).

The joint probability of the observed binary sequence for individual i, given α_i, in the Heckman approach is thus:

$$\Phi\left[(z_{i1}'\pi + \theta\alpha_i)(2y_{i1} - 1)\right] \prod_{t=2}^{T} \Phi\left[(\gamma y_{it-1} + x_{it}'\beta + \alpha_i)(2y_{it} - 1)\right]. \tag{8}$$

Hence for a random sample of individuals the likelihood to be maximized is given by

$$\prod_i \int_{\alpha^*} \left\{ \Phi\left[(z_{i1}'\pi + \theta\sigma_\alpha\alpha^*)(2y_{i1}-1)\right] \prod_{t=2}^{T} \Phi\left[(\gamma y_{it-1} + x_{it}'\beta + \sigma_\alpha\alpha^*)(2y_{it}-1)\right] \right\} dF(\alpha^*) \tag{9}$$

where F is the distribution function of $\alpha^* = \alpha/\sigma_\alpha$. Under the normalization used, $\sigma_\alpha = \sqrt{\lambda/(1-\lambda)}$. With α taken to be normally distributed, the integral over α^* can be evaluated using Gaussian–Hermite quadrature. The program **redprob** provides this Maximum Likelihood estimator. See Stewart (2005) for an application of the estimator in the context of an investigation of the dynamics of the conditional probability of unemployment.

2 The redprob command

2.1 Syntax

`redprob` *depvar varlist* (*varlist*$_{init}$) [`if` *exp*] [`in` *range*] [, `i(`*varname*`)` `t(`*varname*`)` `quadrat(`#`)` `from(`*matname*`)`]

The lagged dependent variable must be constructed by the user and must appear as the first variable in *varlist.* It is the user's responsibility to ensure that both this variable and *depvar* are binary 0/1 variables. *varlist* should additionally contain the variables in x. *varlist*$_{init}$ should contain the variables in z.

2.2 Options

`i(`*varname*`)` specifies the variable name that contains the cross-section identifier, corresponding to index i.

`t(`*varname*`)` specifies the variable name that contains the time-series identifier, corresponding to index t.

`quadrat(`#`)` specifies the number of Gaussian–Hermite quadrature points in the evaluation of the integral.

`from(`*matname*`)` specifies a matrix containing starting values for the parameters of the model. Use this option to check that a global maximum has been found. Also use to reduce required number of iterations or to restart a previously halted run. The default uses a pooled probit for $t \geq 2$ and separate probit for the initial period reduced form.

2.3 Sample output

To give an example of the use of the command and the output produced, this section uses the union data (http://www.stata-press.com/data/r9/union.dta) used in [R] **xtprobit** to model the probability of union membership. The data are for US young women and are from the NLSY. A subsample of the dataset is used here: (1) only data from 1978 onwards are used, (2) the data for 1983 are dropped, and (3) only those individuals observed in each of the remaining 6 waves are kept:

```
drop if year<78
drop if year==83
by idcode: gen nwav=_N
```

```
keep if nwav==6
```

This gives a balanced panel with $N = 799$ individuals observed in each of $T = 6$ waves and hence a sample size of $NT = 4,794$. In the example here the observations for 85 and 87 are implicitly treated as if they were for 84 and 86 respectively, which would give 6 waves at regular 2-year intervals.

In addition to the lagged dependent variable, the model used includes `age` (age in current year), `grade` (years of schooling completed), and `south` (1 if resident in south). These variables are contained in x in the specification of the model given above. The vector z additionally contains the variable `not_smsa` (1 if living outside a standard metropolitan statistical area). This variable has a significant negative effect on the probability of union membership in the initial period reduced form, whether estimated as a separate probit or as part of the full model.

The output from the `redprob` program is as follows:

```
. sort idcode year

. by idcode: gen tper = _n

. by idcode: gen Lunion = union[_n-1]
(799 missing values generated)

. redprob union Lunion age grade south (age grade south not_smsa), /*
> */  i(idcode) t(tper) quadrat(24) from(bstart1)

(output deleted)

Random-Effects Dynamic Probit Model                  Number of obs   =       4794
                                                     Wald chi2(4)    =      82.04
Log likelihood = -1860.2152                          Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
union        |
      Lunion |   .6344453    .098323     6.45   0.000     .4417357    .8271548
         age |  -.0285863    .009217    -3.10   0.002    -.0466512   -.0105213
       grade |  -.0539267   .0269226    -2.00   0.045     -.106694   -.0011595
       south |  -.4883257   .1238919    -3.94   0.000    -.7311494    -.245502
       _cons |   .5632897   .4798547     1.17   0.240    -.3772082    1.503788
-------------+----------------------------------------------------------------
rfper1       |
         age |   .0081099   .0238047     0.34   0.733    -.0385464    .0547663
       grade |  -.0064163   .0340848    -0.19   0.851    -.0732214    .0603888
       south |  -.7260671   .1650785    -4.40   0.000    -1.049615   -.4025192
    not_smsa |  -.4152246   .1644004    -2.53   0.012    -.7374435   -.0930057
       _cons |  -.9597118   .8413652    -1.14   0.254    -2.608757    .6893338
-------------+----------------------------------------------------------------
   /logitrho |   .8453732    .163999     5.15   0.000     .5239411    1.166805
     /ltheta |  -.1461108   .1267408    -1.15   0.249    -.3945182    .1022966
-------------+----------------------------------------------------------------
         rho |   .6995957   .0344663    20.30   0.000     .6280689    .7625671
       theta |    .864062   .1095119     7.89   0.000     .6740047    1.107712
------------------------------------------------------------------------------
```

2.4 Comments

It is instructive to compare the estimates from **redprob** with those from using other estimators. Estimation of the corresponding model as a simple pooled probit, i.e. ignoring the cross-correlation between the composite error term in different time periods for the same individual, gives an estimate of γ of 1.88, much larger than that here. However it is important to note that the random effects models and the pooled probit model use different normalizations. The random effects models use a normalization of $\sigma_u^2 = 1$, while the pooled probit estimator uses $\sigma_v^2 = 1$. When comparing them with pooled probit estimates, random effects model estimates therefore need to be multiplied by an estimate of $\sigma_u/\sigma_v = \sqrt{1-\lambda}$.

Use of the Stata **xtprobit** command allows individual-specific effects in the equation, but takes the initial condition to be exogenous. This results in a considerable reduction in the estimate of γ compared with pooled probit, to 1.15. Allowing for the different normalizations, the scaled estimate of the coefficient on lagged union membership is 0.79, less than half the pooled probit estimate.

Using the Heckman estimator in the program **redprob**, which allows for the endogeneity of the initial conditions, results in a further reduction in the estimate of γ compared to the **xtprobit** estimates, from 1.15 to 0.63, a reduction by almost half. The rescaled estimate declines from 0.79 to 0.35. It is therefore less than half the **xtprobit** estimate and less than one fifth the pooled probit estimate.

The estimated effects of the x-variables are all greater (in absolute value) with the **xtprobit** random effects estimator than the pooled probit estimator, and greater again when the endogeneity of the initial conditions is allowed for with the **redprob** estimator.

Exogeneity of the initial conditions in the random effects model can be viewed as resulting from imposing $\theta = 0$ on the model estimated by **redprob**. Testing this hypothesis must allow for the fact that it is on the boundary of the parameter space. Never-the-less it is clear that in the example considered here the exogeneity hypothesis is strongly rejected.

3 References

Heckman, J.J. (1981), "The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process", in C.F. Manski and D. McFadden (eds.). *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

Stewart, M.B. (2005), "The inter-related dynamics of unemployment and low-wage employment", forthcoming *Journal of Applied Econometrics.*