DPB: Dynamic Panel Binary data models in gretl

Riccardo Lucchetti
Università Politecnica delle Marche
r.lucchetti@univpm.it

Claudia Pigini
Università Politecnica delle Marche
pigini@stat.unipg.it

Given the increasing availability of panel datasets, software procedures to estimate non-linear models for binary longitudinal data are becoming essential for microeconometric applications, especially because dynamic binary choice models lend themselves to an interpretation in terms of *true state dependence* (Heckman, 1981a), i.e. an event in the past affects the probability of the same event occurring in the future. These models have been employed in several microeconomic fields: employment, more specifically female labour supply (Heckman and Borjas, 1980; Hyslop, 1999; Carrasco, 2001; Arulampalam, 2002; Stewart, 2007; Keane and Sauer, 2009), health (Contoyannis et al., 2004; Heiss, 2011; Halliday, 2008; Carro and Traferri, 2012), poverty transitions (Cappellari and Jenkins, 2004; Biewen, 2009), unionised workers (Stewart, 2006), product purchase behaviour (Wunder and Riphahn, 2014), welfare participation (Wunder and Riphahn, 2014), remittances (Bettin and Lucchetti, 2012), and credit constraints of both households (Brown et al., 2012; Giarda, 2013) and firms (Pigini et al., 2014).

While static models are relatively mainstream and are supported by most of the statistical and econometric software, dynamic models are more complex to implement and, therefore, estimation routines are not always readily available to the practitioner. Dealing with unobserved heterogeneity in these models considerably complicates their estimation compared to their static counterparts, as they raise several issues in terms of both the modelling and computation.

The main modelling issue lies in the so-called "initial condition problem", that is how the outcome variable relates to the process before the observations started being available. Random-Effects (RE henceforth) approaches "solve" the initial condition problem by modelling the joint distribution of the outcomes for all occasions or the outcomes distributions conditional on the initial value. Historically, the first proposal is due to Heckman (1981b) who, building on the static RE estimator, proposed a model for the joint distribution for the response variable $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{iT}]$, specifying a linearised reduced form equation for the initial observation y_{i1}. If the individual unobserved heterogeneity α_i is assumed to be normally distributed and error terms are assumed to be serially independent, the integral over α_i may be evaluated by means of Gauss–Hermite quadrature (Butler and Moffitt, 1982) and model parameters can be estimated by Maximum Likelihood (ML). Generalisations of Heckman's estimator were proposed by Hyslop (1999), who introduced autoregressive error terms, and Keane and Sauer (2009), who extended Hyslop (1999) to a model with a correlation parameter for the autoregressive error terms and a distinct parameter for the correlation with the initial condition error term. In these cases, multivariate normal integrals need to be evaluated by simulation techniques such as GHK (Geweke, 1989; Hajivassiliou and McFadden, 1998; Keane, 1994) which allows for any arbitrary correlation structure among error terms and unobserved heterogeneity.

Alternatively, Wooldridge (2005) proposed modelling the outcome distribution conditional on the initial value of the response variable and on the history of covariates instead of dealing with the joint distribution of all outcomes.[1] Wooldrigde's estimator employs techniques for dealing with the initial observation problem in such a way that estimation can be carried out through ordinary RE probit routines with the addition of some ad-hoc explanatory variables.

The initial condition problem can be circumvented by employing a Fixed-Effects (FE) approach which allows for consistent estimations of the regression parameters without making distributional assumptions on the unobserved heterogeneity and the logistic distribution hypothesis makes it possible to define a Conditional Maximum Likelihood (CML) estimator. In the dynamic context, however, the FE approach has not become as popular as the RE one in empirical works since it cannot be easily generalised to every time-configuration of the panel and requires strong restrictions to the model specification. Moreover, these models generally require that at least a transition between the states 0 and 1 is observed for the individual to contribute to the likelihood. As a results, the number of usable observations often reduces drastically compared to the sample size, especially if there is a strong persistence in the outcome of interest.

The first proposal of a FE logit model can be found in Chamberlain (1985): estimation relies on conditional inference and, therefore, is rather simple to perform. Exogenous covariates, however, cannot be included and the proposed sufficient statistic for incidental parameters needs to be determined on a case-wise basis according to the time-series length. Honoré and Kyriazidou (2000) extended Chamberlain's formulation in order to include explanatory variables; this approach, however, requires a non-negligible computational effort due to the nonparametric evaluation of covariates distribution. In addition, time-dummies have to be excluded form the model specification. Recently, Bartolucci and Nigro (2010) defined a dynamic model which belongs to the quadratic exponential family and it has a similar formulation to that of a dynamic logit model. The sufficient statistics for unobserved heterogeneity parameters are the total scores for every time length of the panel series and CML estimation can be implemented in software by a suitable modification of ordinary static FE logit algorithms. Differently form Honoré and Kyriazidou (2000), time-dummies can be included in the model specification. [2]

In this work we present the `gretl` implementation of the available set of tools to estimate dynamic models for binary panel data by both fixed- and random-effects approaches, collected in the `DPB` function package. The random-effects models contained in `DPB` are the dynamic probit with linearised initial condition proposed in Heckman (1981b) and the generalisations by Hyslop (1999) and Keane and Sauer (2009). Compared to the available estimators based on a RE approach, Heckman's estimator is hardly biased in small samples (Miranda, 2007) and widely used in microeconomic applications. Nevertheless, we also show how to implement in `gretl` the estimator proposed by Wooldridge (2005) by means of the available routine to estimate static RE probit and suitable panel data functions. `DPB` also

[1] Similar approaches have been proposed by Orme (1997, 2001) and Arulampalam and Stewart (2009) which we refer to for a more detailed discussion.

[2] Also, there are estimators based on the FE approach for long panels ($T \to \infty$) such as Hahn and Newey (2004), Carro (2007), Fernández-Val (2009) Hahn and Kuersteiner (2011), Bartolucci et al. (2014).

contains the software for the estimation of the quadratic exponential model in Bartolucci and Nigro (2010), which has certain advantages compared to Honoré and Kyriazidou (2000), namely no restrictions on time-varying covariates are needed. We provide a detailed illustration of the features of `DPB` and a thorough discussion of implementation and computational issues. Finally, we provide an empirical application based on a dataset of unionised workers extracted from the Panel Survey of Income Dynamics.

References

Arulampalam, W. (2002). State dependence in unemployment incidence: evidence for british men revisited. Technical report, IZA Discussion paper series.

Arulampalam, W. and Stewart, M. B. (2009). Simplified implementation of the heckman estimator of the dynamic probit model and a comparison with alternative estimators*. *Oxford bulletin of economics and statistics*, 71(5):659–681.

Bartolucci, F., Bellio, R., Sartori, N., and Salvan, A. (2014). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, in press.

Bartolucci, F. and Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, 78:719–733.

Bettin, G. and Lucchetti, R. (2012). Intertemporal remittance behaviour by immigrants in germany. Working Papers 385, Universita' Politecnica delle Marche (I), Dipartimento di Scienze Economiche e Sociali.

Biewen, M. (2009). Measuring state dependence in individual poverty histories when there is feedback to employment status and household composition. *Journal of Applied Econometrics*, 24(7):1095–1116.

Brown, S., Ghosh, P., and Taylor, K. (2012). The existence and persistence of household financial hardship. Working papers, The University of Sheffield, Department of Economics.

Butler, J. S. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, 50(3):761–64.

Cappellari, L. and Jenkins, S. P. (2004). Modelling low income transitions. *Journal of Applied Econometrics*, 19(5):593–610.

Carrasco, R. (2001). Binary choice with binary endogenous regressors in panel data. *Journal of business & economic statistics*, 19(4).

Carro, J. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140:503–528.

Carro, J. M. and Traferri, A. (2012). State dependence and heterogeneity in health using a bias-corrected fixed-effects estimator. *Journal of Applied Econometrics.*

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In Heckman, J. J. and Singer, B., editors, *Longitudinal analysis of labor market data.* Cambridge University Press: Cambridge.

Contoyannis, P., Jones, A. M., and Rice, N. (2004). Simulation-based inference in dynamic panel probit models: an application to health. *Empirical Economics*, 29(1):49–77.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1):71–85.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):1317–1339.

Giarda, E. (2013). Persistency of financial distress amongst italian households: Evidence from dynamic models for binary panel data. *Journal of Banking & Finance*, 37(9):3425 – 3434.

Hahn, J. and Kuersteiner, G. (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, 27(06):1152–1191.

Hahn, J. and Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72:1295–1319.

Hajivassiliou, V. A. and McFadden, D. L. (1998). The method of simulated scores for the estimation of ldv models. *Econometrica*, 66(4):863–896.

Halliday, T. J. (2008). Heterogeneity, state dependence and health. *The Econometrics Journal*, 11(3):499–516.

Heckman, J. J. (1981a). Heterogeneity and state dependence. *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press: Cambridge MA. Manski CF, McFadden (eds).

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

Heckman, J. J. and Borjas, G. J. (1980). Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica*, 47(187):pp. 247–283.

Heiss, F. (2011). Dynamics of self-rated health and selective mortality. *Empirical economics*, 40(1):119–140.

Honoré, B. E. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68:839–874.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, 67(6):1255–1294.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, 62:95–116.

Keane, M. P. and Sauer, R. M. (2009). Classification error in dynamic discrete choice models: Implications for female labor supply behavior. *Econometrica*, 77(3):975–991.

Miranda, A. (2007). Dynamic probit models for panel data: A comparison of three methods of estimation. In *United Kingdom Stata Users' Group Meetings*, volume 11.

Orme, C. D. (1997). The initial conditions problem and two-step estimation in discrete panel data models. mimeo, University of Manchester.

Orme, C. D. (2001). Two-step inference in dynamic non-linear panel data models. mimeo, University of Manchester.

Pigini, C., Presbitero, A. F., and Zazzaro, A. (2014). State dependence in access to credit. Mo.Fi.R. Working Papers 102, Money and Finance Research group (Mo.Fi.R.) - Univ. Politecnica Marche - Dept. Economic and Social Sciences.

Stewart, M. (2006). Maximum simulated likelihood estimation of random-effects dynamic probit models with autocorrelated errors. *Stata Journal*, 6(2):256–272.

Stewart, M. B. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of applied econometrics*, 22(3):511–531.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics*, 20(1):39–54.

Wunder, C. and Riphahn, R. T. (2014). The dynamics of welfare entry and exit amongst natives and immigrants. *Oxford Economic Papers*, 66(2):580–604.