Analysis of binary panel data by static and dynamic logit models

Francesco Bartolucci
University of Perugia
bart@stat.unipg.it

Preliminaries

- Longitudinal (or panel) data consist of *repeated observations on the some subjects at different occasions*
- Data of this type are commonly used in many fields, especially in *economics* (e.g. analysis of labor market, analysis of the customer behavior) and in *medicine* (e.g. study of aging, efficacy of a drug)
- Many longitudinal datasets are now available:
  - National Longitudinal Surveys of Labor Market Experience (NLS)
  - Panel Study of Income Dynamics (PSID)
  - European Community Household Panel (ECHP)
  - The Netherlands Socio-Economic Panel (SEP)
  - German Social Economic Panel (GSOEP)
  - British Household Panel Survey (BHPS)

- With respect to cross-sectional data, longitudinal data have the advantage of allowing one to study (or to take into account in a natural way):
  - *unobserved heterogeneity*
  - *dynamic relationships*
  - *causal effects*
- Longitudinal studies suffer from *attrition*
- We will study, in particular, models for the analysis of *binary* response variables

Basic notation

- There are n *subjects* (or individuals) in the sample, with:
  - T_i: number of occasions at which subject i is observed
  - y_{it}: response variable (binary or categorical) for subject i at occasion t
  - $\mathbf{x}_{it}$: vector of covariates for subject i at occasion t
- The dataset is said *balanced* if all subjects are observed at the same occasions ($T_1 = \cdots = T_n$); otherwise, it is said *unbalanced*
- Usually, the dataset is unbalanced because of *attrition*; particular care is needed in this case, especially when the non-responses are not ignorable
- For simplicity, we will usually refer to the *balanced case* and we will denote by T the number of occasions (common to all subjects)

Example (similar to Hyslop, 1999)

- We consider a sample of $n = 1908$ women, aged 19 to 59 in 1980, who were followed from 1979 to 1985 (source *PSID*)
- *Response variable*: y_{it} equal to 1 if woman i has a job position during year t and to 0 otherwise
- *Covariates:*
  - age in 1980 (time-constant)
  - race (dummy equal to 1 for a black; time-constant)
  - educational level (number of year of schooling; time-constant)
  - number of children aged 0 to 2 (time-varying), aged 3 to 5 (time-varying) and aged 6 to 17 (time-varying)
  - permanent income (average income of the husband from 1980 to 1985; time-constant)
  - temporary income (difference between income of the husband in a year and permanent income; time-varying)

Homogeneous static logit and probit models

- These are *simple models* for the probabilities

$$\pi(\mathbf{x}_{it}) = p(y_{it} = 1 \mid \mathbf{x}_{it})$$

- These probabilities are modeled so that they always belong to [0,1]; this is obtained by a *link function* of type logit or probit:

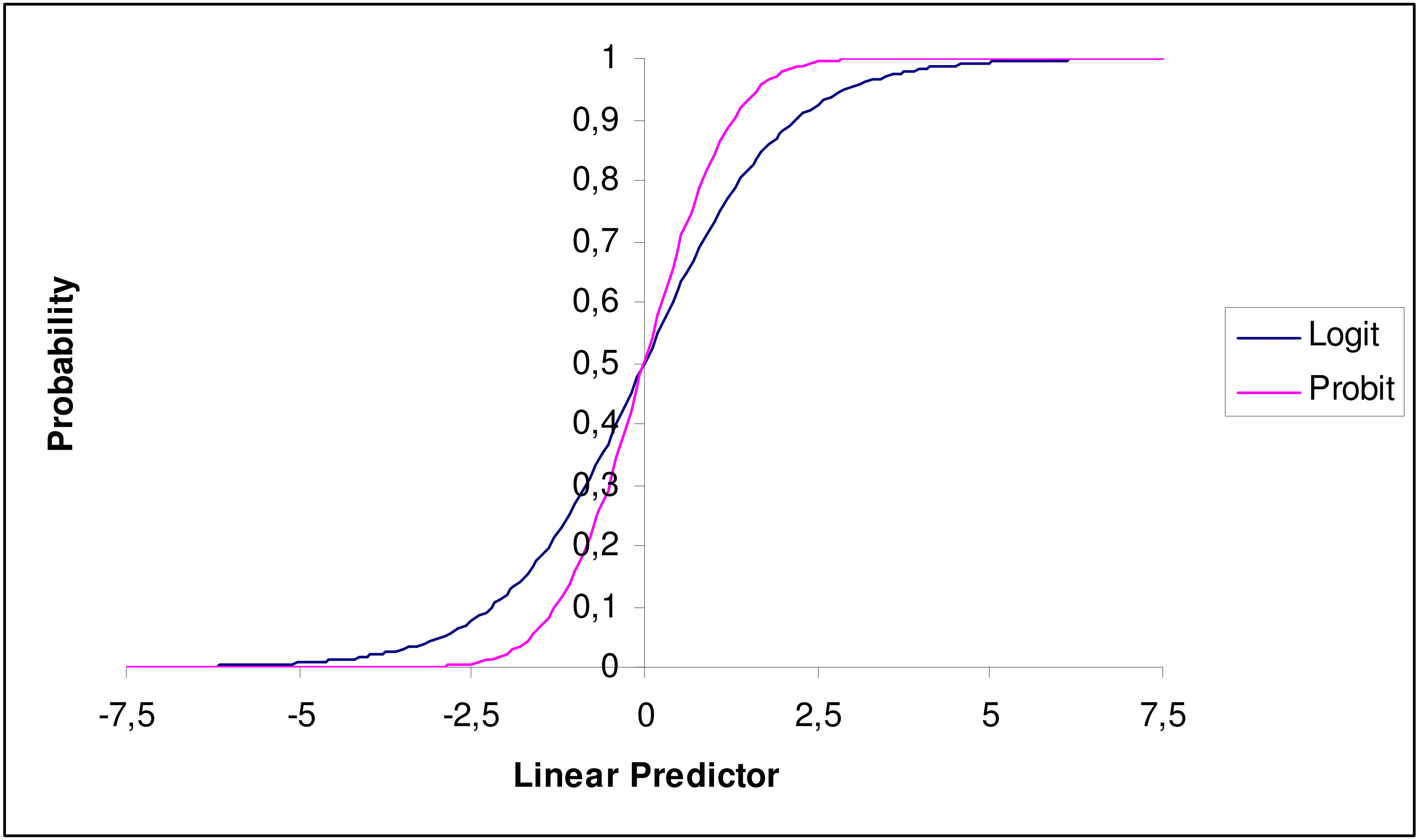
    - logit: $$\log \frac{\pi(\mathbf{x}_{it})}{1-\pi(\mathbf{x}_{it})} = \mathbf{x}_{it}'\boldsymbol{\beta}$$

    - probit: $$\Phi^{-1}[\pi(\mathbf{x}_{it})] = \mathbf{x}_{it}'\boldsymbol{\beta}$$

    - $\Phi^{-1}(\cdot)$: inverse of the distribution function of the standard normal distribution

- The *inverse link function* is:
  - logit: $$\pi(\mathbf{x}_{it}) = \frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta})}$$
  - probit: $$\pi(\mathbf{x}_{it}) = \Phi(\mathbf{x}_{it}'\boldsymbol{\beta})$$
  - $\Phi(\cdot)$: distribution function of the standard normal distribution

- Other *basic assumptions* of the models:
  - Independence between the response variables given the covariates (*static models*)
  - The heterogeneity between subjects is only explained on the basis of observable covariates and then unobserved heterogeneity is ruled out (*homogeneous models*)


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Probability
-7,5
-5
-2,5
0
2,5
5
7,5
Linear Predictor
Logit
Probit

Threshold model

- Logit and probit models may be interpreted on the basis of an *underlying linear model* for the propensity to experience a certain situation:

$$y_{it}^{*} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

  - ε_{it}: error term with standard normal or logistic distribution

- The situation is experienced ($y_{it} = 1$) only if $y_{it}^{*} \geq 0$ (*threshold*), i.e.

$$y_{it} = 1(y_{it}^{*} \geq 0) = \begin{cases} 1 & \text{if } y_{it}^{*} \geq 0 \\ 0 & \text{if } y_{it}^{*} < 0 \end{cases}$$

- Since the distribution of ε_{it} is symmetric, we have that

$$p(y_{it} = 1 \mid \mathbf{x}_{it}) = p(y_{it}^{*} \geq 0 \mid \mathbf{x}_{it}) = p(\mathbf{x}_{it}'\boldsymbol{\beta} \geq -\varepsilon_{it} \mid \mathbf{x}_{it}) = p(\varepsilon_{it} \leq \mathbf{x}_{it}'\boldsymbol{\beta} \mid \mathbf{x}_{it})$$

corresponding to the *logistic or standard normal distr. function*

Model estimation

- The most used method to fit logit and probit models is the *maximum likelihood method*, which is based on the maximization of the log-likelihood:

$$L(\boldsymbol{\beta}) = \sum_i \sum_t y_{it} \log[\pi(\mathbf{x}_{it})] + (1 - y_{it}) \log[1 - \pi(\mathbf{x}_{it})]$$

- Maximization of $L(\boldsymbol{\beta})$ can be performed by the *Newton-Raphson algorithm*. Starting from an initial estimate $\boldsymbol{\beta}^{(0)}$, the algorithm consists of updating the estimate at step h as

$$\boldsymbol{\beta}^{(h)} = \boldsymbol{\beta}^{(h-1)} + \mathbf{J}\left(\boldsymbol{\beta}^{(h-1)}\right)^{-1} \mathbf{s}\left(\boldsymbol{\beta}^{(h-1)}\right)$$

  - $\mathbf{s}(\boldsymbol{\beta}) = \dfrac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$: *score vector*
  - $\mathbf{J}(\boldsymbol{\beta}) = -\dfrac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$: *observed information matrix*

- An alternative algorithm is the *Fisher-scoring* which uses the *expected information matrix*

$$\mathbf{I}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)$$

  instead of the observed information matrix

- *Standard errors* for each element of $\hat{\boldsymbol{\beta}}$ is computed as the square root of the corresponding diagonal element of $\mathbf{I}(\boldsymbol{\beta})^{-1}$

- For the *logit model* we have

$$L(\boldsymbol{\beta}) = \sum_i \sum_t y_{it}\mathbf{x}_{it}'\boldsymbol{\beta} - \log[1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta})]$$

$$\mathbf{s}(\boldsymbol{\beta}) = \sum_i \sum_t [y_{it} - \pi(\mathbf{x}_{it})]\mathbf{x}_{it} ,$$

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbf{I}(\boldsymbol{\beta}) = \sum_i \sum_t \pi(\mathbf{x}_{it})[1 - \pi(\mathbf{x}_{it})]\mathbf{x}_{it}\mathbf{x}_{it}'$$

Example

- Maximum likelihood estimates for the PSID dataset (*logit model*)

Parameter	Estimate	*s.e.*	*t*-statistic	*p*-value
Intercept	-0.6329	0.3093	-2.0464	0.0407
Age	0.0923	0.0172	5.3750	0.0000
Age^2/100	-0.1694	0.0221	-7.6496	0.0000
Race	0.3161	0.0517	6.1188	0.0000
Education	0.3278	0.0152	21.5510	0.0000
Kids 0-2	-0.7810	0.0447	-17.4890	0.0000
Kids 3-5	-0.6450	0.0406	-15.8920	0.0000
Kids 6-17	-0.1400	0.0201	-6.9497	0.0000
Perm. inc.	-0.0215	0.0014	-15.5820	0.0000
Temp. inc.	-0.0070	0.0023	-2.9860	0.0028

- Maximum likelihood estimates for the PSID dataset (*probit model*)

Parameter	**Estimate**	***s.e.***	***t*-statistic**	***p*-value**
Intercept	-0.3770	0.1843	-2.0451	0.0408
Age	0.0548	0.0103	5.3308	0.0000
Age^2/100	-0.1009	0.0133	-7.5893	0.0000
Race	0.1990	0.0304	6.5362	0.0000
Education	0.1921	0.0089	21.6380	0.0000
Kids 0-2	-0.4666	0.0266	-17.5360	0.0000
Kids 3-5	-0.3871	0.0242	-15.9790	0.0000
Kids 6-17	-0.0846	0.0120	-7.0353	0.0000
Perm. inc.	-0.0115	0.0008	-14.3010	0.0000
Temp. inc.	-0.0027	0.0013	-2.0524	0.0401

- By a general rule the estimate of $\boldsymbol{\beta}$ under the logit model is approx. equal to 1.6 times the estimate of $\boldsymbol{\beta}$ under the probit model

Heterogeneous static logit and probit models

- A method to incorporate unobserved heterogeneity in a logit or probit model is to include a set of *subject-specific parameters* α_i and then assuming that

$$\pi(\alpha_i, \mathbf{x}_{it}) = \frac{\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})}{1+\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})} \quad \text{or} \quad \pi(\alpha_i, \mathbf{x}_{it}) = \Phi(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})$$

  - $\pi(\alpha_i, \mathbf{x}_{it}) = p(y_{it} = 1 | \alpha_i, \mathbf{x}_{it})$: conditional probability of success given α_i and $\mathbf{x}_{it}$

- The parameters α_i may be treated as fixed or random:
  - *fixed*: the response variables y_{it} are still assumed independent
  - *random*: the response variables y_{it} are assumed conditionally independent given α_i

- The most used *estimation methods* of the model are:
  - joint maximum likelihood (fixed-parameters)
  - conditional maximum likelihood (only for the logit model)
  - marginal maximum likelihood (random-parameters)

Joint maximum likelihood (JML) method

- It consists of maximizing the log-likelihood

$$L(\boldsymbol{\alpha},\boldsymbol{\beta})=\sum_i\sum_t y_{it}\log[\pi(\alpha_i,\mathbf{x}_{it})]+(1-y_{it})\log[1-\pi(\alpha_i,\mathbf{x}_{it})]$$

  with respect to (*jointly*) $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)'$ and $\boldsymbol{\beta}$

- The method is simple to implement for both logit and probit models

- It is usually based on an iterative algorithm which alternates Newton-Raphson (or Fisher scoring) steps for updating the estimate of each α_i with Newton-Raphson (or Fisher scoring) steps for updating the estimate of $\boldsymbol{\beta}$

- The JML estimator:
  - *does not exist* (for α_i) when $y_{i+} = 0$ or $y_{i+} = T$, with $y_{i+} = \sum_t y_{it}$
  - is not consistent with T fixed as n grows to infinity and so a JML estimate is not reliable for small T even if n is very large; this is because the number of parameters increases with n (*incidental parameters problem*; Neyman and Scott, 1948)
- For the heterogeneous logit model we must *solve the equations*:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \alpha_i} = \sum_t [y_{it} - \pi(\alpha_i, \mathbf{x}_{it})] = 0, \qquad i = 1, \ldots, n$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \sum_t [y_{it} - \pi(\alpha_i, \mathbf{x}_{it})]\mathbf{x}_{it} = \mathbf{0}$$

Conditional maximum likelihood (CML) method

- This estimation method may be used only for the *logit model*

- For the logit model we have that, for $i=1,\dots,n$, y_{i+} is a *sufficient statistic* for the subject specific-parameter α_i and, consequently, we can construct a conditional likelihood which does not depend on these parameters but only on $\boldsymbol{\beta}$

- The *conditional log-likelihood* may be expressed as

$$L_c(\boldsymbol{\beta}) = \sum_i \log[p(\mathbf{y}_i \mid \mathbf{X}_i, y_{i+})], \qquad \mathbf{y}_i = (y_{i1},\dots,y_{iT})'$$

- From the maximization of $L_c(\boldsymbol{\beta})$ we obtain the *CML estimator* of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_c$, which is consistent for fixed T as n grows to infinity; this maximization may be performed on the basis of a *Newton-Raphson* algorithm which also produces standard errors for $\hat{\boldsymbol{\beta}}_c$

- An important *drawback*, common to all fixed-parameters approaches, is that the regression parameters for the time-constant covariates are not estimable

- The *probability of the response configuration* $\mathbf{y}_i$ may be expressed as

$$p(\mathbf{y}_i \mid \mathbf{X}_i) = \prod_t \frac{\exp[y_{it}(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]}{1+\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})} = \frac{\exp(y_{i+}\alpha_i + \sum_t y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\prod_t [1+\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]}$$

- The *probability of the sum of the responses* y_{i+} is then equal to

$$p(y_{i+} \mid \mathbf{X}_i) = \frac{\exp(y_{i+}\alpha_i)}{\prod_t [1+\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta})]} \sum_{\mathbf{z}(y_{i+})} \exp(\textstyle\sum_t z_t \mathbf{x}_{it}'\boldsymbol{\beta})$$

  - $\sum_{\mathbf{z}(y_{i+})}$ is extended to all the response configurations $\mathbf{z} = (z_1, \ldots, z_T)'$ with sum $z_+ = y_{i+}$

- The *conditional probability of the response configuration* $\mathbf{y}_i$ given y_{i+} is then

$$p(\mathbf{y}_i \mid \mathbf{X}_i, y_{i+}) = \frac{\exp(\sum_t y_{it}\mathbf{x}_{it}'\boldsymbol{\beta})}{\sum_{\mathbf{z}(y_{i+})}\exp(\sum_t z_t \mathbf{x}_{it}'\boldsymbol{\beta})}$$

which is equal to 1 for $y_{i+} = 0$ or $y_{i+} = T$ regardless of the value of $\boldsymbol{\beta}$

- The *conditional log-likelihood* is equal to

$$L_c(\boldsymbol{\beta}) = \sum_i 1(0 < y_{i+} < T)\left\{\sum_t y_{it}\mathbf{x}_{it}'\boldsymbol{\beta} - \log\left[\sum_{\mathbf{z}(y_{i+})}\exp(\sum_t z_t \mathbf{x}_{it}'\boldsymbol{\beta})\right]\right\}$$

- *Score* and observed *information matrix*, to be used within the Newton-Raphson algorithm and to compute the standard errors for $\hat{\boldsymbol{\beta}}_c$:

$$\mathbf{s}_c(\boldsymbol{\beta}) = \frac{\partial L_c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i 1(0 < y_{i+} < T)\mathbf{X}_i'\left[\mathbf{y}_i - \mathrm{E}_{\boldsymbol{\beta}}(\mathbf{y}_i \mid y_{i+})\right]$$

$$\mathbf{J}_c(\boldsymbol{\beta}) = -\frac{\partial L_c^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \sum_i 1(0 < y_{i+} < T)\mathbf{X}_i'\mathrm{V}_{\boldsymbol{\beta}}(\mathbf{y}_i \mid y_{i+})\mathbf{X}_i$$

- JML and CML estimates for the PSID dataset (*logit model*)

Parameter	**JML estimate**	**CML estimate**	***s.e.***	***t-statistic***	***p*-value**
Intercept	-	-	-	-	-
Age	-	-	-	-	-
Age^2/100	-	-	-	-	-
Race	-	-	-	-	-
Education	-	-	-	-	-
Kids 0-2	-1.3660	-1.1537	0.0899	-12.8290	0.0000
Kids 3-5	-0.9912	-0.8373	0.0840	-9.9638	0.0000
Kids 6-17	-0.2096	-0.1764	0.0637	-2.7691	0.0056
Perm. inc.	-	-	-	-	-
Temp. inc.	-0.0162	-0.0136	0.0033	-4.1186	0.0000

(-) not estimable

Marginal maximum likelihood (MML)

- This estimation method may be used for both *logit and probit models*

- It is based on the assumption that the subject-specific parameters α_i are *random parameters* with the same distribution $f(\alpha_i)$ which is independent of $\mathbf{X}_i$

- It is also assumed that the response variables $y_{i1},\ldots,y_{iT}$ are *conditionally independent* given α_i, so that

$$p(\mathbf{y}_i \mid \mathbf{X}_i) = \int p(\mathbf{y}_i \mid \alpha_i, \mathbf{X}_i) f(\alpha_i) d\alpha_i, \quad p(\mathbf{y}_i \mid \alpha_i, \mathbf{X}_i) = \prod_t p(y_{it} \mid \alpha_i, \mathbf{x}_{it}),$$

  where the integral must usually be computed by a numerical method (e.g. quadrature)

- The *marginal log-likelihood* is then

$$L_m(\boldsymbol{\beta}) = \sum_i \log[p(\mathbf{y}_i \mid \mathbf{X}_i)],$$

which can be maximized, with respect to $\boldsymbol{\beta}$ and (possibly) the parameters of the distribution of the random effects, by a Newton-Raphson algorithm

Logit model with normal random effects

- Under the assumption $\alpha_i \sim N(\mu, \sigma^2)$, for the *logit model* we have

$$p(\mathbf{y}_i \mid \mathbf{X}_i) = \int p(\mathbf{y}_i \mid w, \mathbf{X}_i)\phi(w)dw$$

  - $p(\mathbf{y}_i \mid w, \mathbf{X}_i) = \prod_t \frac{\exp[y_{it}(\mu + w\sigma + \mathbf{x}_{it}'\boldsymbol{\beta})]}{1+\exp(\mu + w\sigma + \mathbf{x}_{it}'\boldsymbol{\beta})} = \prod_t \frac{\exp\{y_{it}[\mathbf{z}_{it}(w)'\boldsymbol{\gamma}]\}}{1+\exp[\mathbf{z}_{it}(w)'\boldsymbol{\gamma}]}$
  - $\phi(w)$: density function of the standard normal distribution
  - $\mathbf{z}_{it}(w) = (1 \quad w \quad \mathbf{x}_{it}')', \qquad \boldsymbol{\gamma} = (\mu \quad \sigma \quad \boldsymbol{\beta}')'$

- The *score vector* and the (empirical) *information matrix* are given by

$$\mathbf{s}_m(\boldsymbol{\gamma}) = \frac{\partial L_m(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \sum_i \mathbf{s}_{m,i}(\boldsymbol{\gamma}), \qquad \mathbf{s}_{m,i}(\boldsymbol{\gamma}) = \frac{1}{p(\mathbf{y}_i \mid \mathbf{X}_i)} \int \frac{\partial p(\mathbf{y}_i \mid w, \mathbf{X}_i)}{\partial \boldsymbol{\gamma}} \phi(w)dw$$

$$\tilde{\mathbf{J}}_m(\boldsymbol{\gamma}) = \sum_i \mathbf{s}_{m,i}(\boldsymbol{\gamma})\mathbf{s}_{m,i}(\boldsymbol{\gamma})' - \frac{1}{n}\mathbf{s}_m(\boldsymbol{\gamma})\mathbf{s}_m(\boldsymbol{\gamma})'$$

Pros and cons of MML

- The MML method is more complicate to implement than fixed-effects methods (JML, CML), but it allows us to estimate the regression parameters for *both time-fixed and time-varying covariates*

- The MML also allows us to *predict future outcomes*

- Special care has to be used for the *specification of the distribution of the random effects*. It may be restrictive to assume:
  - a specific parametric function for these effects, such as the normal distribution
  - that the distribution does not depend on the covariates

- The approach may be extended to overcome these drawbacks:
  - a discrete distribution with free support points and mass probabilities may be used for the random effects; the approach is in this case of *latent class* type and requires the implementation of an EM algorithm (Dempster *et al.*, 1977) and the choice of the number of support points
  - the parameters of the distribution of the random effects are allowed to depend on the covariates; one possibility is the *correlated effect model* of Chamberlain (1984)

- JML, CML and MML-normal estimates for the PSID dataset (*logit model*); MML algorithm uses 51 quadrature points from –5 to 5

Parameter	**JML estimate**	**CML estimate**	**MML estimate**	***s.e.***	***t*-*statistic***	***p*-value**
Intercept	-	-	-2.9448	1.3461	-2.1876	0.0287
Std.dev (σ)			3.2196	0.1066	30.2090	0.0000
Age	-	-	0.2652	0.0712	3.7243	0.0002
Age^2/100	-	-	-0.4285	0.0906	-4.7271	0.0000
Race	-	-	0.6800	0.2162	3.1449	0.0017
Education	-	-	0.6737	0.0643	10.4810	0.0000
Kids 0-2	-1.3660	-1.1537	-1.3418	0.0773	-17.3490	0.0000
Kids 3-5	-0.9912	-0.8373	-1.0260	0.0635	-16.1680	0.0000
Kids 6-17	-0.2096	-0.1764	-0.2533	0.0438	-5.7775	0.0000
Perm. inc.	-	-	-0.0427	0.0036	-11.9610	0.0000
Temp. inc.	-0.0162	-0.0136	-0.0110	0.0023	-4.7554	0.0000

Summary of the models fit

- Estimates for the PSID dataset (*logit model*):

Method	Log-likelihood	n. parameters	AIC	BIC
Homogenous	-7507.3	10	15034.6	15090.1
Heterogeneous-JML	-2986.3	1912	9796.6	20415.5
Heterogeneous-CML*	-2128.5	4	4265.0	4287.2
Heterogeneous-MML-normal	-5264.4	11	10550.8	10611.9

(*) not directly comparable with the others

- AIC: *Akaike Information Criterion* (Akaike, 1973)

$$\mathrm{AIC} = -2(\mathrm{max.\,log-likelihood}) + 2(\mathrm{n.\,parameters})$$

- BIC: *Bayesian Information Criterion* (Schwarz, 1978)

$$\mathrm{BIC} = -2(\mathrm{max.\,log-likelihood}) + \log(n)(\mathrm{n.\,parameters})$$

Dynamic models

- Previous models are *static*: they do not include the lagged response variable among the regressors

- The *dynamic* version of these models is based on the assumption that, given $y_{i,t-1}$ and α_i, every y_{it} is conditionally independent of $y_{i1},\dots,y_{i,t-2}$ and that

$$\pi(\alpha_i, \mathbf{x}_{it}, y_{i,t-1}) = \frac{\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)}{1+\exp(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)}$$

or

$$\pi(\alpha_i, \mathbf{x}_{it}, y_{i,t-1}) = \Phi(\alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)$$

➢ $\pi(\alpha_i, \mathbf{x}_{it}, y_{i,t-1}) = p(y_{it} = 1 \mid \alpha_i, \mathbf{x}_{it}, y_{i,t-1})$: conditional probability of success

- The initial observation y_{i0} must be known. When the parameters α_i are random, the *initial condition problem* arises. The simplest approach, which however can lead to an biased estimator of $\boldsymbol{\beta}$ and γ, is to treat y_{i0} as an exogenous covariate

- Dynamic models have the great advantage of allowing us to distinguish between:
  - *true state dependence* (Heckman, 1981): effect that experimenting a certain situation in the present has on the propensity of experimenting the same situation in the future
  - *spurious state dependence*: propensity common to all occasions which is measured by α_i and the time-constant covariates

Estimation of dynamic models

- The subject-specific parameters α_i may be considered as *fixed* or *random*

- With *fixed parameters* α_i, the conditional probability of a response configuration $\mathbf{y}_i$ given y_{i0} is:

$$p(\mathbf{y}_i \mid \alpha_i, \mathbf{X}_i, y_{i0}) = \prod_t p(y_{it} \mid \alpha_i, \mathbf{x}_{it}, y_{i,t-1})$$

- The *random-parameters* approach requires to formulate a distribution for the parameters α_i, so that

$$p(\mathbf{y}_i \mid \mathbf{X}_i, y_{i0}) = \int p(\mathbf{y}_i \mid \alpha_i, \mathbf{X}_i, y_{i0}) f(\alpha_i) d\alpha_i,$$

$$p(\mathbf{y}_i \mid \alpha_i, \mathbf{X}_i, y_{i0}) = \prod_t p(y_{it} \mid \alpha_i, \mathbf{x}_{it}, y_{i,t-1})$$

- The most used *estimation methods* for dynamic models are the same as for static models:
  - joint maximum likelihood (fixed-parameters)
  - conditional maximum likelihood (only for the logit model)
  - marginal maximum likelihood (random-parameters)

Joint maximum likelihood (JML) method

- The *log-likelihood* has again a simple form:

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)=\sum_i\sum_t y_{it}\log[\pi(\alpha_i,\mathbf{x}_{it},y_{i,t-1})]+(1-y_{it})\log[1-\pi(\alpha_i,\mathbf{x}_{it},y_{i,t-1})]$$

  and must be jointly maximized with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and γ

- Maximizing the *log-likelihood* may be performed by using a Newton-Raphson (or Fisher scoring) algorithm which alternates a step in which the estimate of each parameter α_i is updated with a step in which the estimates of $\boldsymbol{\beta}$ and γ are updated

- The algorithm is essentially the same as that used for static models, but with $y_{i,t-1}$ included among the covariates $\mathbf{x}_{it}$

- The JML estimator has the same drawbacks it has for static models:
  - it *does not exist* (for α_i) when $y_{i+}=0$ or $y_{i+}=T$, with $y_{i+}=\sum_t y_{it}$
  - it is *not consistent* with T fixed as n grows to infinity

- JML estimates for the PSID dataset (*static and dynamic logit models*)

Parameter	Static logit	Dynamic logit
Kids 0-2	-1.3660	-1.2688
Kids 3-5	-0.9912	-0.8227
Kids 6-17	-0.2096	-0.1730
Temp. inc.	-0.0162	-0.0112
Lagged response	-	0.5696

- A positive state dependence is observed and the *fit of the logit model* improves considerably by including the lagged response variable

Model	Log-likelihood	n. parameters	AIC	BIC
Static logit	-2986.3	1912	9796.6	20415.5
Dynamic logit	-2317.9	1913	8461.8	19086.2

Conditional maximum likelihood (CML) method

- The CML method may be used to estimate the dynamic logit model *only in particular circumstances*

- Under these circumstances, the method is difficult to implement since the sum of the response variables y_{i+} is not a *sufficient statistic* for the subject specific-parameter α_i

- The CML approach may be used when $T = 3$ and there are *no covariates*, so that

$$p(\mathbf{y}_i \mid \alpha_i, y_{i0}) = \frac{\exp(y_{i+}\alpha_i + y_{i*}\gamma)}{\prod_t [1 + \exp(y_{it}\alpha_i + y_{i,t-1}\gamma)]}, \qquad y_{i*} = \sum_t y_{i,t-1} y_{it}$$

- The response configurations $\mathbf{y}_i = (0 \quad 1 \quad y_{i3})'$ and $\mathbf{y}_i = (1 \quad 0 \quad y_{i3})'$ have *conditional probability*

$$p[(0 \quad 1 \quad y_{i3})' | \alpha_i, y_{i0}] = \frac{\exp[(1+y_{i3})\alpha_i + y_{i3}\gamma]}{[1+\exp(\alpha_i + y_{i0}\gamma)][1+\exp(\alpha_i)][1+\exp(\alpha_i+\gamma)]}$$

$$p[(1 \quad 0 \quad y_{i3})' | \alpha_i, y_{i0}] = \frac{\exp[(1+y_{i3})\alpha_i + y_{i0}\gamma]}{[1+\exp(\alpha_i + y_{i0}\gamma)][1+\exp(\alpha_i+\gamma)][1+\exp(\gamma)]}$$

- We can then condition on $y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}$ obtaining the *conditional probabilities*

$$p[(0 \quad 1 \quad y_{i3}) | \alpha_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}] = \frac{\exp(y_{i3}\gamma)}{\exp(y_{i3}\gamma) + \exp(y_{i0}\gamma)} = \frac{1}{1+\exp[(y_{i0} - y_{i3})\gamma]}$$

$$p[(1 \quad 0 \quad y_{i3}) | \alpha_i, y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}] = \frac{\exp(y_{i0}\gamma)}{\exp(y_{i3}\gamma) + \exp(y_{i0}\gamma)} = \frac{\exp[(y_{i0} - y_{i3})\gamma]}{1+\exp[(y_{i0} - y_{i3})\gamma]}$$

- The corresponding *conditional log-likelihood* is

$$L_c(\gamma) = \sum_i d_i (y_{i1}(y_{i0} - y_{i3})\gamma - \log\{1 + \exp[(y_{i0} - y_{i3})\gamma]\})$$

$$d_i = 1(y_{i1} + y_{i2} = 1),$$

  which may be maximized by a simple Newton-Raphson algorithm; it results a consistent estimator of γ (Chamberlain, 1993)

- The conditional approach may also be implemented for $T > 3$ on the basis of the *pairwise conditional log-likelihood*

$$L_{pc}(\gamma) = \sum_i \sum_{s<t<T} 1(y_{is} + y_{it} = 1)(y_{is}(y_{i,s-1} - y_{i,t+1})\gamma - \log\{1 + \exp[(y_{i,s-1} - y_{i,t+1})\gamma]\})$$

  the resulting estimator has the same properties it has for $T = 3$ and, in particular, it is consistent for T fixed as n grows to infinity

- The conditional approach may also be used in the presence of covariates, provided that:
    - the probability that each *discrete covariate* is time-constant is positive (this rules out the possibility of time dummies)
    - the support of the distribution of the *continuous* covariates satisfies suitable conditions
- The *algorithm* to be implemented in this case is rather complicate and leads to a consistent estimator of $\boldsymbol{\beta}$ and γ which, however, is not $\sqrt{n}$-consistent (Honoré and Kyriazidou, 2000)
- The CML approach has the advantage, over the MML approach, of not requiring to formulate the *distribution of the subject-specific parameters*. It also does not suffer from the *initial condition problem* and y_{i0} may be treated as an exogenous covariate

Marginal maximum likelihood (MML) method

- This estimation method may be used for both *dynamic logit and probit models*

- The algorithm is essentially the same as that for static models, but we have to use an *extended vector of covariates* which includes the lagged response variable

- For the *dynamic logit model with normal random effects* we have to maximize

$$L_m(\tilde{\boldsymbol{\gamma}}) = \sum_i p(\mathbf{y}_i \mid \mathbf{X}_i, y_{i0}), \qquad p(\mathbf{y}_i \mid \mathbf{X}_i, y_{i0}) = \int p(\mathbf{y}_i \mid w, \mathbf{X}_i, y_{i0}) \phi(w) dw,$$

  - $p(\mathbf{y}_i \mid w, \mathbf{X}_i, y_{i0}) = \prod_t \dfrac{\exp[y_{it}(\mu + w\sigma + \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)]}{1 + \exp(\mu + w\sigma + \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma)}$

- MML-normal estimates for the PSID dataset (*static and dynamic logit models*)

Parameter	**Static logit**	**Dynamic logit**	***s.e.***	***t-statistic***	***p*-value**
Intercept	-2.9448	-2.3313	0.6609	-3.5275	0.0004
Std.dev (σ)	3.2196	1.1352	0.0930	12.2060	0.0000
Age	0.2652	0.1037	0.0360	2.8820	0.0040
Age^2/100	-0.4285	-0.1813	0.0464	-3.9096	0.0001
Race	0.6800	0.3011	0.1054	2.8573	0.0043
Education	0.6737	0.3034	0.0332	9.1456	0.0000
Kids 0-2	-1.3418	-0.8832	0.0825	-10.7010	0.0000
Kids 3-5	-1.0260	-0.4390	0.0736	-5.9629	0.0000
Kids 6-17	-0.2533	-0.0819	0.0393	-2.0831	0.0372
Perm. inc.	-0.0427	-0.0189	0.0019	-10.1030	0.0000
Temp. inc.	-0.0110	-0.0036	0.0030	-1.1783	0.2387
Lagged response	-	2.7974	0.0653	42.8420	0.0000

- For the above example, a much *stronger state dependence effect* is observed with the MML method with respect to the JML method ($\hat{\gamma} = 2.7974$ vs. $\hat{\gamma} = 0.5696$)

- The suspect is that with the MML method the parameter γ is overestimated and this is because the assumptions on the distribution of the parameters α_i are *restrictive*

- A simple way to *give more flexibility* to the approach is to allow the mean of the normal distribution assumed on the parameters α_i to depend (through a linear regression model) on the initial observation y_{i0} and the corresponding time-varying covariates $\mathbf{x}_{i0}$

- MML-normal estimates for the PSID dataset (*dynamic and extended dynamic logit models*)

Parameter	**Dynamic logit**	**Exteded dynamic logit**	***s.e.***	***t*-statistic**	***p*-value**
Intercept	-2.3313	-3.4484	0.8942	-3.8566	0.0001
Std.dev (σ)	1.1352	1.6473	0.0900	18.2930	0.0000
Age	0.1037	0.1103	0.0502	2.1970	0.0280
Age^2/100	-0.1813	-0.1902	0.0647	-2.9410	0.0033
Race	0.3011	0.2744	0.1374	1.9971	0.0458
Education	0.3034	0.2864	0.0419	6.8412	0.0000
Kids 0-2	-0.8832	-1.0498	0.0917	-11.4470	0.0000
Kids 3-5	-0.4390	-0.5865	0.0871	-6.7369	0.0000
Kids 6-17	-0.0819	-0.1213	0.0624	-1.9426	0.0521
Perm. inc.	-0.0189	-0.0164	0.0031	-5.3094	0.0000
Temp. inc.	-0.0036	-0.0049	0.0032	-1.5133	0.1302
Lagged response	2.7974	1.8165	0.0824	22.0550	0.0000

- The estimate for the *state dependence effect seems now more reliable* ($\hat{\gamma} = 1.8164$ vs. $\hat{\gamma} = 2.7974$) even if it is strongly positive
- Estimates of the parameters for the mean of the distribution for α_i

Parameter	Estimate	*s.e.*	*t-statistic*	*p*-value
Kids 0-2	0.2669	0.1284	2.0787	0.0376
Kids 3-5	0.2424	0.1221	1.9864	0.0470
Kids 6-17	0.1299	0.0680	1.9102	0.0561
Temp. inc.	0.0116	0.0058	2.0180	0.0436
Initial observation (y_{i0})	2.5915	0.1586	16.3450	0.0000

Model	Log-likelihood	n. parameters	AIC	BIC
JML	-2317.9	1913	8461.8	19086.2
MML	-4188.1	12	8400.2	8466.8
MML extended	-3976.2	17	7986.4	8080.8

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *Second International symposium on information theory*, Petrov, B. N. and Csaki F. (eds), pp. 267-281.

Bartolucci, F. (2006), Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities, *Journal of the Royal Statistical Society, series B,* **68**, pp. 155-178.

Chamberlain, G. (1984), Panel data, in *Handbook of Econometrics, vol. 2*, Z. Griliches and M.D. Intriligator (eds.), Elsevier Science, Amsterdam, pp. 1247-1318.

Chamberlain, G. (1993), Feedback in Panel Data Models, Mimeo, Department of Economics, Harvad University.

Dempster A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, pp. 1-38.

Elliot, D. S., Huizinga, D. and Menard, S. (1989), *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*, Springer-Verlag, New York.

Frees, E. W. (2004), *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*, Cambridge University Press.

Heckman, J. J. (1981a), Heterogeneity and state dependence, in *Structural Analysis of Discrete Data*, McFadden D. L. and Manski C. A, Cambridge, MA, MIT Press, pp. 91-139.

Heckman, J. J. (1981b), The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, in *Structural Analysis of Discrete Data with Econometric Applications*, C.F. Manski and D. McFadden (eds.), MIT Press, Cambridge, USA, 179-195.

Hsiao, C. (2005), *Analysis of Panel Data, 2nd edition*, Cambridge University Press.

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

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

Manski, C.F. (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, **3**, pp. 205-228.

Neyman, J. and Scott, E. (1948), Consistent estimates based on partially consistent observations, *Econometrica*, **16**, pp. 1-32

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, pp. 461-464.