Impact evaluation of job training programs by a latent variable model

Francesco Bartolucci and Fulvia Pennoni

Abstract We introduce a model for categorical panel data which is tailored to the dynamic evaluation of the impact of job training programs. The model may be seen as an extension of the dynamic logit model and, as such, it allows us to disentangle true from spurious state dependence. The unobserved heterogeneity between subjects is taken into account by formulating the conditional distribution of the response variables given a discrete latent variable. For the estimation of the model parameters we use an EM algorithm and we compute standard errors on the basis of the numerical derivative of the score vector of the complete data log-likelihood. The approach is illustrated through the analysis of a dataset containing the work histories of the employees of the private firms of the province of Milan between 2003 and 2005, some of whom attended job training programs supported by the European Social Fund.

Key words: Latent class model; Panel data; State dependence; Unobserved heterogeneity

1 Introduction

We develop an approach to study the effect of job training programs on the type of employment. The approach is used to analyse a longitudinal dataset containing the work histories of a large group of subjects who are resident in the Province of Milan (Italy), which includes 189 towns and municipalities.

Francesco Bartolucci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia
e-mail: bart@stat.unipg.it

Fulvia Pennoni
Dipartimento di Statistica, Università degli Studi di Milano-Bicocca
e-mail: Fulvia.Pennoni@unimib.it

The model we introduce may be seen as an extension of the dynamic logit model; see [5], [4]. As such, it is based on subject-specific intercepts to account for the unobserved heterogeneity between subjects and it includes, among the regressors, the lagged response variable. This allows us to estimate the effect of the true *state dependence* [4], i.e. the actual effect that experiencing a certain situation in the present has on the probability of experiencing the same situation in the future. Differently from more common approaches, we assume that the random intercepts have a discrete distribution, following in this way a formulation similar to that of the latent class model [6]. This formulation avoids to specify any parametric assumption on the distribution of the random intercepts. Among the regressors, we also include a set of dummies for having attended the training program. These dummies are time-specific, so that we can also evaluate whether the program has or not a constant effect during the chosen period of study.

Maximum likelihood estimation of the model parameters is carried out through an Expectation-Maximization (EM) algorithm [3]. On the basis of the score vector of the complete data log-likelihood, which is obtained as a by-product of the EM algorithm, we compute standard errors for the parameter estimates; see also [2].

The paper is organized as follows. In the next section we describe in more detail the dataset mentioned above. In Section 3 we outline the latent variable model and in Section 4 we discuss the main results from the application of this model to the dataset described in Section 2.

2 The dataset

The dataset we analyse is extracted from a database derived from the merge of two administrative archives. The first archive is made by the mandatory announcements of the employers to the public employee service registers (employment offices) operating on the Province of Milan about hiring (new contract) or firing (expired contract). It is then possible to obtain, for every employee working in a private firm, relevant data on his/her employment trajectories, such as the number of events, type and duration of the contract, sector, and qualification. Since 2000, this archive is updated at any change of the job career.

The second archive contains information about the voluntary participants to the courses supported by the European Social Fund which took place in Lombardy between 2000 and the first quarter of 2007. We select, among the programs designed at that time, those aimed at favouring: (*i*) first time employment, (*ii*) return to work, and (*iii*) acquisition of additional skills for young employees. Most participants are young, with an age between 18 and 35. The courses lasted on average less than six months and ranged from broadly oriented to relatively specialized topics, thus having a different case-mix of attendants among workers.

With the data at hand, we choose to evaluate the impact of job-training programs on the probability of improving in the type of contractual category. We select three main categories: (*i*) temporary agency, (*ii*) temporary (fixed term), and (*iii*) permanent (open ended) job contract. We also choose to study the impact of those programs taking place in the first quarter of 2004 and to restrict the analysis to Italian employees aged 20 to 35 in 2004. We end up with a group of 370,869 workers: 4,146 trained subjects (1.12%) and 366,723 untrained subjects (98.88%).

Note that from the administrative archives, the employment status of a subjects is not available if he/she is: (*i*) not employed, (*ii*) employed outside the Province of Milan, (*iii*) self-employed, or (*iv*) employed in the public sector or with a coordinated and continued collaboration type of contract. Therefore, for each period of interest we consider a response variable having four levels: 0 if the labour state of the subject is unknown (he/she is not in the archive at this time), 1 if he/she is employed with a temporary agency contract, 2 if he/she is employed with a fixed term contract, 3 if he/she is employed with a permanent contract.

We are interested in estimating the early effects of the training. For this reason, we consider the response variable three months before and six, nine, twelve and fifteen months after the beginning of the program. We then have five response variables for each subjects, which are denoted by $y_1,\dots,y_5$. Table 1 reports the frequency of each category of these variables among trained and untrained subjects. Note that categories 1 to 3 are ordered, with the last one corresponding to the most stable type of contract in the Italian system. Moreover, Table 2 reports descriptive statistics referred to the available covariates.

	Trained					*Untrained*				
Outcome	y_1	y_2	y_3	y_4	y_5	y_1	y_2	y_3	y_4	y_5
0	58.18	44.19	43.63	42.69	41.51	34.16	26.19	21.27	19.12	21.97
1	2.82	3.84	3.50	3.09	3.69	3.75	3.47	4.43	4.28	3.63
2	6.68	9.38	9.31	9.67	9.24	12.14	13.16	14.47	14.62	13.45
3	32.32	42.60	43.56	44.55	45.56	49.94	57.18	59.82	61.99	60.95

Table 1 Frequency (%) of each outcome category for trained and untrained subjects.

		Trained	*Untrained*
Males (%)		48.50	54.87
Age in 2003 (mean)		27.76	27.95
Level of Education:	missing (%)	26.75	41.87
	none or primary school (%)	0.72	1.19
	middle school (%)	21.23	23.26
	high school (%)	37.65	25.16
	college degree (%)	13.65	8.51
	higher (%)	0.00	0.02

Table 2 Descriptive statistics for the covariates of trained and untrained subjects.

3 The statistical approach

For each subject i in the sample, $i = 1, \dots, n$, we denote by y_{i0} and y_{i1} the labour state observed, respectively, six and three months before the first quarter of 2004 (period of the beginning of the job training program). We also denote by $y_{i2}, \dots, y_{i5}$ the labour state observed, respectively, six, nine, twelve and fifteen months after the first quarter of 2004.

3.1 Model assumptions

Given the nature of the response variables, we use a model based on nested logits; see [1]. For each variable we have three logits. The first one compares the probability of entering the database against not entering, i.e. category 0 against all the other categories. At nested level, we use two cumulative logits for modelling the conditional probability of each category larger than 0, because these categories are ordered.

The model accounts for unobserved heterogeneity and state dependence by the inclusion of subject-specific intercepts and the lagged response variable among the regressors. The intercepts are treated as random parameters having a discrete distribution with k support points, which identify k latent classes in the population. The model considers the first response variable y_{i0} as given, whereas the distribution of y_{i1} is modelled as follows

$$\log\frac{p(y_{i1}>0|c_i,\mathbf{x}_{i1},y_{i0})}{p(y_{i1}=0|c_i,\mathbf{x}_{i1},y_{i0})} = \alpha_{1c_i} + \mathbf{x}_{i1}'\boldsymbol{\beta}_{11} + \sum_{j=1}^{3} d_{ij0}\beta_{1,j+1},$$

$$\log\frac{p(y_{i1}>1|c_i,\mathbf{x}_{i1},y_{i0},y_{i1}>0)}{p(y_{i1}\leq 1|c_i,\mathbf{x}_{i1},y_{i0},y_{i1}>0)} = \alpha_{2c_i} + \mathbf{x}_{i1}'\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij0}\beta_{2,j+1},$$

$$\log\frac{p(y_{i1}>2|c_i,\mathbf{x}_{i1},y_{i0},y_{i1}>0)}{p(y_{i1}\leq 2|c_i,\mathbf{x}_{i1},y_{i0},y_{i1}>0)} = \alpha_{2c_i} + \tau + \mathbf{x}_{i1}'\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij0}\beta_{2,j+1},$$

where $\mathbf{x}_{i1}$ is the vector of exogenous covariates at the first occasion, and c_i is the latent class of subject i. Moreover, α_{1c} and α_{2c} are the support points associated to latent class c, $c = 1, \dots, k$, τ is the shift parameter for the third logit with respect to the second, and d_{ijt} is a dummy variable equal to 1 if $y_{it} = j$ and to 0 otherwise. The probability of each latent class c will be denoted by π_c.

For what concerns the distribution of y_{it}, $t = 2, \dots, 5$, we assume

$$\log\frac{p(y_{it}>0|c_i,\mathbf{x}_{it},y_{i,t-1},z_i)}{p(y_{it}=0|c_i,\mathbf{x}_{it},y_{i,t-1},z_i)} = \alpha_{1c_i} + \mathbf{x}_{it}'\boldsymbol{\beta}_{11} + \sum_{j=1}^{3} d_{ij,t-1}\beta_{1,j+1} + z_i\gamma_{1t},$$

$$\log \frac{p(y_{it} > 1|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)}{p(y_{it} \leq 1|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)} = \alpha_{2c_i} + \mathbf{x}_{it}'\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij,t-1}\beta_{2,j+1} + z_i\gamma_{2t},$$

$$\log \frac{p(y_{it} > 2|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)}{p(y_{it} \leq 2|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)} = \alpha_{2c_i} + \tau + \mathbf{x}_{it}'\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij,t-1}\beta_{2,j+1} + z_i\gamma_{2t},$$

where the vector of covariates $\mathbf{x}_{it}$ at occasion t also includes time dummies. Note that the parameters γ_{1t} and γ_{2t}, $t = 2, \ldots, 5$, measure the dynamic effect of the job training program for each period, as they correspond to the difference (on the logit scale) of the probability of success between trained and untrained subjects, all other factors remaining constant; see [7].

Finally, for the binary variable z_i equal to 1 if subject i attends the job training program and to 0 otherwise, we assume

$$\log \frac{p(z_i = 1|c_i, \mathbf{x}_{i1}, y_{i0})}{p(z_i = 0|c_i, \mathbf{x}_{i1}, y_{i0})} = \alpha_{3c_i} + \mathbf{x}_{i1}'\boldsymbol{\delta}_1 + \sum_{j=1}^{3} d_{ij0}\delta_{j+1},$$

with α_{3c}, $c = 1, \ldots, k$, being support points associated to the latent classes.

3.2 Maximum likelihood estimation

Estimation of the model parameters is based on the maximization of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_i \log[p(y_{i1}, z_i, \mathbf{y}_{i2}|\mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})],$$

by an EM algorithm; see [3]. In the expression above, $\boldsymbol{\theta}$ denotes the vector of all model parameters, $\mathbf{X}_{i2} = (\mathbf{x}_{i2}, \ldots, \mathbf{x}_{i5})'$, and $\mathbf{y}_{i2} = (y_{i2}, \ldots, y_{i5})'$.

As usual, this algorithm alternates two steps (E-step and M-step) until convergence and it is based on the *complete data log-likelihood.* On the basis of the dummy variables u_{ic}, the latter may be expressed as

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) &= \sum_i \sum_c u_{ic} \log[p(y_{i1}, z_i, \mathbf{y}_{i2}|c, \mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})\pi_c] = \\ &= \sum_i \sum_c u_{ic} \log[p(y_{i1}|c, \mathbf{x}_{i1}, y_{i0})] + \sum_i \sum_c u_{ic} \log[p(z_i|c, \mathbf{x}_{i1}, y_{i0})] + \\ &\quad + \sum_i \sum_c u_{ic} \sum_{t>1} \log[p(y_{it}|c, \mathbf{x}_{it}, y_{i,t-1}, z_i)] + \sum_i \sum_c u_{ic} \log(\pi_c), \end{aligned} \tag{1}$$

where u_{ic} is equal to 1 if subject i belongs to latent class c and to 0 otherwise.

At the E-step, the EM algorithm computes the conditional expected value of u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, k$, given the observed data and the current value of the parameters. This expected value is denoted by $\hat{u}_{ic}$ and is proportional to

$$p(y_{i1}, z_i, \mathbf{y}_{i2}|c, \mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})\pi_c.$$

The M-step consists of maximizing the expected value of the complete data log-likelihood, obtained by substituting in (1) each u_{ic} by the corresponding expected value computed as above. In this way we update the parameter estimates. In particular, to update the probabilities of the latent class we have an explicit solution given by $\pi_c = \sum_i \hat{u}_{ic}/n$, $c = 1, \ldots, k$. For the other parameters we need an algorithm to maximize the weighted log-likelihood of a logistic model.

A crucial point is the initialization of the EM algorithm. Different strategies may be used in order to overcome the problem of multimodality of the likelihood. As usual, it is convenient to use both deterministic and stochastic rules to choose the starting values and to take, as maximum likelihood estimate of the parameters, $\hat{\boldsymbol{\theta}}$, the solution that at convergence corresponds to the highest value of $\ell(\boldsymbol{\theta})$.

Finally, in order to compute the standard errors for the parameter estimates, we rely on an approximation of the observed information matrix $\mathbf{J}(\boldsymbol{\theta})$, which is obtained as in [2]. In practice, we obtain this matrix as minus the numerical derivative of the score of $\ell(\boldsymbol{\theta})$. The latter is equal to the expected value of the score of $\tilde{\ell}^*(\boldsymbol{\theta})$. The expected value is conditional on the observed data (as that computed at the E-step of the EM algorithm) and is evaluated at the same point $\boldsymbol{\theta}$.

4 Results

In order to illustrate the approach based on the model outlined above, we analyse the dataset described in Section 2 and we compare the results obtained with $k = 3$ latent classes with those obtained from the model without unobserved heterogeneity, i.e. when $k = 1$.

The model with three latent classes has 49 parameters and maximum log-likelihood equal to $-1,027,393$. This value is much higher than that of the model without unobserved heterogeneity; for the latter, the maximum log-likelihood is equal to $-1,043,618$, with 41 parameters. For both models, the parameter estimates are reported in Tables 3 and 4. Note that for the model with unobserved heterogeneity, the three classes have estimated probabilities equal to 0.090, 0.036 and 0.874.

The most interesting aspect is that the estimates of the parameters γ_{ht}, which measure the dynamic impact of the training program, considerably change when unobserved heterogeneity is taken into account, i.e. when we use three latent classes instead of one. In particular, the estimates for the first logit ($h = 1$), which concerns the probability of entering the archive, are always negative with $k = 1$ and become positive with $k = 3$. Less evident is the difference in the estimates of these parameters for the second and third logits ($h = 2$). With both one and three latent classes, these estimates indicate that the training program has a significant effect on the probability of improving in the contractual level only for the first period after the beginning of the program ($t = 2$). There is no evidence of a significant effect for the other periods.

For what concerns the parameters measuring the effect of the individual covariates on the response variables, we do not observe a great difference between the model with three latent classes and that with one latent class. For both models, we note that males tend to improve more easily than females in the contractual level.

		First logit							
		$k=3$				$k=1$			
Effect		estimate	s.e.	t-statistic	p-value	estimate	s.e.	t-statistic	p-value
intercepts	(α_{11})	-3.446	0.019	-177.54	0.000	-1.222	0.013	-97.49	0.000
	(α_{12})	-1.575	0.024	-65.77	0.000	-	-	-	-
	(α_{13})	-0.870	0.015	-57.30	0.000	-	-	-	-
time dummies	(β_{111})	0.401	0.007	59.23	0.000	0.414	0.007	61.01	0.000
	(β_{112})	0.490	0.008	64.63	0.000	0.427	0.007	60.56	0.000
	(β_{113})	0.457	0.008	56.22	0.000	0.362	0.007	50.30	0.000
	(β_{114})	-0.068	0.008	-8.49	0.000	-0.081	0.007	-11.36	0.000
gender*	(β_{115})	-0.020	0.006	-3.58	0.000	-0.025	0.005	-5.54	0.000
age†	(β_{116})	0.029	0.001	44.69	0.000	0.024	0.001	45.22	0.000
dummy educ.‡	(β_{117})	0.137	0.014	9.61	0.000	0.186	0.012	16.00	0.000
education	(β_{118})	0.054	0.005	11.21	0.000	0.075	0.004	19.04	0.000
lag response	(β_{12})	2.214	0.012	190.91	0.000	2.186	0.010	217.87	0.000
	(β_{13})	2.521	0.008	330.85	0.000	2.642	0.007	394.91	0.000
	(β_{14})	3.858	0.011	564.53	0.000	3.818	0.006	673.75	0.000
training	(γ_{12})	1.136	0.067	16.92	0.000	-0.264	0.041	-6.49	0.000
	(γ_{13})	0.639	0.072	8.86	0.000	-0.919	0.044	-20.84	0.000
	(γ_{14})	0.761	0.071	10.66	0.000	-0.819	0.044	-18.49	0.000
	(γ_{15})	1.375	0.070	19.69	0.000	-0.339	0.044	-7.63	0.000
		Second, third logits							
intercepts	(α_{21})	4.092	0.070	58.64	0.000	3.421	0.019	180.42	0.000
	(α_{22})	-3.421	0.045	-76.56	0.000	-	-	-	-
	(α_{23})	4.439	0.022	198.93	0.000	-	-	-	-
shift	(τ)	-4.363	0.010	-445.61	0.000	-3.647	0.007	-518.28	0.000
time dummies	(β_{211})	0.439	0.011	41.24	0.000	0.480	0.010	48.80	0.000
	(β_{212})	-0.130	0.011	-11.74	0.000	-0.128	0.010	-12.63	0.000
	(β_{213})	0.084	0.011	7.50	0.000	0.072	0.010	7.11	0.000
	(β_{214})	-0.032	0.012	-2.59	0.010	-0.028	0.011	-2.60	0.009
gender*	(β_{215})	0.116	0.007	16.41	0.000	0.094	0.006	14.78	0.000
age†	(β_{216})	0.018	0.001	21.70	0.000	0.019	0.001	24.48	0.000
dummy educ.‡	(β_{217})	0.210	0.018	11.64	0.000	0.235	0.016	14.59	0.000
education	(β_{218})	-0.012	0.006	-1.91	0.056	-0.043	0.005	-7.75	0.000
lag response	(β_{22})	-5.701	0.017	-333.33	0.000	-5.056	0.014	-348.97	0.000
	(β_{23})	-2.102	0.009	-233.04	0.000	-1.493	0.008	-194.60	0.000
	(β_{24})	4.447	0.013	349.05	0.000	4.683	0.012	396.20	0.000
training	(γ_{22})	0.219	0.068	3.24	0.001	0.183	0.062	2.96	0.003
	(γ_{23})	0.195	0.092	2.11	0.035	0.138	0.083	1.67	0.094
	(γ_{24})	0.091	0.093	0.97	0.330	-0.002	0.082	-0.03	0.979
	(γ_{25})	0.096	0.095	1.01	0.312	-0.062	0.082	-0.75	0.454

Table 3 Estimates of the parameters obtained with the proposed model for the distribution of the response variables (*dummy equal to 1 for a male and for a female; †minus average age; ‡dummy for the category of education missing).

Moreover, age has a positive effect on the first logit and also on the second and third logits, whereas the number of years of education have a clear positive effect only on the first logit. A strong state dependence is also observed since all the parameters associated to the lagged responses are highly significant, indicating a strong persistence on the same type of contract.

		$k=3$				$k=1$			
Effect		estimate	s.e.	t-statistic	p-value	estimate	s.e.	t-statistic	p-value
intercepts	(α_{31})	-2.139	0.159	-13.48	0.000	-4.458	0.078	-56.94	0.000
	(α_{32})	-5.592	0.092	-60.91	0.000	-	-	-	-
	(α_{33})	-5.381	0.100	-53.57	0.000	-	-	-	-
gender*	(δ_{11})	-0.161	0.034	-4.77	0.000	-0.153	0.032	-4.84	0.000
age†	(δ_{12})	-0.003	0.004	-0.71	0.475	-0.001	0.004	-0.31	0.759
dummy educ.‡	(δ_{13})	0.114	0.084	1.36	0.175	0.031	0.079	0.39	0.694
education	(δ_{14})	0.316	0.027	11.84	0.000	0.249	0.025	10.04	0.000
init. period	(δ_2)	-1.158	0.111	-10.48	0.000	-0.678	0.104	-6.51	0.000
	(δ_3)	-1.436	0.069	-20.76	0.000	-0.921	0.064	-14.50	0.000
	(δ_4)	-1.307	0.043	-30.58	0.000	-0.813	0.035	-23.28	0.000

Table 4 Estimates of the parameters obtained with the proposed model for the probability of attending the training program (*dummy equal to 1 for a male and for a female; †minus average age; ‡dummy for the category of education missing).

It also emerges that the covariates that have a significant effect on the propensity to attend the job training program are gender, years of education and the response at the initial period. In particular, female have a higher propensity to attend the program, as well as subjects with higher educational level and with a less favourable contract position at the beginning of the period of observation.

Finally, the estimates of the random intercepts (α_{hc}) and the corresponding class probabilities (π_c) indicate that there is one main group of subjects corresponding to the third class; these subjects have the highest propensity to enter the archive and to improve in the contractual level. They also have a lower propensity to attend the training program with respect to subjects in the first class and a propensity to attend the program similar to subjects in the second class.

Acknowledgements We acknowledge the financial support of the Einaudi Institute for Economics and Finance. We are grateful to Prof. Mario Mezzanzanica and to Dr. Matteo Fontana, Department of Statistics, University of Milano-Bicocca, for providing the dataset.

References

1. Agresti A. (2002). *Categorical Data Analysis, 2nd Edition.* Canada: Wiley.
2. Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure, *Journal of the American Statistical Association*, **104**, 816–831.
3. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. series B*, **39**, 1–38.
4. Heckman, J. (1981). Heterogeneity and state dependence. In *Studies of Labor Markets*, Ed. S. Rosen. University of Chicago Press, 91–139.
5. Hsiao, C. (2003). *Analysis of Panel Data, 2nd Edition.* Cambridge University Press, New York.
6. Lazarsfeld P. F. and Henry N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.
7. Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.