Quaderni di STATISTICA

VOLUME 14 - 2012

LIGUORI EDITORE

Quaderni di Statistica
Vol. 14, 2012

Evaluation of the degree effect on the work path by a latent variable causal model

Francesco Bartolucci
Department of Economics, Finance and Statistics, University of Perugia
E-mail: bart@stat.unipg.it

Fulvia Pennoni
Department of Statistics, University of Milano-Bicocca
E-mail: fulvia.pennoni@unimib.it

Giorgio Vittadini
Department of Quantitative Methods for Business and Economic Sciences
E-mail: giorgio.vittadini@unimib.it

Summary: We formulate a causal model for the effect of the degree program on the work path of graduate students, which is based on a Markov process to represent latent characteristics of the subjects. A latent Markov model with covariates and mixed-type responses results, which is estimated by an EM algorithm. We illustrate the proposed approach through an application to a dataset deriving from administrative panel data which concern labor market in Lombardy and allows us to evaluate the effectiveness, in terms of education, of certain Universities in Milan.

Keywords: Causal Inference, Education, Labor Market, Latent Markov Model.

1. Introduction

In this paper, we propose a causal model to evaluate the effect of degree programs on the employment status in terms of income, easiness in switching between types of position, and employment skills of the graduates. The approach is motivated by a dataset deriving from the following databases: (*i*) database of the observatory of the Lombardy labor market, (*ii*) database of graduates from the largest universities in Milan, and (*iii*) database of the office of revenues.

The causal model is formulated following the approach of Heckman (2010) that, in a general context, shows the equivalence between econometric structural models and potential outcome models (Rubin, 1974). The model takes into account the longitudinal structure of the data through a Markov chain that has the role of representing individual characteristics which are not directly observable (latent characteristics). Moreover, it allows for individual covariates, which may directly affect the outcomes or the parameters of the Markov chain, and for response variables of different types, essentially categorical and continuous.

The model based on the above formulation is in practice a latent Markov model (Wiggins, 1973, Bartolucci *et al.*, 2010) with covariates and mixed-type responses. Therefore, from the methodological point of view our main contribution is that of casting this model in the causal literature providing an interpretation of its parameters in terms of causal effects that may be used for evaluation purposes. The approach based on this model may be then compared with recent causal inference approaches for longitudinal data; see, among others, Gill and Robins (2001), Abbring and Van Den Berg (2003), and Aalen *et al.* (2012).

2. *The dataset*

The dataset concerns 1258 young individuals who graduated in 2004 from the three main universities of Milan: Milano-Bicocca, Milano-Statale, and Cattolica del Sacro Cuore. They have been followed along 20 quarters after the graduation date and four quarters before, covering the years 2003-2008. The choice of the specific 2004 cohort is motivated by the availability of the data of the employment offices from 2004 to 2009. The response variables are: (*i*) annual income in Euro referred to the previous year, (*ii*) employment status, (*iii*) type of position indicating whether a subject is employed with a temporary or permanent contract, and (*iv*) job quality, measured by the skill level of the job (low, medium, or high). The last one is derived by a categorization of the job qualification made by the Italian National Institute of Statistics. The available covariates concern individual characteristics such as: (*i*) gender, (*ii*) age, (*iii*) number of family components, (*iv*) family income, (*v*) place of birth, (*vi*) type of high school, (*vii*) employement during the graduate studies, (*viii*) place of graduation, (*ix*) type of degree program (scientific, humanistic, or social science and business), (*x*) examination grades, (*xi*) final grade at college.

In Table 1 we report descriptive statistics for the distribution of some of the available covariates, whereas in Table 2 we report the descriptive statistics for the response variables for each year of observation.

The main research question concerns the evaluation of the degree program effect on the labor market transitions of the graduates. A question of this type may be addressed in the framework of causal inference by the model for longitudinal data that we illustrate in the following.

Table 1. Descriptive statistics for the distribution of the covariates

Covariate		%	mean	st.dev.
Male		39.11		
Age in 2004			25.59	1.82
Degree grade			101.91	7.28
Family component in 2004			3.19	1.14
Degree type	*Scientific*	15.66		
	Humanistic	25.36		
	Social S. - Bus.	58.98		
Empl. before 2004	*Part-time*	15.02		
	Full-time	5.64		

Table 2. Frequency of every response variable for each year of observation

Year	Average Income	Temporary cont.%	Skill	
			high %	medium %
2004	6890	53.88	49.33	46.67
2005	11100	54.01	59.17	38.06
2006	15170	50.55	64.47	33.76
2007	18700	48.40	63.94	34.57
2008	20040	42.86	52.44	28.35

3. The causal model

With reference to a subject in the panel, let $\boldsymbol{Y}^{(t)}$ denote the vector the response variables of interest at time occasion t, $t = 1, \ldots, T$, and $\boldsymbol{X}^{(t)}$ denote the corresponding vector of covariates. The proposed model assumes the existence of a latent process $U = (U^{(1)}, \ldots, U^{(T)})$ which affects the distribution of the response variables. The main assumption of the model is that the vectors $\boldsymbol{Y}^{(t)}$, $t = 1, \ldots, T$, are conditionally independent given the latent process and the covariate vectors $\boldsymbol{X}^{(t)}$, $t = 1, \ldots, T$. In such a context the latent variables summarize the observed outcomes. The latent process is assumed to follow a first-order Markov chain with state space $\{1, \ldots, k\}$, where k is the number of latent states. Note that, contrary to other latent Markov formulations, we do not assume that the response variables in $\boldsymbol{Y}^{(t)}$ are conditionally independent given the corresponding latent variable $U^{(t)}$, but, as in Bartolucci and Farcomeni (2009), these variables may be dependent.

We admit that certain covariates affect the distribution of the response variables given the latent process, whereas other covariates directly affect the response variables. Among these covariates we include dummies for the type of degree and for the University where the degree was earned. We show that the regression coefficients for these dummy variables have a causal interpretation in the sense of Heckman (2010). More-

over, among the covariates we include the lagged response variables, so that we account for state dependence. This means accounting for the effect of having a certain income and employment skill at a given occasion on the probability of having the same income and skill at the following occasion, once observable covariates and subject specific unobservable factors are taken into account.

Maximum likelihood estimation of the model parameters is carried out by the EM algorithm. As usual, this algorithm alternates two steps until convergence in the likelihood. At the first step (E-step), the posterior probability of every latent state and pair of consecutive latent states is computed for every subject in the sample. At the M-step, the expected value of the complete log-likelihood, computed on the basis of these posterior probabilities, is maximized by standard rules.

Acknowledgements: We are grateful to Prof. M. Mezzanzanica and Dr. M. Fontana, CRISP, University of Milano-Bicocca, for providing the dataset.

References

Aalen O. O., Røysland K., Gran J. M., Ledergerber B. (2012), Causality, mediation and time: a dynamic viewpoint, *Journal of the Royal Statistical Society, Series A*, DOI: 10.1111/j.1467-985X.2011.01030.x.

Abbring J., Van Den Berg G. (2003), The nonparametric identification of treatment effects in duration models, *Econometrica*, 71, 1491–1517.

Bartolucci F., 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.

Bartolucci F., Farcomeni A., Pennoni, F. (2010), An overview of latent Markov models for longitudinal categorical data, http://arxiv.org/abs/1003.2804.

Gill R.D., Robins, J. M. (2001), Causal inference for complex longitudinal data: the continuous case, *Annals of Statistics*, 29, 1785–1811.

Heckman J. J. (2010), Building bridges between structural and program evaluation approaches to evaluating policy, *Journal of Economic Literature*, 46, 356–398.

Rubin D.B. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, 66, 688–701.

Wiggins L. M. (1973), *Panel Analysis: Latent probability models for attitude and behaviour processes*, Elsevier, Amsterdam.