A Multilevel Chain Graph Model for the Analysis of Graduates' Employment

Anna Gottard, Leonardo Grilli, Carla Rampichini

Statistics Department "Giuseppe Parenti", University of Florence, Italy

Summary. The main goal of the present paper is the analysis of the working position of graduates using multilevel and chain graph models, extended to the case of correlated data. After a brief introduction to multilevel modelling and a description of the conditional independence implied by the model, we describe chain graphs for multilevel models. The model put forward can analyse the factors influencing the graduates' job position, using the data collected on students of the University of Florence who graduated in the year 2000.
Keywords: Chain graph models; Logistic regression; Multilevel models; University system evaluation.

1. Multilevel and chain graph modelling

The university system evaluation requires *ad hoc* methods and statistical models able to capture the complexity of the phenomenon. Such a complexity originates from many facets, such as:

(*a*) the hierarchical structure of the data, entailing a correlation among the observations and requiring the consideration of effects at different levels of the hierarchy. This hierarchical structure (students within classes, classes within study programmes, and so on) is substantial for the analysis and the underestimation of cluster effects and the fact that some of the assumption of the usual regression models are not satisfied, may lead to incorrect standard errors of the estimated coefficients;

(*b*) the presence of variables referring to different moments along the students' careers (e.g. parents education, high school grades, exam grades, graduation grades). This aspect implies a logical and temporal order among the involved variables that must be taken into consideration to shed light on the way students achieve their final result (e.g. getting a

job), and to distinguish between direct and indirect effects.

Multilevel models (Snijders & Bosker, 1999; Goldstein, 2003) allow us to cope with the intra-class correlation and to analyse in a proper way the cluster effects. For such reasons, multilevel models are widely applied in the education evaluation framework. Chain graph models (Cox & Wermuth, 1996) are a useful tool for the representation of the process described at point (b).

In the following, we propose a method for the integration of multilevel and chain graph models that allows:

(i) to properly model the relationships among the probability of finding a job after the degree, the students' careers and their individual characteristics;

(ii) to stress the contribution of the study programme to the student's success in the labour market;

(iii) to distinguish among direct and indirect effects of the background and career variables.

In Section 2, we describe the two-level linear model and its extension to a binary response. In Section 3 the multilevel graph model derived from the integration among the multilevel model and the chain graph model is illustrated. In the fourth Section, we present the data at hand and the main results of the empirical analysis, and in the fifth we conclude by giving some lines for future research.

2. The linear random intercept model

Let us consider a two-level hierarchical structure, where Y_{ij} is the response variable for the i-th subject (first level unit) of the j-th cluster (second level unit), $i=1,2,\ldots,n_j$, $j=1,2,\ldots,J$. For each subject, a vector $\mathbf{X}_{ij}$ of individual (e.g. gender, high school rank) and cluster (e.g. number of enrolled students for each program course) variables is available.

Let us assume Y_{ij} is a continuous variable. If the relationship between the response Y_{ij} and the covariates $\mathbf{X}_{ij}$ is linear, it is possible to specify the following linear random intercept model:

$$\begin{aligned} Y_{ij} &= \alpha_j + \boldsymbol{\beta}' \mathbf{X}_{ij} + \varepsilon_{ij} \\ \alpha_j &= \alpha + U_j \end{aligned} \tag{1}$$

where ε_{ij} are the first level residuals, while U_j are the second level ones. Residuals are assumed to be independent and normally distributed, with zero mean and variances $\mathrm{Var}(\varepsilon_{ij})=\sigma^2$ at subject level and $\mathrm{Var}(U_j)=\tau^2$ at cluster level. Moreover, as it is common in regression models, the correlations among residuals at both levels and the covariates are assumed null. The independence hypotheses among the observations following from this model are: