The Determinants of Graduates' Placement. Analysis of Interactions Using Boolean Logit Models

Mariano Porcu, Giuseppe Puggioni, Nicola Tedesco[1]

Department of Economic and Social Sciences, University of Cagliari, Italy

Summary. In this analysis of the occupational placement of graduates, we define the role played by some covariates assembled to predict the dichotomous event occupied/unoccupied. These covariates influence the response variable singularly and jointly. This work aims to evaluate this joint effect by means of a recently developed technique known as Boolean logit. We applied an exploratory binary segmentation analysis to support the analysis.
Keywords: Graduates' placement; Segmentation analysis; Boolean regression analysis; Logit regression analysis.

1. Introduction

In the evaluation of the performances of the university educational system, the search for the determinants of the occupational placement of graduates is an important issue at stake. The issue has been approached with different methodologies (Chiandotto, 2004; Civardi & Zavarrone, 2004). An approach that appears to have an important role is the logit model, based on causal dependence between a response variable and a set of predictors. The dichotomous variable *employed*/*unemployed* is considered as dependent on a set of p predictors

$$y = f(x_1,\ldots,x_p).$$

[1] This paper is the result of the joint research of the three authors. M. Porcu was responsible for the final editing of Sections 1, 2, 5 and 6, whereas N. Tedesco was responsible for Section 4, and G. Puggioni for Section 3. The authors wish to thank the anonymous referees for their precious suggestions.

Predictors influence the response variable singularly, and in combination with each other. Such responses outline a framework of analysis based on the conceptual category of *causal complexity*. According to Braumoeller (2003), causal complexity is a concept in which "*multiple causes interact with one other, and the way in which they interact is described by the logical operators* and *and* or".

A number of concepts can be considered as special cases of causal complexity, that is:

- multiple conjoint causation: X_1 and X_2 and X_3 produce Y;
- substitutability: X_1 or X_2 or X_3 produce Y;
- contexts: X_2 produces Y but only in the presence of X_1;
- necessary and sufficient conditions: X_1 and X_2 produce Y, either X_1 or X_2 produce Y;
- INUS conditions[2]: (X_1 and X_2) or (X_3 and X_4) produce Y.

Complex causation is a problem for the majority of standard statistical techniques. The problem is that causal complexity implies non-additivity, which arises from the cumulative process of the influence of the independent variables on the response variable. This means that the presence or absence of one independent variable mitigates (or even nullifies) the impact of another. So, from a practical point of view, the problem arises of how to "capture" causal complexity with standard statistical techniques.

A number of methodological proposals have been put forward and much attention has been paid to studying them (Frosini, 2004). With reference to the dichotomous event *employed*/*unemployed,* we observe that in several research studies it has been stated that the event could be considered as the outcome of a process of causal complexity.

2. Modelling the interactions

The logistic regression model is often used to model the probability of a certain event as a function of a set of explanatory variables. The influence of the explanatory variables on the response is considered linear on a logit scale

$$\log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ .$$

The possible joint effects among covariates are usually taken into account by fitting the product among the variables into the model itself (Hosmer & Lemeshow, 1989):

$$\log(\pi/(1-\pi)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \{X_1 \times X_2\} \ .$$

[2] The acronym INUS, created by Mackie, defines a particular kind of causal relationship. It makes reference to "an insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result" (Braumoeller, 2003).

This modelling forces the researcher to keep the interactions among the variables on a very simple level, generally one-way or two-ways, for both technical (e.g. sparsity of data, power of tests) and theoretical reasons (e.g. parsimony principle). Consequently, only the main effects are usually included in the model, although the joint effects are more useful in predicting or in selecting groups, especially for social analysis purposes.

2.1 The Boolean logit

A method that can take into account the relationships among variables that are rooted in the concepts of causal complexity is the so-called *Boolean logit* (Braumoeller, 2003). It allows the researcher to estimate the influence of the interactions among predictors on the binary response Y. It is postulated that Y is thought to be produced by a *Boolean* or *logic* combination of some conditions $A_1, \ldots, A_k, \ldots$, e.g.:

$$A_1 \text{ and } (A_2 \text{ or } A_3) \rightarrow Pr(Y=1) = \pi = Pr(A_1) \times Pr(A_2 \cup A_3).$$

where *Pr(.)* denotes the probability of the argument. The probability of occurrence for each condition

$$Pr(A_K) = p_k$$

is expressed by means of a logit (or probit) model (Braumoeller, 2003):

$$\boldsymbol{p_k} = \frac{\exp(\boldsymbol{\beta_k X})}{1+\exp(\boldsymbol{\beta_k X})}.$$

The k index means that each "condition" depends on its own explanatory variables $X = \{X_j\}$ through corresponding parameters β_k. The same X_j can be included in different p_k with no multicollinearity arising. Obviously, if there is only one "condition", Boolean logit reverts to the standard logit.

Boolean logit can be helpful in solving the problem of statistical estimation when causal complexity is present. To fit it, we need to postulate a model for π and to express π as a function of explanatory variables and relevant parameters through the p_k. As an example, if we assume that

$$\pi = Pr(A_1) \times Pr(A_2),$$
$$\text{logit}(p_1) = \mathbf{x}_1' \boldsymbol{\beta}_1; \quad \text{logit}(p_2) = \mathbf{x}_2' \boldsymbol{\beta}_2,$$

the model takes the form

$$\pi_i = p_{1i} \times p_{2i}$$

and the likelihood is

$$\boldsymbol{Lik}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n} (\boldsymbol{p}_{1i} \times \boldsymbol{p}_{2i})^{y_i} (1 - \boldsymbol{p}_{1i} \times \boldsymbol{p}_{2i})^{1-y_i}.$$

So, once the occurrence of the event is explained in the language of *complex causation*, the consequent hypothesis can be expressed in probabilistic terms[3]. In conclusion, it might be worthwhile noting that the brackets are quite important since the Boolean statement "(*A and B*) *or C*" is different from "*A and* (*B or C*)".

3. The data

Data were collected with a CATI survey carried on in November 2003. A sample of 1,112 graduate students of the University of Cagliari was selected among those who graduated in the years 1999 and 2000. At the end of the survey, interviewees were classified according to their occupational status: 823 were employed, 150 unemployed, 137 were enrolled in postgraduate educational programmes (2 were missing). The whole group of employed was then split into two subsets, one of which included graduates (756) that found employment after graduation, the other graduates that were in work before they finished university studies (67).

Before performing the analysis, we decided to fix some eligibility criteria for the observation, i.e.

occupational status:

- employed;
- unemployed;

for the employed:

- job found after graduation;
- no more than 36 months to get the job.

Because of such criteria, and in order to model the dichotomous event "*Y*" *employed*/*unemployed*, we considered a total amount of 837 observations. Among them, 687 were employed (*Y*=1), 150 unemployed (*Y*=0).

During the survey, a lot of information on demographic as well social information on the graduates was collected. Evaluations on their educational programmes and previous working experiences were also collected together with data on time spent looking for a job (Porcu & Tedesco, 2004; Porcu & Puggioni, 2004).

Some exploratory analysis was carried out. The results (not shown here) led to the selection of a set of variables that could be particularly relevant in undertaking the present analysis, namely *sex, high school attended, mark gained in the high school leaving examinations, degree*[4], *age at graduation, final*

[3] A possible alternative to Boolean logit to model causal complexity is the so called *Logic Regression* recently proposed by Ruczinski *et al*. (2003).

[4] The study programmes were classified into four groups. a) Economic-Legal-Social (ELS); b) Scientific-Technical (SCT); c) Health and Life Sciences (SHL); d) Humanities and Behavioural Sciences (HBS).

mark at graduation, post-graduation studies. We will dichotomise all these variables.

4. Groups of variables for the analysis of interactions

One of the major drawbacks of the Boolean logit is the arbitrariness in the choice of the combination of predictors. A good set of predictors is the researcher's beliefs and opinions; such ideas are obviously rooted in her/his own experience on the topic that is being investigated. In any case, the uncertainty and the subjectivity of such a method for choosing the combinations of predictors may weaken the final model.

Moreover, the Boolean logit models require numerous computational resources and a model choice based on comparisons among the results gained with different combinations of predictors that could be excessively time consuming.

In the following, we propose a method to choose the combination of predictors. We will apply an exploratory segmentation analysis to give the researcher useful hints on the influence exercised by a set of predictors on a response variable and on the relationship between them. With such a method, results could vary with respect to the choice of the kind of segmentation (especially considering the chosen "*criterion function*").

Nevertheless, the choice for a binary segmentation analysis based on the likelihood criterion function could be considered suitable in that it allows the researcher not to choose a distance function (Tedesco, 2002). Segmentation could be considered as a representation of the causal complexity data as well. Therefore, our aim is not to lose this kind of information while building up Boolean groups. Essentially, we try not to make the abstract ideas of the researcher prevail but let the data 'speak for itself'.

The software used for segmentation was SAS-RECPAM (Carinci & Pellegrini, 2001; Ciampi, 1991). The chosen criterion function is to maximise the likelihood ratio of the logit *employed*/*unemployed* with respect to all the possible two by two combinations of the predictors; a minimum number of 40 observations per node/leaf was also incorporated into the model (minimum 10 of them employed). The split was set to α=0,05. Such a restrictive criterion was chosen to form a concise tree with not too many branches. The main objective was, indeed, to explore data for the next Boolean logit analysis.

The possible predictors we set in the segmentation were: the *attending of a post graduate programme* (Yes/No); *sex* (Male/Female), *type of high school attended* (Lyceum[5]/Other); *kind of degree* (ELS, SCT, SHL, HBS); *mark*

[5] The Italian "Lyceum" provides a classical education such as the one offered by the old British "Grammar Schools".

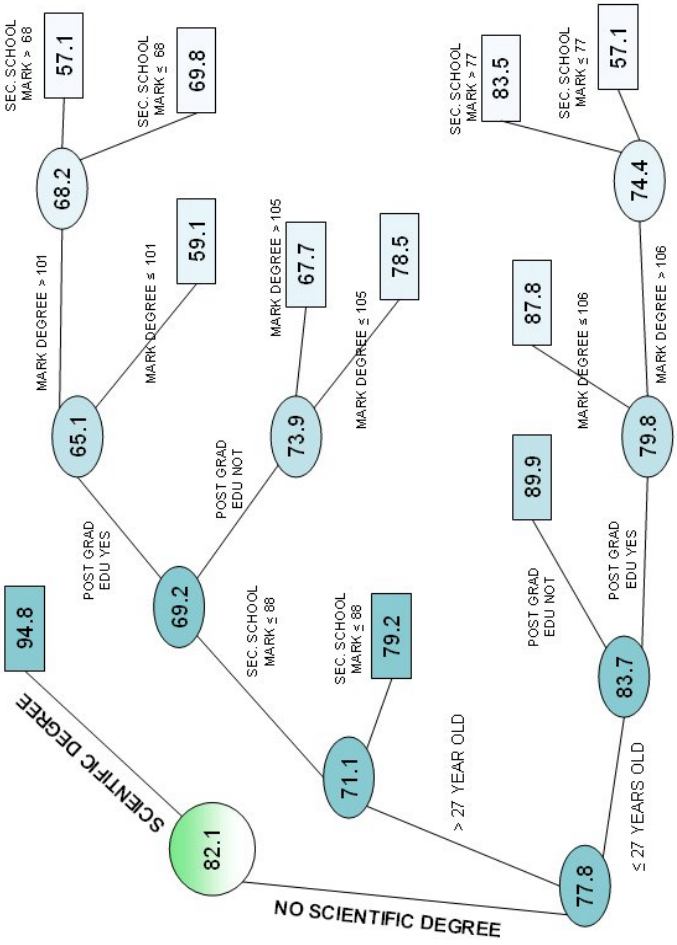

Figure 1. Segmentation tree

gained in the school leaving examinations ; *age at graduation*, *final mark at graduation*. We considered these variables on a continuous scale so to obtain guidance for the setting of thresholds useful to collapse them into categories.

The tree diagram (Figure 1) shows some interesting results. First, we perceive a marked asymmetry because graduates in the SCT group have a high employment rate (94.8%) and do not split further. This means that for these graduates the kind of degree they gained is the only important factor of success in getting employment.

On the other hand, for all the other graduates, the total employment rate is lower (77.8%) but it reaches notably higher values, in particular, for those younger (<27 years old) and who did not attend postgraduate programmes of study (89.9%). Among older graduates (≥27 years old), what seemed to be the

cause of notable disadvantage was a low mark at high school final examination, the attendance of post graduates studies and a low mark at graduation.

It is clear that among graduates that are not in the SCT group, the low age at graduation is more useful than postgraduate studies for getting a job. Maybe, this is because post-graduate education is sometime a sort of standby status for those seeking occupation, and training is provided directly by companies.

The mark gained at high school leaving examinations also seems interesting: it occurs several times in segmentation, especially if this evidence is compared with the non-significant role played by variables such as sex, or type of high school. It may suggest that the "quality" of a graduate lies not only in the university education but also in the high school curriculum, as it plays an important role in determining the occupational placement of a graduate.

Focusing again our attention on the non-SCT group and on younger and with a postgraduate education, we highlight that the graduation mark has little influence (87.8% is the percentage of employed with a graduation mark ≤106/110). Among those with a high graduation mark (>106), it is the mark gained at high school leaving examinations that seems to play a role.

In conclusion, for the non-SCT group (considering the values of the GPI[6], see Table 1) marks for high school leaving examinations and the age at graduation play a joint effect on the response variable.

Table 1. GPI values

Covariates	GPI
High school leaving examination mark	100
Degree	96
Age at degree	90
Degree mark	60
Postgraduate studies	55
Type of high school attended	28
Sex	27

5. Modelling the event employment

To model the event *employed* (Y=1) *vs unemployed* (Y=0) we considered the results of the segmentation analysis and we treated them as dichotomous variables (1=Yes; 0=Not):

[6] The GPI (Global Predictive Index) is a measure of the predictive power of a covariate based on increments of the LRS for each covariate in each node with respect to the value of LRS obtained without that predictor. It is a measure of the informative power of the *i*-th predictor. Once the *i* sums are calculated, the bigger value is set at 100 and the remaining are rescaled in relation to this (Ciampi, 1991).

- male sex (SEXM);
- "lyceum" attended as high school (LICCS);
- mark at high school final examination ≥90/100 (DIP90);
- degree Technical-Scientific (SCIEN);
- age at degree ≤ 26 years (LAU26);
- final mark at graduation ≥108 (VOTOHIGH);
- postgraduate studies = Yes (CORPOST).

5.1 Fitting of a standard logit

We will first present the results of the fit of a standard logit model and then the fit of the Boolean logit. The results are summarised in Table 2. We can see that the only significant ($\alpha = 0.05$) variables[7] are DIP90, SCIEN, LAU26 and CORPOST.

Table 2. Point estimates ($\hat{\boldsymbol{\beta}}$) and z-scores ($z = \hat{\boldsymbol{\beta}} / \boldsymbol{SE}(\hat{\boldsymbol{\beta}})$) for the standard logit model (*Log*Lik=-362,937)

Covariates	$\hat{\boldsymbol{\beta}}$	|*z*|-*score*
SEXM	0.1967	0.916
LICCS	-0.2801	1.417
DIP90	0.5453	**2.068**
SCIEN	1.4855	**4.315**
LAU26	0.6134	**2.875**
VOTOHIGH	-0.2575	1.301
CORPOST	-0.4493	**2.309**

Table 3. Point estimates ($\hat{\boldsymbol{\beta}}$) and z-scores ($z = \hat{\boldsymbol{\beta}} / \boldsymbol{SE}(\hat{\boldsymbol{\beta}})$) of some standard logit models

Covariates	*Basic model*		*Without SCIEN*		*With SEXM×SCIEN*	
	$\hat{\boldsymbol{\beta}}$	|*z*|-*score*	$\hat{\boldsymbol{\beta}}$	|*z*|-*score*	$\hat{\boldsymbol{\beta}}$	|*z*|-*score*
SEXM	0.1967	0.916	0.5464	0.916	0.1780	0.787
LICCS	-0.2801	1.417	-0.2030	1.417	-0.2770	1.398
DIP90	0.5453	2.068	0.8068	2.068	0.5454	2.069
SCIEN	1.4855	4.315	–	–	1.3758	2.544
LAU26	0.6134	2.875	0.5251	2.470	0.6128	2.873
VOTOHIGH	-0.2575	1.301	-0.3212	1.633	-0.2579	1.303
CORPOST	-0.4493	2.309	-0.5115	2.668	-0.4445	2.274
SEXM × SCIEN	–	–	–	–	0.1778	0.256
log*Lik*	-362.937		-374.985		-362.904	

[7] Negative values for $\hat{\beta}$ indicate a lower probability for the event.

After the fitting a second model that took into account the first order interactions among predictors, we observed that none of them seemed to be significant for the response (Table 3).

The causal relationship as the one just described poses the researcher with the problem of how to interpret the effect exercised by the set of predictors on the response variable *Y*. For example, if from the base model in Table 2 we do not include the SCIEN predictor, we can note that the variable SEXM has a significant influence (even though the model is poorer in terms of log*Lik*). Nevertheless, if we fit a new model that considers the interaction term between SCIEN and SEXM, it is not significant.

5.2 Fitting a Boolean logit

To fit a Boolean logit[8] model we have considered the same predictors used for the standard logit (namely, SEXM, LICCS, DIP90, SCIEN, LAU26, VOTOHIGH, CORPOST). To fit the Boolean logit model it is necessary to posit some preliminary conditions. Taking into account the results we obtained with the binary segmentation analysis we decided to set the following conditions:

- A_1 = "ownership of *winning* qualifications for the job market"
- A_2 = "ownership of characteristics pertaining to the educational profile".

A_1 is defined by a set of covariates that refers to some of the most notable characteristics of someone who is seeking to enter the job market, namely *age* and *skills*: LAU26 and SCIEN. A_2 is defined by a set of covariates that refers to the educational background of the graduate, plus the variable sex, namely, DIP90, LICCS, VOTOHIGH, CORPOST and SEXM.

The probability of being employed, $Pr(Y=1)=\pi$ is modelled as the interaction among A_1 and A_2:

$$\pi = Pr(A_1) \times Pr(A_2)$$

The conditions A_1 and A_2 are expressed as an additive function of the explanatory variables considered:

- A_1 = LAU26 + SCIEN
- A_2 = SEXM + DIP90 + LICCS + VOTOHIGH + CORPOST.

As we can note from Table 4, the results are rather similar to those we had with the standard logit both in terms of log likelihood and parameter estimates. Nevertheless, the underlying models are quite different: in the standard logit, no interaction term shows significant effects on the response and therefore the model suggests that no variable influences the probability of dropping-out independently of the remaining ones.

On the contrary, in the Boolean logit, the model shows that the response is

[8] To process the data we applied the "Boolean" library available for R package (R Development Core Team, 2003).

Table 4. Point estimates ($\hat{\boldsymbol{\beta}}$) and z-scores ($z = \hat{\boldsymbol{\beta}} / \boldsymbol{SE}(\hat{\boldsymbol{\beta}})$) for the standard logit model and two Boolean logit models

Covariates	*Standard*		*Boolean 1*		*Boolean 2*	
	$\hat{\boldsymbol{\beta}}$	*|z|-score*	$\hat{\boldsymbol{\beta}}$	*|z|-score*	$\hat{\boldsymbol{\beta}}$	*|z|-score*
LAU26	0.6134	2.875	1.0330	2.996	1.0363	1.904
SCIEN	1.4855	4.315	2.2442	2.813	2.2573	1.192
LICCS	-0.2801	1.417	-1.4221	1.094	0.0070	0.008
SEXM	0.1967	0.916	0.5035	0.796	0.4973	0.501
DIP90	0.5453	2.068	1.3693	1.715	1.3640	1.326
VOTOHIGH	-0.2575	1.301	-1.2291	1.251	-1.2127	0.527
CORPOST	-0.4493	2.309	-1.6930	1.066	-1.6632	0.408
LICCS	–	–	–	–	-1.4294	0.904
log*Lik*	-362.937		-360.640		-360.639	

produced by multi-vector interactions, namely interactions of vectors of variables. Thus, it has been assumed that a student's dropout from University depends jointly on A_1 and A_2, with the parameters tacitly maintaining an "interaction meaning". The parameters DIP90 and CORPOST show no significant influence on the response Y and this means that in interacting with the other variables these covariates lose their predictive power.

The Boolean logit allows the researcher to fit models where the same covariate is set in more than one "condition". For example, the variable *lyceum as high school* (LICCS) could be inserted both in condition A_1 and in condition A_2. The results of the fitting are in Table 4 in the "Boolean 2" columns. In spite of the variable LICCS being statistically not significant, it acts on the response in opposite directions if considered in A_1 or in A_2.

6. Conclusions

The use of standard logit to model the probability of a dichotomous event as the effect of a causal relationship from a set of predictors gives researchers some useful tools to understand social phenomena. These tools (namely the log-odds ratios) allow the researcher to interpret the role played by each predictor on the response controlling for the remaining parameters.

Modelling the probability of graduates of finding an occupation, the standard logit allows us to identify factors that negatively influence the probability of 'getting an occupation'. Among these factors, the graduates who obtained high marks at graduation and in their postgraduate studies are likely to be older and therefore less attractive for the companies when they try to enter the job market.

We have seen that other factors play an opposite role (they contribute to increasing the event probability). Among these, we can highlight the case of science graduates and the importance of completing a study programme when still relatively young. We should remember that the fitted standard logit does not take into account possible interaction among covariates. Such an assumption entails a form of additive causal dependence that makes it difficult for the model to catch the full complexity of the phenomenon studied.

Boolean logit is neither an alternative nor a method better than the standard logistic one, but it does offer an advantage: it allows the researcher to consider models that consider causal complexity. Causal complexity mechanisms make it possible to improve the predictive power of the variable response models. Their major drawback lays in the subjective choice of the probability statements that lead to the combination of predictors.

However, the likelihood-based criteria for choosing the best model tend to mitigate this subjectivity. Another drawback is that (unlike standard logit) the parameters are not (log) odds-ratios for the response. Finally, the method is computationally time consuming.

Nevertheless, considering the encouraging results gained in this and in other applications (Muggeo & Porcu, 2004) we can conclude that Boolean logit is a useful tool for implementing sensitivity analyses of other models and for re-enforcing the evidence that emerges regarding the meaning of the predictors studied.

References

BRAUMOELLER B.F. (2003) Causal complexity and the study of politics, *Political Analysis*, **11**: 209-233.

CARINCI F., PELLEGRINI F. (2001) *RECPAM/SAS (Recursive Partitioning and Amalgamation): a statistical tool for criterion-driven data-mining*, Technical Report, in http://med.monash.edu.au\publichealt.

CHIANDOTTO B. (2004) La situazione occupazionale dei laureati: dall'indagine alla pianificazione degli interventi sui percorsi formativi. In: CIVARDI M. (ed) *Transizione Università-Lavoro: la definizione delle competenze*, CLEUP, Padova: 1-18.

CIAMPI A. (1991) Generalized Regression Tree, *Computational Statistics and Data Analysis*, **12(1)**: 57-78.

CIVARDI M., ZAVARRONE E. (2004) Proposta di un modello generatore delle competenze acquisite attraverso la formazione universitaria. In: AURELI CUTILLO E. (ed) *Strategie metodologiche per lo studio della transizione Università-Lavoro*, CLEUP, Padova: 141-152.

FROSINI B.V. (2004) Causality and Causal Models. In: *Atti della XLII Riunione della Società Italiana di Statistica*, *Volume 1*, CLEUP, Padova: 3-32.

HOSMER D.W., LEMESHOW S. (1989) *Applied Logistic Regression*, John Wiley & Sons, New York.

MUGGEO V, PORCU M. (2004) Factors that Cause University Students to Drop Out. An Alternative Modelling of Interaction Terms in Logistic Regression Models. In: *Atti della XLII Riunione della Società Italiana di Statistica*, *Volume 2*, CLEUP, Padova: 511-514.

PORCU M., PUGGIONI G. (2004) Condizione occupazionale e prima valutazione del fenomeno dell'emigrazione dei laureati dell'Ateneo di Cagliari. In: D'OVIDIO F. (ed) *Professioni e competenze nel lavoro dei laureati*, CLEUP, Padova: 317-326.

PORCU M., TEDESCO N. (2004) Dall'Università al Lavoro: analisi dei tempi di passaggio dei laureati dell'Ateneo di Cagliari. In: AURELI CUTILLO E. (ed) *Strategie metodologiche per lo studio della transizione Università-Lavoro*, CLEUP, Padova: 281-295.

RUCZINSKI I., KOOPERBERG C., LEBLANC M. (2003) Logic Regression, *Journal of Computational and Graphical Statistics*, **12**: 475-511.

R DEVELOPMENT CORE TEAM (2003) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, in: http://R-project.org.

TEDESCO N. (2002) Analisi di segmentazione di una coorte di immatricolati dell'Università di Cagliari. In: PUGGIONI G. (ed) *Modelli e metodi per l'analisi dei rischi sociali e sanitari*, CLEUP, Padova: 141-160.