

Forecasting Models for University Students Careers

Modelli di Previsione per le Carriere degli Studenti Universitari

Giulio Ghellini

Università degli Studi di Siena, Piazza S. Francesco 8, Siena, ghellini@unisi.it

Francesco Maccari

Università degli Studi di Firenze, Viale Morgagni 59, Firenze, maccari@ds.unifi.it

Riassunto. Il consistente incremento alla propensione di proseguire gli studi universitari, unitamente ai sempre più stringenti vincoli di bilancio, impone oggi alle Università di prestare maggiore attenzione alle attività di monitoraggio e di tutorato al fine di consentire la messa a punto di idonei strumenti che facilitino il rispetto dei tempi di studio degli studenti universitari. In questo lavoro, utilizzando sia basi di dati amministrativi sugli immatricolati e sulla loro carriera, sia indagini *ad hoc* svolte sui medesimi soggetti, viene proposto un modello logistico atto ad individuare un coerente insieme di variabili indicatrici della difficoltà a rispettare i tempi di studio. L'individuazione delle variabili, da un lato può essere vista come un strumento di valutazione dell'Università, dall'altro può consentire la predisposizione di coerenti interventi di orientamento e di tutorato *in itinere*.

Keywords: logistic model, school career, tutorship.

1. Introduction

The ultimate purpose of the paper is the definition of some valid instruments for the detection of variables that may influence the university students career (for a review of application on Italian University see Bottiroli Civardi M. and Camiz S., 1997). The basic idea is to achieve this aim through the observation of the performance of some cohorts of students throughout their first year at the university. So far, the study is limited to the students who attended the first year at the Faculty of Economics in the University of Siena in 2001-02. Performance is measured only in terms of the quote of credits achieved by each student by the end of the autumn session of exams (30th September 2002).

The analysis proposed consists of two Logit models (see Agresti A., 2002 for details), by which we try to estimate the probability that a student will not reach a certain percentage of the total credits of the first year, given the value of some covariates that are available at the moment of enrolment at the university. The data used to apply these models will be described in the following section. The relevance of this work lays in the fact that a sufficiently precise identification of those students who are, in the sense above specified, "at risk" may well stimulate and improve the organization of effective individual tutorial activities by the university, so as to prevent students from having a too long university path or, even worse, from quitting their studies before getting their degree.

2. The data base

In order to perform the analyses, we have taken into account to process four data sets. The first two, gathered from the university administrative data warehouse, are:

- a) the “Exams” data set, containing, for each student of the analysed population, the list of the exams to be taken in the first year. For each exam, the data set reported the amount of credits it gives and, if successfully passed, the mark that the student received;
- b) the “Private Data” data set, reporting, for each student, the enrolment form recording.

In particular, the first data set allowed us to calculate the percentage of credits reached by each student, which we have referred to in order to create the response variable of the two models. From the second data set we have instead extracted some explanatory variables, which had the great advantage to be available for the largest part of the population involved in the research and, therefore, represent a good basis for a first, general model.

In order to increase the number of possible explanatory variables, we have also decided to refer to the results of a not compulsory aptitude test and of a motivation survey (Maccari F., 2002), to which our university students are recommended to participate at the beginning of their university studies. The results of the test and the responses to the survey were gathered into two more data sets, that we processed to obtain some other variables.

The four data sets described above have been gathered from three different sources. “Exams” and “Private Data” data sets, as we have already said, are managed by the Central Administration of the University of Siena and, therefore, contained a very low number of missing data. The only problem we had to face was that some explanatory variables extracted from the “Private Data” were generally missing for foreign students. The major problem with the other two data sets, instead, arose from the fact that the participation to both the test and the survey is optional for the students. In this case, only 170 students (25.1%) attended the test, while 434 (64.2%) took part in the survey (out of a population of 676 students matriculated in 2001). Unfortunately, the very low response rate to the test enabled us to use the variables collected with it; then, the only variable used in the general model is the participation or not to it. Of course, the same variable, in this case referred to the participation to the survey, has been also included in the general model.

Regarding to the second model, it is to be underlined that the additional variables coming from the survey are available only for the respondents; hence the second model could be performed only for the 434 students who took part in the survey.

3. Models specification and estimation

As response variable of the models to be specified a credit threshold has been chosen defined on the basis of the results of the first three examination session (spring, summer and fall). The threshold has been fixed equal to the thirty percent of the total credits forecast for the first university year, so that the student below this threshold could be

considered having great difficulties in following a regular path¹. The first step of the analysis has been devoted to the specification and estimation of a model (the general model) for the entire cohort (n = 676), based on administrative data. In particular, the following variables have been considered: gender, type of secondary school and final mark achieved (maturity exam). In this analysis, also two dummy variables, concerning the participation or not to the two above mentioned surveys, have been added: in fact, previous descriptive analyses had pointed out some relationship between this dummies and the response variable.

After that, the predictive capability of some other variables collected with the motivational survey has been evaluated by means of a new model, estimated only on students who responded to the survey (n = 434). This model also included the general model variables, with the obvious exclusion of the two dummies concerning the response or not to the *ad hoc* surveys. It is to be underlined that preliminary analyses on the survey non-response mechanism show the absence of significant relationship between non response and all private data included in the model. The new explanatory variables are related to the following spheres: difficulties encountered during secondary education, students' expectations from the chosen faculty, participation to university guidance during secondary school, study behaviour, attendance to university lessons.

The choice of introducing the second model has two main objectives: first, we wanted to check whether the selection of only the sub-population of the respondents to the survey would induce a certain variation in the estimates of the parameters that were also included in the general model; on the other hand, we had the opportunity to observe if the additional variables introduced would substantially improve the fitting capability of the model.

The two models have been estimated by means of a stepwise regression procedure, which has led to the specifications reported in Tab. 1.

Tab. 1. Logit models estimates of not-reaching the thirty percent credits threshold

Variables	General model			Model for respondents to the survey		
	Estimate	p-value	Odd ratio	Estimate	p-value	Odd ratio
Intercept	-1.29	***		-1.79	***	
Tecnical Institute	1.17	***	3.22	1.24	***	3.44
Professional Institute	2.63	***	13.84	2.70	***	14.89
Other not Lyceum	1.67	***	5.33	1.75	**	5.73
60-79 Final mark	0.58	***	1.77	1.04	***	2.82
90-100 Final mark	-1.68	***	0.19	-1.53	***	0.22
Test participation	-1.39	***	0.25	n. d.	-	n. d.
Survey participation	-0.55	***	0.58	n. d.	-	n. d.
Working perspective	n. d.	-	n. d.	-0.86	***	0.42
Univ. Guidance	n. d.	-	n. d.	-1.03	***	0.36
Working student	n. d.	-	n. d.	1.33	***	3.79
Adj. R square	0.32			0.43		

P-value: *** = $\alpha < 0.01$; ** = $0.01 \leq \alpha < 0.05$.

¹ Since information about the quote of achieved credits has been gathered before the end of the first academic year, it has not been possible to have any administrative information about students who had already quitted their studies at that moment.

The analysis of the models output leads us to focus on the following major results.

- a) As expected, in both models there is a strong influence of the type of high school attended and of the mark obtained at the maturity exam on the probability of not reaching the thirty percent of credits.
- b) In the general model, both the dummy variables referring to the participation to the test and to the survey are significant and negative. That means that the participants are the least exposed students to have difficulties during their university career.
- c) The only variable that has no significant effect in the general model is the gender of the students.
- d) In the second model, the procedure detects three additional significant variables. In detail, there is a statistical evidence that students who took part to guidance meetings for the selection of the university courses and those who have based their choice of the faculty on future working prospects are more likely to overtake the thirty percent quote of credits in the first year. On the contrary, as one could well expect, students who want to attend university while working are unlikely to reach that threshold.
- e) The second model fits the observed data considerably better than the general one, as it arises from the comparison between the values of adjusted r-square. In general, this result may be influenced by the fact that the second model is estimated on a subgroup of those students included in the first one, but this does not seem the case, for the reasons explained at the beginning of this paragraph. Therefore, it underlines the importance of the role that this kind of surveys may hold among instruments of prediction and control of students' careers. Consequently, it urges to extend the participation to the survey to the largest possible part of students, in order to minimize the risk of selection effects, and, at the same time, it justifies further research work to create other instruments aimed to collect more pertinent variables.

The next steps of our research will concern the application of the same models to cohorts of other Faculties of the Siena University to test the persistence or not of the same variables significance and sign; furthermore, the estimated models could lead us to the estimation of the different probability risk of not reaching the threshold for the subgroups of student identified by the selected variables, giving the possibility of defining more precise tutorial policies.

References

- Agresti A. (2002) *Categorical Data Analysis*, Wiley, Hoboken.
- Bottiroli Civardi M., Camiz S. (1997) *La popolazione studentesca e le Università italiane: indagini, modelli e risultati*, CLEUP, Padova.
- Maccari F. (2002) *Indagine sulla scelta universitaria delle matricole dell'Università degli Studi di Siena*, Università degli Studi di Siena, mimeo.