Predicting students' academic performance: a challenging issue in statistical modelling

Leonardo Grilli, Carla Rampichini and Roberta Varriale

Abstract We discuss some issues related to the statistical modelling of students' academic performance, with special emphasis on the role of predictors such as high school marks and pre-enrolment tests. After a brief review of the literature, we outline the strategies we devised in the analysis of data on the freshmen of the University of Florence.

Key words: binomial mixture model, concomitant variables, excess zeroes, hurdle model, regression chain graph

1 Introduction

Predicting students' academic performance is a key step in order to improve the efficiency of university systems. Indeed, delays or failures are costly for both the students and the administration. Therefore, it is of primary importance to determine the factors associated with the performance in order to plan actions such as guidance, restrictions to the access, and tutoring. To this end, universities can typically rely on information about the high school career, such as the type of school and various measures of proficiency. However, the results at high school are not fully appropriate to predict the academic performance due to several issues, including the possible mismatch between the competencies evaluated at high school and the com-

Leonardo Grilli
Department of Statistics, Informatics, Applications 'G. Parenti', University of Florence, e-mail: grilli@disia.unifi.it

Carla Rampichini
Department of Statistics, Informatics, Applications 'G. Parenti', University of Florence, e-mail: rampichini@disia.unifi.it

Roberta Varriale
Istat, Rome, e-mail: varriale@istat.it

petencies required for a given degree program, and the heterogeneity in the criteria for awarding marks (usually, there is substantial variability across types of schools and across geographical regions). To overcome those issues, some universities devise a pre-enrolment assessment test tailored on the needs of each degree program. However, a quick look at pre-enrolment tests around the world reveals a lack of commonly accepted guidelines and a shortage of empirical evidence about the predictive ability.

The literature on the empirical research about predicting students' academic performance is scattered in various journals, ranging from Psychology to Economics. Some noteworthy papers are Murray-Harvey (1993), Wedman (1994), Hoefer and Gould (2000), Murphy et al. (2001), Maree et al. (2003), Dancer and Fiebig (2004), Win and Miller (2005), Smith and Naylor (2005), Birch and Miller (2006), Birch and Miller (2007), Mills et al. (2009), Mallik and Lodewijks (2010), Bianconcini and Cagnone (2012), Chowdhury and Mallik (2012), Adelfio et al. (2013).

The statistical modelling of the academic career is challenging due to the complexity of the process. For example, the pre-enrolment test is an instrument to measure students' competencies in addition to already known characteristics, such as the high school mark, thus it is important to assess the value added by the test and to disentangle the effect of the high school mark on the academic performance into a direct effect and an indirect effect mediated by the test. To this end, the analyst has to rely on complex approaches such as path models (Murray-Harvey, 1993) or the regression chain graphs discussed in the following.

Another complication for the statistical modelling of gained credits is that the observed distribution is typically quite irregular: in fact, exams yield different number of credits and the sequence of exams varies across students; moreover, the distribution usually has peaks at zero and at the maximum. A simple approach such as OLS regression can still be used to analyze the associations, but it cannot be used to make predictions. To this end, a proper statistical model is required, even if the features of the response variable rule out conventional parametric models. Two effective methods are the quantile regression (Birch and Miller, 2006; Adelfio et al., 2013) and the mixture regression discussed in the following.

In the rest of the paper we focus on a case study about the pre-enrolment test at the University of Florence, illustrating some modelling strategies based on regression chain graphs, mixture models and hurdle models.

2 Data on freshmen at the University of Florence

In the academic year 2008/2009, the School of Economics of the University of Florence introduced a compulsory test to evaluate the background of the students wishing to enrol in a degree program. The test has 3 editions (September, November and December) and it is based on 40 multiple-choice items covering 3 areas: Logic (12 items, 30%), Reading (10 items, 25%) and Mathematics (18 items, 45%). For each item, one out of 5 alternatives is correct, with the following scoring system: 1 if correct, 0 if blank, -0.25 if wrong. Thus the total score ranges from -10 to 40, and the threshold for passing the test is fixed at 9: candidates with a lower total score are advised against enrollment. In such a case, they can still enrol in a degree program of the School of Economics, but they can take examinations only after passing the test during one of the later editions.

We consider the participants to the first edition of the test (September 2008). The data set is obtained by merging data collected at the test with the administrative data of the School of Economics. After deleting 68 foreign students (due to missing information), the data set has 1057 observations. The available students' variables are listed in the following. *Pre-test*: Female, Far-away resident (indicator for residence in the provinces of Massa-Carrara and Grosseto or in a province out of Tuscany), Type of high school (Scientific, Humanities, Technical, Other), High school irregular career (indicator for age at high school diploma > 19), High school grade (from 60 to 100, centered at 80). *Test*: Total test score, Partial test scores (Logic, Reading, Mathematics), Test passed (indicator for total test score ≥ 9). *University career*: Delay in enrollment (indicator for being enrolled one or more years after high school diploma), Degree program (Management, Economics, Tourism, Marketing and Statistics), Credits gained during the first year (from 0 to 60), Second year enrollment at the School of Economics.

The test was passed by 853 candidates (80.7%). The test result is not mandatory for enrollment, but it influences the probability of enrollment: the enrollment rates were 65.3% overall, 67.9% for candidates who passed the test and 54.4% for candidates who did not pass the test.

The analysis is based on 690 students who took the test and then enrolled at the School of Economics. The sample distribution of gained credits after one year (December 2009) has a small percentage at the maximum (0.75% of freshmen gained 60 credits), but it has a peak at the minimum (23% of freshmen did not gain any credit). Therefore, the phenomenon is characterized by a relevant left censoring that needs to be accounted by the model. Moreover, the distribution of positive credits is quite irregular, showing peaks at 6, 15, 24, 36 and 45 credits. This pattern results from the paths followed by students, which can take exams weighting 6, 9 or 12 credits. The distribution of positive credits has a median of 30 and a mean of 29.8.

3 Modelling strategies

To disentangle direct and indirect effects of students background characteristics on the number of gained credits, the result of the admission test can be treated as an intermediate variable in a regression chain graph (Wermuth and Sadeghi, 2012). The specified chain graph model has tree blocks: (*i*) pre-test (exogenous) variables, (*ii*) standardized test scores (intermediate variables), and (*iii*) gained credits after one year (outcome). The test result could be summarized by the total score, but this would obscure some interesting aspects of the phenomenon: first of all, the three areas (Logic, Reading and Math) have different numbers of items; moreover, we wish to evaluate the relationships of each of the three partial scores with pre-test variables and the outcome. Therefore, we consider the Logic, Reading and Math scores as distinct variables, using standardized values to eliminate the effect of the different numbers of items. The three standardized partial scores are jointly regressed on pre-test covariates with a multivariate linear model (as compared to three separate regressions the multivariate regression yields the same point estimates but slightly different standard errors).

The model for regressing gained credits y_i on pre-test variables and test scores entails remarkable difficulties since, as noted in the previous Section, the distribution of y_i is quite irregular and has a large peak in zero. Therefore, conventional parametric models are not suitable. We tried two alternatives that we are going to outline in sequence, namely a binomial mixture model with concomitant variables (Grilli et al., 2013) and a hurdle model (Grilli et al., 2012)

3.1 Binomial mixture model with concomitant variables

The binomial mixture model is explicitly designed for a count variable with a fixed maximum, such as the number of gained credits y_i. This model assumes that the distribution $P(y_i)$ is defined by a finite mixture of conditional distributions $P(y_i \mid u_i)$, where u_i is a categorical latent variable taking values $k = 1, \ldots, K$ with prior probabilities $\pi_k = P(u_i = k)$, where $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

$$P(y_i) = \sum_{k=1}^{K} \pi_k P(y_i \mid u_i = k). \tag{1}$$

where the conditional distributions $P(y_i \mid u_i)$ are binomial with common number of trials t and component-specific probabilities of success θ_k:

$$P(y_i \mid u_i = k) = \binom{t}{y_i} \theta_k^{y_i} (1 - \theta_k)^{t - y_i}. \tag{2}$$

In order to exploit the covariates, we fit a *Concomitant variable mixture model* (Dayton and Macready, 1988), where the component probabilities of the finite mixture vary across subjects according to a vector of covariates $\mathbf{z}_i$ (usually including a constant for the intercept):

$$P(y_i \mid \mathbf{z}_i) = \sum_{k=1}^{K} \pi_{k|\mathbf{z}_i} P(y_i \mid u_i = k), \tag{3}$$

where $\pi_{k|\mathbf{z}_i} = P(u_i = k \mid \mathbf{z}_i)$, with $\pi_{k|\mathbf{z}_i} > 0$ and $\sum_{k=1}^{K} \pi_{k|\mathbf{z}_i} = 1$ for any subject i. Such constraints are satisfied by any model for nominal variables, like the multinomial logit model:

$$\pi_{k|\mathbf{z}_i} = \frac{\exp(\mathbf{z}_i'\boldsymbol{\beta}_k)}{\sum_{l=1}^{K}\exp(\mathbf{z}_i'\boldsymbol{\beta}_l)}, \quad k = 1,\ldots,K, \tag{4}$$

with $\boldsymbol{\beta}_1 = 0$ for model identifiability. Therefore, the prior probabilities of class membership depend on the covariates $\mathbf{z}_i$ through a non-linear function.

For given K, the parameters can be estimated with Maximum Likelihood using the EM algorithm (McLachlan and Peel, 2000).

3.2 Hurdle model

A hurdle or two-part model (Cameron and Trivedi, 2005) can be used to account for the large proportion of students (23%) gaining no credits ($y_i = 0$). Such a proportion should not be regarded as a nuisance, but as a key feature of the phenomenon since those students failed to begin the university career and, indeed, most of them dropped out.

The hurdle model has two components: a logit model for the probability of gaining at least one credit $P(y_i > 0 \mid \mathbf{z}_i)$, and a linear model for the expected number of gained credits $E(y_i \mid y_i > 0, \mathbf{x}_i)$. The linear model is fitted on the subset of students who gained at least one credit. The covariates of the two sub-models, $\mathbf{z}_i$ and $\mathbf{x}_i$, are distinct in principle, but they can even be the same. Since no parametric distribution appropriately describes the pattern of gained credits, we avoid a parametric specification and estimate the parameters via OLS and then compute robust standard errors.

The linear model for positive credits should be regarded as an approximation of the relationship between the mean and the covariates, without trying to model the whole distribution. Indeed, the linear model does not put restrictions on the support of y_i, so that non-integer values and out-of-range values are possible. However, in this application such issues are not critical, since non-integer values are just a problem of rounding, whereas the predicted mean is always within the range [0,60].

References

1. Adelfio, G., Boscaino, G., Capursi, V.: Quantile regression on a new indicator for higher education performance. Working Paper, CNR Solar (2013) http://eprints.bice.rm.cnr.it/id/eprint/5181
2. Bianconcini, S., Cagnone, S.: A General Multivariate Latent Growth Model With Applications to Student Achievement. Journal of Educational and Behavioral Statistics **37**, 339–364 (2012)
3. Birch, E.R., Miller, P.W.: Student Outcomes At University In Australia: A Quantile Regression Approach. Australian Economic Papers, Wiley Blackwell **45**, 1–17 (2006)
4. Birch, E.R., Miller, P.W.: The influence of type of high school attended on university performance. Australian Economic Papers **46**, 1-17 (2007)
5. Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge (2005)

6. Chowdhury, M., Mallik, G.: How Important are Introductory Subjects in Advanced Economics Studies? Economic Papers: A journal of applied economics and policy **31**, 255–264 (2012)
7. Dancer, D.M., Fiebig, D.G.: Modelling Students at Risk. Australian Economic Papers **43**, 158–173 (2004)
8. Dayton, C. M., Macready, G. B.: Concomitant-Variable Latent-Class Models. Journal of the American Statistical Association **83**, 173–178 (1988)
9. Grilli, L., Rampichini, C., Varriale, R.: University admission test and students' careers: an analysis through a regression chain graph with a hurdle model for the credits. 46^{th} Scientific Meeting of the Italian Statistical Society. Rome, 20-22 June, (2012)
10. Grilli, L., Rampichini, C., Varriale, R.: Binomial mixture modelling of university credits. To appear in Communications in Statistics - Theory and Methods (2013)
11. Hoefer, P., Gould, J.: Assessment of Admission Criteria for Predicting Students' Academic Performance in Graduate Business Programs. Journal of Education for Business **75**, 225–229 (2000)
12. Mallik, G., Lodewijks, J.: Student Performance in a Large First Year Economics Subject: Which Variables are Significant? Economic Papers: A journal of applied economics and policy **29**, 80–86 (2010)
13. Maree, J.G., Pretorius, A., Eiselen, R.J.: Predicting success among first-year engineering students at the rand afrikaans university. Psychological Reports **93**, 399–409 (2003)
14. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
15. Mills, C., Heyworth, J., Rosenwax, L., Carr, S., Rosenberg, M.: Factors associated with the academic success of first year Health Science students Advances in Health Science Education **14**, 205–217 (2009)
16. Murphy, M., Papanicolaou, K., McDowell, R.: Entrance score and performance: A three year study of success. Journal of Institutional Research **10**, 32–49 (2001)
17. Murray-Harvey, R.: Identifying characteristics of successful tertiary students using path analysis. Australian Educational Researcher **20**, 63–81 (1993)
18. Smith, J., Naylor, R.: Schooling Effects on Subsequent University Performance: Evidence for the UK University Population'. Economics of Education Review **24**, 549–562 (2005)
19. Wedman, I.: The Swedish Scholastic Aptitude Test: Development, Use, and Research. Educational Measurement: Issues and Practice **13**, 5–11 (1994)
20. Vermunt, J. K., Magidson, J.: LG-Syntax users guide: Manual for Latent GOLD 4.5 Syntax Module. Statistical Innovations Inc., Belmont, MA (2008)
21. Wermuth, N., Sadeghi, K.: Sequences of regressions and their independences. Test **21**, 215–252 (2012)
22. Win, R., Miller, P.W.: The Effects of Individual and School Factors on University Students' Academic Performance. Australian Economic Review **38**, 1–18 (2005)