Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

Predictive Validity of High School Grade and other Characteristics on Students' University Careers using ROC Analysis

Laura Pagani[1] and Chiara Seghieri[2]

Abstract

The aim of this paper is to highlight the ability of high school grade and other students' characteristics (such as socio-economic status, high school of origin and so on) to predict success in university careers. To study the predictive capability we use data based on the records of almost 3200 freshmen who entered the University of Udine during the academic years 1992/93-1997/98. The structure of our data-base (students in high schools) suggests the use of logit multilevel models to relate a response variable (a binary variable that takes value 1 if the student gives at least one exam during the first academic year, 0 otherwise) to independent variables. Once the models are fitted, the estimated probabilities are used to construct an evaluation test to discriminate between two conditions: the student that can successfully enter the University (*positive* subject) and the student with lower abilities (*negative* subject). The evaluation test is obtained using the Receiver Operating Characteristic (ROC) analysis. If this test turns to be a valid measure of predicting success then it can be used to determine students' eligibility for admission (admission policy) or to guide students in their faculty choice.

1 Introduction

The principal aim of this paper is to highlight the ability of high school grades and other student characteristics (such as socio-economic status, high school of origin and so on) to predict success in university careers. In fact a valid measure of predicting success in college can be used to determine students' eligibility for admission (admission policy) or to direct students in their faculty choice.

The structure of our data set suggests the use of logit multilevel models (Goldstein, 1995; Snijders and Bosker, 1999) to relate four response or status variables (binary variables which take value 1 if the student gives respectively at least one,

[1] University of Udine - Department of Statistics - Via Treppo, 18 - 33100 Udine - Italy; pagani@dss.uniud.it

[2] University of Florence - Department of Statistics "G. Parenti" - Viale Morgagni, 59 - 50134 Florence - Italy; seghieri@ds.unifi.it

two, three or four exams during the first academic year, 0 otherwise) to independent variables related to students' characteristics (such as socio-economic status, high school grade, high school of origin and so on).

Once the models have been fitted we to apply the ROC analysis, extensively used in the medical applications, to evaluate the ability of the four different models/*tests*, to correctly discriminate between students with good and bad capabilities.

ROC methodology is based on statistical decision theory and was developed in the context of electronic signal detection and problems with radar in the early 1950s (Metz, 1986). By the mid-1960s, ROC plots had been used in experimental psychology and psychophysics. Following work in psychophysics by Green and Swets (1966) Leo Lusted (1971), a radiologist, suggested using ROC analysis in medical decision making in 1967.

Nowadays ROC curves are widely used in many areas to describe and compare the performance of diagnostic technology and diagnostic algorithms, for example: weather forecasting, information retrieval, medical imaging, material testing and aptitude testing (for students and workers).

2 Data

To fit the four multilevel models we use data based on the records of almost 3200 freshmen who entered the University of Udine during the academic years 1992/93-1997/98.

The students are enrolled in six faculties of the University of Udine (Agriculture, Engineer, Medicine, Arts, Economy, Science) and come from forty-two high schools located in Udine and the surrounding province.

Most of the variables used in the analysis come from a questionnaire the students are required to complete when they enter the university. The questionnaire collects information on the students characteristics such as age, gender, socioeconomic status, and so on.

In addition we have information about university career (numbers of exams passed, regularity of the studies,etc.) from the administrative archives of the university and data regarding some characteristics of the high schools the students attended (name and kind of school, final grade).

3 Statistical models

The structure of the data (students in high schools) and the kind of the response variables (binary) suggest the use of a logit multilevel model. To fit the multilevel models we consider two levels: the lower level units are the students ($i = 1, \ldots, 3243$) while the higher level units are their high school of origin ($j = 1, \ldots, 42$). We use four different kind of binary response variables and consequently four different logit multilevel models. The response variables take value one if the student gives respectively at least one, two, three or four exams during the first year of his/her university career, 0 otherwise. Therefore all the responses can be seen as indicators

of students' success at the university and we assume that this increases when the number of exams passed during the first academic year increases. The hypothesis of the logit multilevel models are:

$$Y_{ij} \sim Binomial(1, \pi_{ij})$$

$$logit(\pi_{ij}) = \gamma_0 + \sum_{h=1}^{H} \gamma_h X_{hij} + U_{0j}$$

$$U_{0j} \sim N(0, \sigma_0^2)$$

We fit several models with different sets of explanatory variables using *MlwiN* (Goldstein et al., 1998) a standard software for multilevel models. The definition of the explanatory variables used in the final models are reported in Table 1.

Table 1: Explanatory variables definition.

Gender	1= male; 0= female
Grade	the standardized mark obtained at the end of high school
Age	the age of the student at his first academic year
Arts	1= the student attended the faculty of Arts; 0= otherwise
Engineering	1= the student attended the faculty of Engineering; 0= otherwise
Medicine	1= the student attended the faculty of Medicine; 0= otherwise
Science	1= the student attended the faculty of Science; 0= otherwise
School	1= the student attended a high school specialized in classic or scient. subjects; 0= otherwise
Father's job1	1= the student's father is a worker; 0= otherwise
Father's job2	1= the student's father is a trader; 0= otherwise
Father's job3	1= the student's father is a manager; 0= otherwise

The results about the final models are shown in Table 2 in which Model 1 stands for the logit model where the response variable indicates if the student passes at least one exam during the first academic year or not; Model 2 stands for the logit model where the binary response variable indicates if the student passes at least two exams during the first academic year or not; response variable in Model 3 indicates if the student passes at least three exams during the first academic year or not; and finally the response variable in the Model 4 if the student passes at least four exams during the first academic year or not.

The main results for all models are:

- the mark obtained at the end of high school has a positive effect on the students' university performance (the higher the grade, the higher is the probability of giving exams in the first academic year);
- the high positive influence of the kind of high school (students attended a high school specialized in classic or scientific subjects perform better than students attending others kinds of high schools);

Table 2: Parameter estimations for the four final models.

Parameter	**Coefficients (S.E.) Model 1**	**Coefficients (S.E.) Model 2**	**Coefficients (S.E.) Model 3**	**Coefficients (S.E.) Model 4**
Constant	1.523 (0.249)	1.182 (0.261)	0.014 (0.205)	0.239 (0.322)
Gender	0.410 (0.063)	-0.032 (0.063)	0.010 (0.063)	0.042 (0.069)
Grade	0.332 (0.029)	0.358 (0.029)	0.468 (0.029)	0.478 (0.035)
Age	-0.082 (0.012)	-0.078 (0.013)	-0.038 (0.009)	-0.065 (0.016)
Arts	0.410 (0.111)	0.597 (0.114)	0.808 (0.115)	0.586 (0.100)
Engineering	0.338 (0.074)	0.373 (0.073)	0.193 (0.069)	0.199 (0.081)
Medicine	-	-	-	0.573 (0.211)
Science	-0.661 (0.096)	-0.455 (0.099)	-0.512 (0.100)	-0.466 (0.126)
School	0.631 (0.100)	0.665 (0.112)	0.856 (0.126)	0.555 (0.121)
Father's job1	0.178 (0.064)	0.140 (0.063)	-	-
Father's job2	-	-	-	-0.183 (0.075)
Father's job3	-	-	-	-0.245 (0.114)
Arts*school	-0.434 (0.189)	-0.490 (0.169)	-0.704 (0.153)	-
Arts*grade	-	-	0.474 (0.152)	0.856 (0.126)
Medicine*grade	-	-	-	0.618 (0.232)
$\hat{\sigma}_0^2$	0.033 (0.016)	0.054 (0.022)	0.096 (0.031)	0.073 (0.028)

Note: - indicates that the parameter is not statistically significant in that model; * indicates interaction.

- the significant effects of the kind of faculty the students attend, i.e. students attending the faculty of Science have a lower performance than the students attending others faculties. As there are no significative interaction between high school grade and faculty or kind of high school and faculty this effect can be explained with the fact that the faculty of Science is more difficult than other faculties;

- the negative effect of the age at which the student starts the university (the older is the student the lower is the probability of giving exams);

- the effect of father's job (proxy of the socioeconomic state) (except in Model 3);

- the random parameter is significant so there are real differences between high schools.

The main difference between the models concerns the different effects of the father's job:

In the Model 3 the variables related to father's job have no effect, while in Model 1 and Model 2 the effect is positive (if the father is a worker the student has high probability of giving exams); in Model 4 the kind of jobs that are statistically significant are trader and manager, both with a negative effect. So the conclusion is that the student with low socioeconomic status performs better than the students with high status.

4 The ROC analysis

The next step of our analysis is to use the estimated probabilities, calculated using the logit multilevel models, to construct an evaluation test (a sort of diagnostic test) that can be used to discriminate between two conditions: the student that can successfully enter the University (*positive* subject) and the student with lower abilities (*negative* subject). If this test is a valid measure of predicting success in the university career then it can be used to determine students' eligibility for admission (admission policy) or to guide students in their faculty choice. A well known procedure used for this kind of problem is the Receiver Operating Characteristic (ROC) analysis. ROC analysis can help to quantify the accuracy, i.e. the ability of any test/measure to discriminate between the relevant alternative, states or characteristics of the subject. The accuracy of a test is expressed by its ability to correctly classify subject, in our case the students, that we know are *positive* or *negative*. Closely connected to the accuracy are the well known concepts of sensitivity and specificity. Sensitivity is the proportion of *positive* subjects that are correctly classified by the diagnostic test (true-positive rate), while specificity is the proportion of *negative* subjects that are correctly classified (true-negative rate). When we fix a set of decision thresholds or cup-points, used to classify the subjects as *positive* or *negative* based on test result, there is a set of pairs, one for each decision threshold, of sensitivity and specificity. Only the entire spectrum of sensitivity/specificity pairs provide a complete picture of test accuracy. This kind of analysis can be summarized by a ROC curve and plotted in a ROC graph. The ROC curves are sets of pairs of sensitivity and (1-specificity), i.e. the false-positive rate, and the ROC graph is the plot of these pairs (sensitivity versus 1-specificity). A test that perfectly discriminates would yield a *curve* that coincides with the left and top sides of the plot. While a test that is completely useless would give a straight line from the bottom left corner to the top right corner. When results from multiple tests are obtained the ROC plots can be graphed together. The relative position of the plots indicate the relative accuracies of the tests. A plot lying above and to the left of another plot indicates grater observed accuracy. Because different groups

Table 3: False positive rate (FPR) and True positive rate (TPR) at different cut points.

	Model 1		**Model 2**		**Model 3**		**Model 4**	
Cut points	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR
0.1	0.001	0.003	0.001	0.003	0.001	0.009	0.001	0.030
0.2	0.002	0.028	0.003	0.044	0.007	0.099	0.014	0.210
0.3	0.011	0.107	0.018	0.154	0.038	0.290	0.120	0.508
0.4	0.035	0.258	0.061	0.341	0.139	0.532	0.334	0.755
0.5	0.120	0.453	0.203	0.598	0.367	0.782	0.619	0.921
0.6	0.345	0.741	0.467	0.858	0.667	0.932	0.829	0.980
0.7	0.684	0.929	0.765	0.960	0.864	0.987	0.963	0.998
0.8	0.927	0.993	0.956	0.996	0.997	0.999	0.975	0.999
0.9	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000

of subjects selected at random from a population can yield different ROC plots, such sampling variability for a single ROC plot is often indicated by reporting the variance or constructing a confidence interval about a point or points on the ROC plot. The most common global measure to quantify the accuracy of a test is the area under the ROC plot. By convention, this area is always greater or equal 0.5. Values range between 1 (perfect separation of the test values of the two groups) and 0.5 (no apparent separation between the two groups). Direct statistical comparison of multiple tests is usually done by comparing the relative ROC plots using the area under the plot as an overall measure. After establishing the set of threshold values, in our case $0.1, 0.2, \ldots, 0.9$, we have calculated the true-positive rate and false-positive rate using the fitted probability obtained after adapting the four logit multilevel models (as shown in Table 3). For example a false-positive is relative to a student who has done no exams in the first year (response variable with value or status value = 0) and at the same time has an estimated probability that is higher than the fixed threshold value.

The values of the four areas of the ROC curves, relative to the four multilevel models, and their standard errors, calculated using the method suggested by Hanley and McNeil (1982) are presented in Table 4, while the corresponding ROC plots are shown in Figure 1. The areas are computed using a non-parametric method based on the trapezoidal rule.

Table 4: Areas under the ROC plots.

Model	Model 1	Model 2	Model 3	Model 4
Area (S.E.)	0.772 (0.009)	0.779 (0.009)	0.788 (0.007)	0.793 (0.009)

As we can see from Table 4 all the four evaluation tests (the estimated probabilities under the four models) have reasonable values of their areas, in particular Model 4 has the highest value one; so it discriminates better than the other tests between *positive* subjects and *negative* subjects, even if they are quite similar. However it seems that there is not a significant difference in accuracy of the four tests. But if we consider the area under a ROC curve as a measure of the ability of the test (in our case an evaluation test) to correctly classify the graduate as a *positive* subject or *negative* subject, our result is not completely satisfactory. In fact we can see the area under a ROC curve as the probability that the result of the evaluation test of a randomly selected *positive* subject will be greater than the result of the same test from a randomly selected *negative* subject. In our case this probability ranges from about 0.77 to 0.80. This means that we have omitted some important information about the students when estimating the parameters of the logistic multilevel models.

5 Conclusion

The analysis of the data about students of Udine using a logit multilevel model leads us to an evaluation test that is not completely satisfactory, suggesting that the problem of measuring the abilities of students is a complex phenomenon. So it

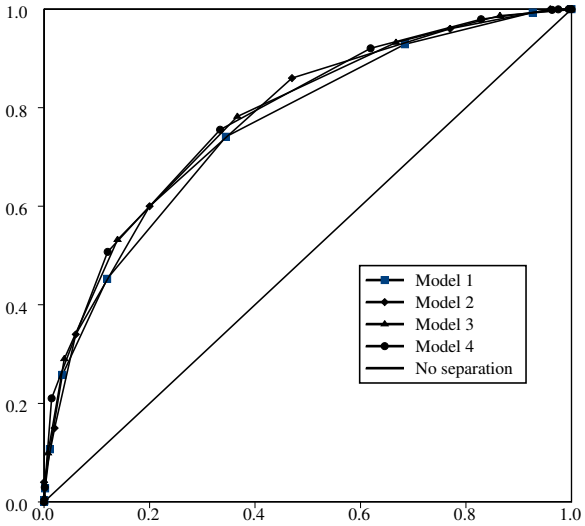

Figure 1: The four ROC plots.

is important to collect additional information about student characteristics (for example: performance during high school, more information about the socio-economic status) and to include some contextual variables (e.g., information about the high schools' teachers, available high schools' fund) in order to improve the performance of the logit multilevel model and therefore the accuracy of the evaluation test.

Acknowledgements

Authors thank the anonymous referees and the editors for very helpful comments and suggestions.

References

[1] Goldstein, H. (1995): *Multilevel Statistical Models*. Edward Arnold, London.

[2] Goldstein, H. et al. (1998): *A User's Guide to Mlwin*, Institute of Education, University of London.

[3] Green, D.M. and Swets, J.A. (1966): *Signal Detection Theory in Psychophysics*. New York: Wiley & Sons.

[4] Hanley, J.A. and McNeil, B.J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

[5] Lusted, L.B. (1971): Decision making studies in patient management. *N Engl J Med*, **284**, 416-424.

[6] Metz, C.E. (1986): Roc methodology in radiology imaging. *Invest Radiol*, **21**, 720-33.

[7] Snijders, T. and Bosker, R. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage Publications.