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

Leonardo Grilli    Carla Rampichini
Dipartimento di Statistica, Informatica, Applicazioni – Università di Firenze

Roberta Varriale
Istat - Roma
Outline

- Introduction
- Literature review
- Case study: performance of freshmen at the University of Florence
- Modelling strategies:
  - Regression chain graph
  - Hurdle model
  - Binomial mixture models with concomitant variables
- Discussion


pre-print at [http://local.disia.unifi.it/grilli/papers.htm](http://local.disia.unifi.it/grilli/papers.htm)
Predicting academic performance
(so important, so difficult...)

- Predicting students' academic performance is a key step in order to improve the efficiency of university systems.
- Universities rely on info about the high school career, e.g. type of school and various measures of proficiency.
- However, the results at high school are *not fully appropriate* to predict the academic performance:
  - mismatch between competencies evaluated at high school and competencies required for a given degree program.
  - heterogeneity in the criteria for awarding marks (variability across types of schools and across geographical regions).
- A partial remedy: *pre-enrolment assessment test* tailored on the needs of each degree program (lack of commonly accepted guidelines and shortage of empirical evidence about the predictive ability).
A look at the literature

- 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)** Identifying characteristics of successful tertiary students using path analysis. *Australian Educational Researcher*
  
  
  
  - **Murphy et al. (2001)** Entrance score and performance: A three year study of success. *Journal of Institutional Research*
  
  - **Maree et al. (2003)** Predicting success among first-year engineering students at the rand afrikaans university. *Psychological Reports*
  
  - **Dancer and Fiebig (2004)** Modelling Students at Risk. *Australian Economic Papers*
A look at the literature (cont.)

- **Win and Miller (2005)** The Effects of Individual and School Factors on University Students’ Academic Performance. *Australian Economic Review*

- **Smith and Naylor (2005)** Schooling Effects on Subsequent University Performance: Evidence for the UK University Population’. *Economics of Education Review*

- **Birch and Miller (2006)** Student Outcomes At University In Australia: A Quantile Regression Approach. *Australian Economic Papers*

- **Mills et al. (2009)** Factors associated with the academic success of first year Health Science students. *Advances in Health Science Education*

- **Mallik and Lodewijks (2010)** Student Performance in a Large First Year Economics Subject: Which Variables are Significant? *Economic Papers*

- **Bianconcini and Cagnone (2012)** A General Multivariate Latent Growth Model With Applications to Student Achievement. *Journal of Educational and Behavioral Statistics*

Freshmen at the University of Florence: Pre-enrolment test

- In a.y. 2008/2009, the School of Economics of the University of Florence introduced a **compulsory pre-enrolment test** to evaluate the background of the students.

- 40 multiple-choice items covering 3 areas: **Logic** (12 items), **Reading** (10 items) and **Mathematics** (18 items)
  - for each item, 1 out of 5 alternatives is correct
  - scoring system: 1 if correct, 0 if blank, -0.25 if wrong

- The test has a main edition in September and several supplementary editions later.

- Candidates with a total score lower than 9 are advised against enrolment: they could still enrol, but they could take examinations only after ‘passing’ the test during one of the later editions.

[www.economia.unifi.it/vp-586-test-di-accesso.html](http://www.economia.unifi.it/vp-586-test-di-accesso.html)
Freshmen at the University of Florence: Administrative data

We analyse data on 690 freshmen of the School of Economics in Florence in a.y. 2008/2009, considering the students who took the pre-enrolment test in September 2008.

The data set is obtained by merging data collected at the test and administrative data.

- Pre-test:
  - Gender
  - High school type (Scientific, Humanities, Technical, Other)
  - High school grade (from 60 to 100, centered at 80)
  - High school irregular career (indicator for age at diploma > 19)
  - Far-away resident

- Test: Partial test scores (Logic, Reading, Mathematics)

- Post-test: Credits gained during the first year (from 0 to 60)
Regression chain graph

- Formal representation of prior knowledge and working hypotheses
- Effective tool to represent model and results
- Disentangling **direct** and **indirect** effects


**Step 0:** collect variables into ordered blocks

**Step 1:** Regress the three (standardized) test scores on pre-test covariates

**Step 2:** Regress gained credits on test scores *and* pre-test covariates
Modelling gained credits

Gained credits after one year are in the interval [0,60]

Exams have different credits (multiples of 3), usually 6, 9 or 12 → the distribution of gained credits is quite irregular!

- peak at the minimum (23% of freshmen did not gain any credit)
- the distribution of positive credits is quite irregular, showing peaks at 6, 15, 24, 36 and 45 credits

- Standard parametric models are not suitable → solutions
  1. Hurdle (or two-part) model
  2. Binomial mixture model
  3. Quantile regression
Modelling gained credits

solution #1: hurdle model

- Our ‘hurdle’ or ‘two-part’ model has two components:
  1. A logit model for the probability of gaining at least one credit
     \[ P(y_i > 0 \mid z_i) \]
  2. A linear model for the expected number of gained credits
     (fitted on the subset of students who gained at least one credit)
     \[ E(y_i \mid y_i > 0, x_i) \]

- The covariates of the two sub-models are distinct in principle, but
  they can even be the same

- No parametric distribution is suitable for the distribution of credits:
  to avoid distributional assumptions, we estimate the parameters of
  the linear model via OLS and use robust standard errors
An arrow is traced when the regression coefficient is statistically significant.
Main findings

- Even controlling for pre-test covariates, the standardized partial test scores have a significant effect on credits:
  - higher **score on Reading** → a higher probability of gaining credits $P(Y>0)$
  - higher **score on Math** → higher expected number of gained credits $E(Y)$

- The **score on Logic** does not help predict the gaining of credits when the scores on Reading and Math are known.

- The effects of pre-test covariates are mediated by the test scores, with the notable exceptions of
  - **high school grade** (positive effect)
  - **irregular career** (negative effect)

Proxies of abilities and attitudes of the students that are not fully captured by the pre-enrolment test
Modelling gained credits
solution #2: binomial mixture model

- Response (count): \( y_i = \frac{\text{credits}_i}{3} \)
- Distribution: \( y_i \sim Bin(t=20, \theta_k) \)
- Mixture components represented by the categorical random variable \( u_i \), taking values \( k = 1, \ldots, K \) with prior probabilities \( \pi_k \)

\[
P(y_i) = \sum_{k=1}^{K} \pi_k P(y_i | u_i = k)
\]

where all the conditional distributions \( P(y_i | u_i) \) are binomial with common number of trials \( t \) and component-specific probabilities of success \( \theta_k \)

\[
P(y_i | u_i = k) = \binom{t}{y_i} \theta_k^{y_i} (1 - \theta_k)^{t-y_i}
\]

Credits range from 0 to 60 in blocks of 3

Binomial mixture model: fit without covariates

- Given $K$ the model can be fitted with ML using the EM algorithm – we used Latent Gold (Vermunt & Magidson, 2008)
  - we later replicated the analysis with the R package `flexmix`: code and data available at http://local.disia.unifi.it/grilli

- Selection of the number of components $K$ with BIC, bootstrap LRT and EM test Li and Chen (2010) ⇒ they all select $K=5$

| Component | $\pi_k$ | $\theta_k$ | $E(\text{credits}|u = k)$ | $P(\text{credits} = 0|u = k)$ | $P(\text{credits} \geq 54|u = k)$ |
|-----------|--------|--------|-----------------|-----------------|-----------------|
| 1         | 0.22   | 0.00   | 0               | 1.000           | 0.000           |
| 2         | 0.15   | 0.14   | 9               | 0.045           | 0.000           |
| 3         | 0.25   | 0.39   | 23              | 0.000           | 0.000           |
| 4         | 0.28   | 0.65   | 39              | 0.000           | 0.012           |
| 5         | 0.10   | 0.85   | 51              | 0.000           | 0.381           |

- The first component (size 0.22) is almost degenerate in 0, accounting for the excess zeroes in the sample distribution:
  
  $P(\text{credits} = 0) \approx 0.22 \times 1.000 + 0.15 \times 0.045 = 0.230$

  (equal to the sample proportion)

- In general, the fit is satisfactory in all the support
Binomial mixture model: fit with covariates (concomitant var.)

- In a *concomitant variable* specification the covariates affect the component probabilities $\pi_k$ (Dayton and Macready, 1988)

\[
P(y_i \mid z_i) = \sum_{k=1}^{K} \pi_{k \mid z_i} P(y_i \mid u_i = k)
\]

\[
\pi_{k \mid z_i} = P(u_i = k \mid z_i) = \frac{\exp(z_i^T \beta_k)}{\sum_{l=1}^{K} \exp(z_i^T \beta_l)}
\]

<table>
<thead>
<tr>
<th>Latent class</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>0.38</td>
</tr>
<tr>
<td>4</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>0.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binomial probability $\theta_k$</th>
<th>0.00</th>
<th>0.15</th>
<th>0.38</th>
<th>0.64</th>
<th>0.85</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial logit model* for $\pi_k$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.03</td>
<td>0.22</td>
<td>0.96</td>
<td>-0.57</td>
<td>0.000</td>
</tr>
<tr>
<td>HS Technical/other</td>
<td>-0.63</td>
<td>0.18</td>
<td>-0.40</td>
<td>-1.43</td>
<td>0.013</td>
</tr>
<tr>
<td>HS irregular career</td>
<td>-0.39</td>
<td>-0.79</td>
<td>-3.08</td>
<td>-0.57</td>
<td>0.012</td>
</tr>
<tr>
<td>HS grade</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.06</td>
<td>0.12</td>
<td>0.000</td>
</tr>
<tr>
<td>Logic (std score)</td>
<td>-0.11</td>
<td>0.21</td>
<td>0.26</td>
<td>-0.34</td>
<td>0.052</td>
</tr>
<tr>
<td>Reading (std score)</td>
<td>0.51</td>
<td>0.33</td>
<td>0.29</td>
<td>0.79</td>
<td>0.001</td>
</tr>
<tr>
<td>Math (std score)</td>
<td>-0.09</td>
<td>0.00</td>
<td>0.25</td>
<td>1.10</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Effect of test scores on \( E(\text{credits}) \)

*Expected number of gained credits* by test scores

(the value in zero refers to the *baseline* student: HS Scientific/Humanities, HS grade at midpoint, regular career)
Effect of test scores on $P(\text{first class})$

*Probability of belonging to the zero-credit latent class* by test scores
(the value in zero refers to the *weak* student: HS Technical/other, HS grade at minimum, irregular career)
The hurdle model (logit+linear) is **simple** and it may be used for studying associations.

In our application it yields the same findings as the binomial mixture model about the pre-enrolment test, namely:

- a **low Reading score** is related to a **difficult start-up** of the university career.
- a **low Math score** is related to a **slow progression**, likely for problems encountered in Math and Statistics (which are often the hardest exams).

However, the hurdle model should not be used for making predictions: unbounded response → **non-admissible predictions**, e.g. negative number of gained credits.
Can we really predict gained credits?

- The linear part of the hurdle model has R-squared = 0.24
- Binomial mixture model → Mean Absolute Error of prediction (10-fold cross-validation):
  - Null model: MAE = 15.7
  - Model with only background characteristics: MAE = 13.3 (-15%)
  - Model with background char. + test scores: MAE = 12.7 (-4%)
- In terms of prediction ability, the background characteristics give a relevant contribution
- The pre-enrolment test yields a further slight improvement, even if the predictive ability remains modest (students' careers are difficult to predict!)
Tests vs unstructured interviews

- The results about the predictive ability of pre-enrolment tests are not exciting... what about unstructured interviews?

- Apart from the high expense, unstructured interviews are ineffective in predicting the students performance:
    http://www.sas.upenn.edu/~danajd/interview.pdf

In addition to the vast evidence suggesting that unstructured interviews do not provide incremental validity, we provide direct evidence that they can harm accuracy. [...] interviewers are likely to feel they are getting useful information from unstructured interviews, even when they are useless. *Our simple recommendation for those who make screening decisions is not to use them.*
Thanks for your attention!
grilli@disia.unifi.it
rampichini@disia.unifi.it
varriale@istat.it