Tree-based methods for a flexible analysis of OECD-PISA 2015 data across countries

LEER Workshop on Education Economics
March 2017
Chiara Masci, Geraint Johnes & Tommaso Agasisti
Tree-based methods for a flexible analysis of PISA 2015 data across countries

Chiara Masci  
MOX, Department of Mathematics, Politecnico di Milano, Italy

Geraint Johnes  
LUMS, Lancaster University Management School, Lancaster, UK

Tommaso Agasisti  
School of Management, Politecnico di Milano, Italy
Outline

• Introduction & Motivation
• Research question
• Methodology: Multilevel regression trees & Boosting
• Data
• Results
• Conclusions
OECD PISA assesses 15 years-old students in reading, mathematics and science within 72 world countries, every three years (started in 2000).

Students and school principals have to fill out questionnaires:

- Student level variables
- School level variables
- ≫ hierarchical structure of the data

We focus on 9 world countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK and USA.
Research questions

- Which students’ characteristics are related to students’ performances and how?

- How to measure the school value-added?

- Which school factors are related to the school value-added?

- How are these factors interrelated?

- Do these relations change across countries?
The educational systems are complex, unknown and very heterogeneous around the world.

Weak points of the statistical methodologies:
- Parametric assumptions on the educational production function
- Lack of interaction between different level variables

Relaxing the parametric assumptions
Allowing interaction

TREE-BASED METHODS
(MACHINE LEARNING APPROACH)
• **Statistical approaches**
  • start by *assuming an appropriate data model* and parameters for this model are then estimated from the data
  • focus on questions such as *what model will be postulated* (e.g. are the effects additive, or are there interactions?), how the response is distributed, and whether observations are independent.

• **ML methods**
  • use an algorithm to *learn the relationship* between the response and its predictors (Breiman 2001)
  • Assume that the *data-generating process* (in the case of ecology, nature) is *complex and unknown*, and try to learn the response by observing inputs and responses and finding dominant patterns.
Methodology

Two-stage analysis:

• 1° stage:
  • *Mixed-effects regression trees (RE-EM trees)* to estimate (i) the *students’ characteristics* associations to students’ performances and (ii) *school value-added*.

• 2° stage:
  • *Regression trees and Boosting* to (Inter)relate school value-added to *schools’ characteristics*. 
Tree-based methods for regression and classification involve stratifying or segmenting the predictor space into a number of simple regions. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs.

\[ y_i = f(x_{i1}, \ldots, x_{ip}) + \epsilon_i \quad i = 1, \ldots, N \]

\[ p = \text{number of predictors} \]

The predictor space is divided into J distinct and non-overlapping regions - high dimensional rectangles (or boxes) - where the goal is to minimize the Residual Sum of Square (RSS).
Introduction to Regression trees: An example

Outcome variable: Student test score

Average score of students that are in this branch
1° stage: Mixed-effects Regression trees (RE-EM trees)

Two-level regression trees: students (level 1) nested within schools (level 2)

\[ y_{ij} = f(x_{1ij}, ..., x_{ pij}) + b_j + \varepsilon_{ij} \]

Where \( f(x_{1ij}, ..., x_{ pij}) \) is the partition of the covariate space
\( y_{ij} \) is the PISA test score of student i within school j
\( (x_{1ij}, ..., x_{ pij}) \) are the p covariates at student level
\( b_j \) is the school value-added
\( b_j \sim N(0, \sigma^2_b) \) \( \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon) \) \( \Rightarrow PVRE = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\varepsilon} \)
2nd Stage: Regression trees
Advantages and Disadvantages

- **Advantages:**
  - Trees can be displayed *graphically* and are easily interpretable
  - Trees do not force any type of *functional relationship* between the outcome variable and the covariates
  - Trees can easily handle *qualitative predictors*
  - Trees allow *interaction* between the variables

- **Disadvantages:**
  - Trees generally suffer from *high variance*
  - Trees are very sensitive to *outliers*
2nd Stage: Regression trees and Boosting

Regression trees suffer from **high variance** and they are really sensitive to **outliers**

- **Boosting**: sequential, stepwise procedure
  - Trees are grown *sequentially*
  - Boosting is a numerical optimization technique for minimizing the residual function by adding, at each step, a new tree that best reduces the residual function (‘*Functional gradient descendent*’)

Data: Student and school levels variables

• **Student level variables:**
  
  • *dummy/categ*: gender, video games, sport, immigrant status, parents’ education level
  
  • *continuous*: maths test score, socio-economical index, time of homework, self motivation, self belonging, cooperation in class, anxiety, teacher/parents support perception, cultural possession, home educational resources

• **School level variables:**
  
  • *dummy/categ*: management, private, inadequacy of materials/struct
  
  • *continuous*: school size, computer/stud ratio, teacher/stud ratio, % disadvantaged/special needs students, % funds from govern, students truancy, teachers absenteeism
Data: Sample size and average mathematics students’ score across countries

<table>
<thead>
<tr>
<th>Country</th>
<th># Students</th>
<th># Schools</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>14,530</td>
<td>758</td>
<td>481.587</td>
</tr>
<tr>
<td>Canada</td>
<td>20,058</td>
<td>759</td>
<td>505.021</td>
</tr>
<tr>
<td>France</td>
<td>6,108</td>
<td>252</td>
<td>496.997</td>
</tr>
<tr>
<td>Germany</td>
<td>6,504</td>
<td>256</td>
<td>509.170</td>
</tr>
<tr>
<td>Italy</td>
<td>11,583</td>
<td>474</td>
<td>500.235</td>
</tr>
<tr>
<td>Japan</td>
<td>6,647</td>
<td>198</td>
<td>532.66</td>
</tr>
<tr>
<td>Spain</td>
<td>6,736</td>
<td>201</td>
<td>491.361</td>
</tr>
<tr>
<td>UK</td>
<td>14,157</td>
<td>550</td>
<td>490.765</td>
</tr>
<tr>
<td>USA</td>
<td>5,712</td>
<td>177</td>
<td>467.383</td>
</tr>
</tbody>
</table>
1° stage results: RE-EM trees across countries
How much of the total variability can we explain?

<table>
<thead>
<tr>
<th>Country</th>
<th>$\sigma^2_e$</th>
<th>$\sigma^2_b$</th>
<th>PVRE</th>
<th>PV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>0.690</td>
<td>0.125</td>
<td>15.41%</td>
<td>33.59%</td>
</tr>
<tr>
<td>Canada</td>
<td>0.724</td>
<td>0.143</td>
<td>16.49%</td>
<td>29.93%</td>
</tr>
<tr>
<td>France</td>
<td>0.464</td>
<td>0.419</td>
<td>47.47%</td>
<td>55.28%</td>
</tr>
<tr>
<td>Germany</td>
<td>0.525</td>
<td>0.437</td>
<td>45.44%</td>
<td>50.17%</td>
</tr>
<tr>
<td>Italy</td>
<td>0.568</td>
<td>0.395</td>
<td>41.04%</td>
<td>45.57%</td>
</tr>
<tr>
<td>Japan</td>
<td>0.510</td>
<td>0.437</td>
<td>46.13%</td>
<td>50.32%</td>
</tr>
<tr>
<td>Spain</td>
<td>0.706</td>
<td>0.068</td>
<td>0.08%</td>
<td>30.11%</td>
</tr>
<tr>
<td>UK</td>
<td>0.695</td>
<td>0.162</td>
<td>18.97%</td>
<td>32.51%</td>
</tr>
<tr>
<td>USA</td>
<td>0.689</td>
<td>0.132</td>
<td>16.15%</td>
<td>33.45%</td>
</tr>
</tbody>
</table>
1° stage results: Tree of the fixed-effects in Italy
1° stage results: Tree of the fixed-effects in UK

```
ESCS < 0.2617
  ├── disc_climate < -0.3975
  │    └── motiv > 0.1232
  │         ├── -0.3742
  │         │    └── -0.2989
  │         │        ├── -0.1007
  │         │        │    └── 0.2385
  │         │        └── anxtest <= -0.04505
  │         │                └── -0.1171
  │         │                   └── 0.2385
  │         └── anxtest >= -0.042
  │              └── disc_climate <= -0.5505
  │                   ├── -0.1171
  │                   │    └── 0.1542
  │                   └── 0.4123
```

POLITECNICO MILANO 1863
2° stage results: Regression trees & Boosting
How to read results of Boosting

For each country, applying regression trees and Boosting, we obtain:

- The ranking of **covariates’ importance** in explaining the response variable (school value-added)
- The Percentage of Explained Variability (PVE) by the model
- Single **partial plots** of each covariate and the response (to explore the marginal effect)
- **Joint partial plots** of couples of covariates and the response

<table>
<thead>
<tr>
<th></th>
<th>Australia</th>
<th>Canada</th>
<th>France</th>
<th>Germany</th>
<th>Italy</th>
<th>Japan</th>
<th>Spain</th>
<th>UK</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVE</td>
<td>40.36%</td>
<td>28.09%</td>
<td>59.13%</td>
<td>53.08%</td>
<td>28.09%</td>
<td>30.87%</td>
<td>14.15%</td>
<td>39.12%</td>
<td>35.81%</td>
</tr>
</tbody>
</table>
2° stage: Variables importance in Italy

Variables importance in Italy

- % disadvantage students
- % parents speak teachers
- # students
- % students special needs
- ratio computers-stud
- students truancy
- ratio stud-teachers
- % parents in school govern
- infrastructures inadequacy
- ratio stud-teachers5

Relative influence
2° stage: Variables importance in UK

Variables importance in England

- % disadvantage students
- % students special needs
- Ratio computers-stud
- # students
- % funds given by the govern
- Stud no respect teachers
- Stud admit record
- % parents speaking with teach
- Ratio stid-teachers5
- Students truancy

Relative influence
2° stage results: Single partial plots in Italy

Partial Dependence on % of disadvantaged students

Partial Dependence on school size

Partial Dependence on % parents speaking with teachers

Partial Dependence on % students with special needs
2° stage results: Single partial plots in UK

- Partial Dependence on % of disadvantaged students
- Partial Dependence on % students with special needs
- Partial Dependence on Ratio computers/students
- Partial Dependence on school size
2° stage results: Joint partial plots in Italy
2° stage results: Joint partial plots in UK
Conclusions

• Tree-based methods (ML approach) are a useful tool to analyse worldwide education systems, standing on their unknown and complex nature.

• It is possible to identify certain students’ and schools’ characteristics that are associated to students’ and schools’ performances in many countries, but often, they interact with responses in different ways.

• It is therefore worth to consider interactions among predictors.
References

• James, G., Witten, D., Hastie, R., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: Springer


THANKS FOR YOUR ATTENTION