Technology Reviews

- PISAnalyticTool

Nixi Wang, Hongbin Qu, Luyu Xu, Shenghao Xie

Introduction

PISA

International assessment of 15-year-old students' capabilities in three subjects and a range of factors of interest.

- Year 2015, stratified and clustered data across over 70 countries
  519,334 student cases, 921 variables on student-level
  17,908 schools, 273 variables on school-level

World Bank

Gender parity index (GPI) in Multiple Indicator Cluster Survey and Urban Informal Settlement Survey

- 264 countries

Inquiry: Educational equity across countries

Package requirements

- Retrieving sav and csv files
- Visual and statistic descriptives of achievement scores, student/family characteristics, attitudes and learning strategies, etc.
- Modeling schooling outcomes differentiated in gender by contributing characteristics

Data Retrieving

Data Retrieving Tools Evaluation: savReaderWriter

- A cross-platform Python interface to the IBM SPSS Statistics Input Output Module.
- Read or Write SPSS system files (.sav, .zsav).
- SPSS is short for Statistical Package for the Social Sciences.
- Created for the management and statistical analysis of social science data.
- Statistics Program, Modeler Program, Text Analytics for Surveys Program, Visualization Designer

In [1]:

```
import pandas as pd
import numpy as np
import savReaderWriter as spss

file = "data\CY6_MS_CMB_TCH_QQQ.sav"

records = []
with spss.SavReader(file) as reader:
    print(file)
    for line in reader:
        records.append(line)
df = pd.DataFrame(records)
df.head()
```

```
data\CY6_MS_CMB_TCH_QQQ.sav
```

Out[1]:

	0	1	2	3	4	5	6	7	8	9	...	247	248	249	250	251	252	253	254	
0	36.0	b'AUS'	3600001.0	3605909.0	NaN	b'06MS'	b'003600'	3600.0	b'AUS0421'	b'0360000'	...	9.0	9.0	9.0	9.0	9.0	9.0	99.0000	99.0000	1.15
1	36.0	b'AUS'	3600001.0	3607746.0	4.0	b'06MS'	b'003600'	3600.0	b'AUS0421'	b'0360000'	...	9.0	9.0	9.0	9.0	9.0	9.0	99.0000	99.0000	1.15
2	36.0	b'AUS'	3600001.0	3609742.0	5.0	b'06MS'	b'003600'	3600.0	b'AUS0421'	b'0360000'	...	3.0	0.0	2.0	2.0	3.0	2.0	0.2810	-0.8246	1.15
3	36.0	b'AUS'	3600001.0	3612774.0	5.0	b'06MS'	b'003600'	3600.0	b'AUS0421'	b'0360000'	...	0.0	1.0	0.0	0.0	0.0	0.0	0.4777	0.4838	1.15
4	36.0	b'AUS'	3600001.0	3603132.0	4.0	b'06MS'	b'003600'	3600.0	b'AUS0421'	b'0360000'	...	9.0	9.0	9.0	9.0	9.0	9.0	99.0000	99.0000	1.15

5 rows × 257 columns

Data Processing

Merge data:
Pandas

- Short for “panel data”
- An open-source Python Library providing high-performance data manipulation and analysis tool
- load, prepare, manipulate, model, and analyze data
- Providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive

Data Processing: Numpy vs Scipy

Numpy:

- A basic library for mathematical or numerical calculation
- Basic operations: indexing, sorting, reshaping, basic elementwise functions
- A non-vectorized operation will typically run slowly, while vectorization may increase memory complexity

Scipy:

- Builds on the NumPy array object
- Supports linear algebra, integration, FFT, ODE solvers and others
- Contains more modules and has better performance than Numpy

Functions for data

- PISA: We are interested in four terms: differences of two genders, 10% score, mean and 90% score. Use a tuple to store them: (differences, 10% score, mean, 90% score)
- WB: Get the value of a given variable in world bank data for different countries

- An example of HLM model proposed would be:

$$y_{ij} = \beta^T x_{ij} + b_j^T x_{ij} + \varepsilon_{ij}$$

Where

y_{ij} = DiffScores in gender for school i in country j

x_{ij} = SchoolClimate, FamilyWealth, StudentInterests, SchoolResources, SchoolType, ESCS, CEI, %FemaleTeachers, log(StudentsPerClassTeacher), log(TotalEnrollmentFemale)

HLM package overview

- Interacting with R, lmer 4 package

```
%load_ext rpy2.ipython
```

- Use Jupyter with the IR Kernel - combining Python and R language
- PyMC3:
  - Bayesian modeling
  - Depends on Numpy, Scipy, Pandas, Matplotlib
  - GLM subcomponents depend on patsy
- statsmodels
  - using familiar R-style formulas

Appeal: both well-documented, statsmodels has more active support group (google group)

Data visualization

Package: Bokeh

- A Python interactive visualization library that targets modern web browsers for presentation.
- Elegant, concise construction of novel graphics in the style of D3.js
- Easily create interactive plots, dashboards, and applications.

Interactive plots example

Bokeh vs. Plotly

Pros:

- Publication quality
- Embed your visualizations in applications
- Flexible, capable, pythonic
- Open source

Cons:

- Fancier than matplotlib but more work

Pros:

- Built-in 3d plots
- compatibility with number of different languages
- Simple syntax

Cons:

- Community version
- Upper limit on API calls and coloring options

Challenges

- New technology for all team members, may take more time to learn functions before application
- Examples are not enough online
- We need to think about the layout of the interface.
- Alternatives with ‘ggplot’ and ‘matplotlib’