GitHub Integration

I've created a git repository for my existing code. You can view all existing and future codes from github.com/berkanteber/cs210project.

I've also created a project page on GitHub. It has still some work to do, but you can reach it from berkanteber.github.io/cs210project.

April 26, 2017 / Leave a comment

Part 1 Overview

(Code: Part1Overview.pdf)

In the first part of the project, I've examined the nation-wide education and R&D data with GDP per capita and GINI index.

I've used the data on education spendings and PISA scores from OECD, and I've used R&D, GDP and GINI index data from World Bank.

Then, I've merged all data as follows:

- for education data, I used the same-year data,
- for PISA scores, I used the latest data,
- for other data, I used the same-year data if possible; then I first looked at the following year's data and then the previous year's data.

Finally, I've eliminated the countries with missing data.

As a result, I've come up with a dataset with 26 countries with the attributes below:

1. Primary to Non-tertiary Education (% of GDP)
2. Public Tertiary Education (% of GDP)
3. Private Tertiary Education (% of GDP)
4. PISA Scores (Reading)
5. PISA Scores (Math)
6. PISA Scores (Science)
7. R&D Expenditures (% of GDP)
8. Researchers in R&D (per million people)
9. High Technology Exports (% of manufactured exports)

10. Resident Patent Numbers (per 100 people)
11. Non-resident Patent Numbers (per 100 people)
12. GDP per capita
13. GINI index

Then, I've applied linear regression on these attributes and I've reached the following results:

- The spendings on primary to non-tertiary education spendings have no significant effect on PISA scores.
- Public tertiary education spendings and PISA math scores explain 54% of the variation in R&D expenditures.
- Public tertiary education spendings explain 34% of the variation in researchers in R&D.
- Education and R&D data are jointly insignificant to explain the variation in high technology data.
- Private tertiary education spendings and R&D expenditures explain 59% of the variation in resident patents.
- Private tertiary education spendings and R&D expenditures explain 63% of the variation in nonresident patents.
- Primary to non-tertiary education spendings, private tertiary education and non-resident patent number explains 58% of the variation in GDP per capita.
- PISA reading and math scores explain 73% of the variation in GINI index.

Some comments:

- The results saying that spendings on primary to non-tertiary education doesn't explain the variation in PISA scores was unexpected for me. Since, we see PISA scores explaining some other variables in other regressions, my interpretation is that PISA scores are not solely related to the spendings but the quality also.
- The result regarding to high technology exports was also unexpected.
- Public tertiary spendings explain the variation in R&D data, however they are insignificant to explain the variation in patent data. My interpretation is that since the variables are not truely independent R&D expenditure also contains information about the public tertiary education spendings.
- We can see that 58% of the variation in GDP per capita is explained by primary to non-tertiary education spendings, private tertiary education spendings and nonresident patent numbers. However, we cannot decide whether high GDP per capita increases the percentage of the education spendings or higher education spendings increase GDP per capita. I am also indecisive about this question. On the other hand, I think, the relation between nonresident patent numbers and GDP per capita is an indicator of brain drain (human capital flight).

April 24, 2017 / Leave a comment

Planning for the rest of the project

So far, I've analysed the data sets from OECD and World Bank about education, R&D, GDP per capita and GINI index. I have performed some linear regressions on these data sets. There have been both some significant and insignificant results.

My next move will be finalizing the interpretations on these data.

Then, I will move on to a new data set from the site below:

https://archive.ics.uci.edu/ml/datasets/Student+Performance#

The reason of that is to analyze the effect of individual differences on education.

Previously, I analyzed the nation-wide education data and it's effects nation-wide.

Now, it's time to focus on a more individualistic study.

To do so, I will have been analyzed education on wider aspects.

April 24, 2017 / Leave a comment

Regression (New)

I performed the regressions using statsmodels this time.
Since I have test statistics, I performed inference more accurately.

Below, you can find the resulting regressions for each possible output. I have chosen the least restrictive regression according to f-test. You can find the intermediary steps are intermediary steps in the code.

Regressions:

• pryntry on pisa_read: non-significant
• pryntry on pisa_math: non-significant
• pryntry on pisa_science: non-significant

• pisa_math and try_public on rd_exp: significat, R-squared = 0.543
• try_public on pisa_researchers_rd: significant, R-squared = 0.342
• pisa_read, pisa_math, pisa_science, pryntry, try_public, try_private, rd_exp, researchers_rd on high_tech_exports: non-significant

• try_private, rd_exp on patent_res_per100people: significant, R-squared = 0.593
• try_private, rd_exp on patent_nres_per100people: significant, R-squared = 0.632

• pryntry, try_private, patent_nres_per100people on gdp_percapita_dividedby1000: significant, R-squared = 0.577
• pisa_read, pisa_math on gini_index: significant, R-squared = 0.725

Code: LinearRegressionNew.pdf

April 14, 2017 / Leave a comment

Cluster

I clustered the countries into 2 groups according to GDP per Capita and GINI index.

When I looked at the data, I figured out that the clusters are almost like the below:

Cluster 1: Developed Countries Cluster 2: Developing Countries

I plan to using other data points such as education, R&D or patent numbers to predict the cluster.

Code: Cluster.pdf

Data: MinimizedDataFullClustered.pdf

April 13, 2017 / Leave a comment

Regressions and Some Plots

(Note: I used sklearn for this regressions. Since there is no test statistics there, I used only R-squared to interpret. I will use statsmodels for more accurate interpretations.)

I performed some linear regressions on the data I have.

Below, you can find X, Y and R-squared for each regression:

1. Primary to Non-Tertiary Education Spendings on PISA Scores
   (R-squared = 0.029) (Plot Available)

   Comment: Spendings on primary to non-tertiary education explain a very little portion of PISA scores.

2. Education Spendings on R&D Expenditures
   (R-squared = 0.080) (Plot Available)

   Comment: Education spendins explain a little portion of R&D expenditures.

3. PISA Scores and Tertiary Education Spendings on R&D Expenditures
   (R-squared = 0.566)

   Comment: If we replace primary to non-tertiary education with PISA scores, R-squared increases significantly and now a large portion of R&D expenditures is explained.

4. PISA Scores and Education Spendings on R&D Expenditures
   (R-squared = 0.571)

   Comment: If we add spendings on primary to non-tertiary education to the previous regression R-squared increases a little bit.

5. Education Spendings on Researchers in R&D
   (R-squared = 0.025) (Plot Available)

   Comment: Education spendins explain a little portion of researchers in R&D.

6. PISA Scores and Tertiary Education Spendings on Researchers in R&D
   (R-squared = 0.138)

   Comment: If we replace primary to non-tertiary education with PISA scores, R-squared increases significantly and now some portion of R&D expenditures is explained but not much.

7. PISA Scores and Education Spendings on Researchers in R&D
   (R-squared = 0.204)

   Comment: If we add spendings on primary to non-tertiary education to the previous regression R-squared increases from 13.8% to 20.4%. It is significant.

8. PISA Scores, Education Spendings, R&D Expenditures and Researchers in R&D on High Technology Exports
   (R-squared = 0.356)

   Comment: Education data and R&D data explains a significant portion of high technology exports.

9. PISA Scores and Education Spendings on High Technology Exports
   (R-squared = 0.336)

   Comment: Only education data explains a little less, however it is still significant.

10. PISA Scores, Education Spendings, R&D Expenditures and Researchers in R&D on Patent Numbers
    (R-squared = 0.685)

    Comment: Education data and R&D data explains a large portion of patent numbers.

11. PISA Scores and Ecucation Spendings on Patent Numbers
    (R-squared = 0.460)

    Comment: Only education data explains a little less, however it is still significant.

12. PISA Scores, Education Spendings, R&D Expenditures and Researchers in R&D on GDP per Capita
    (R-squared = 0.515)

    Comment: Education data and R&D data explains a large portion of GDP per capita.

13. PISA Scores and Education Spendings on GDP per Capita
    (R-squared = 0.341)

Comment: Only education data explains a little less, however it is still significant.

14. PISA Scores and Education Spendings on GINI index
    (R-squared = 0.795)

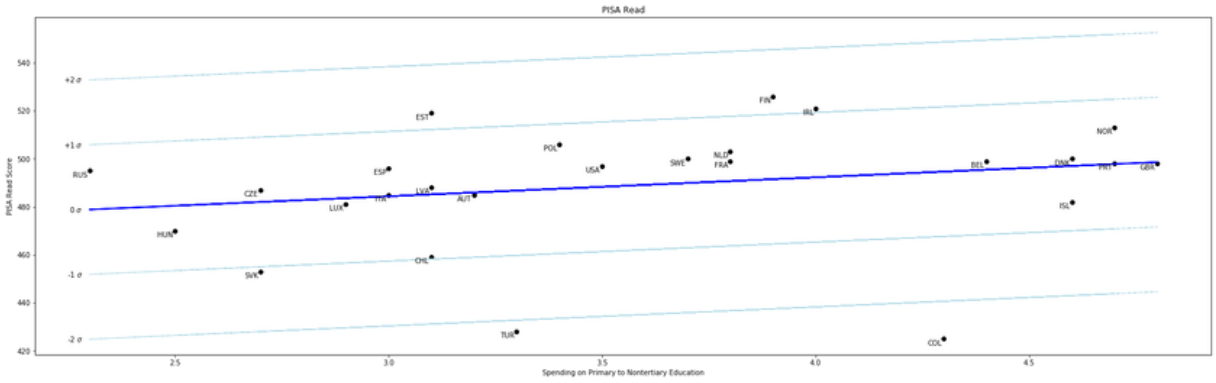
    Comment: Education data explains a very large portion of GINI index.

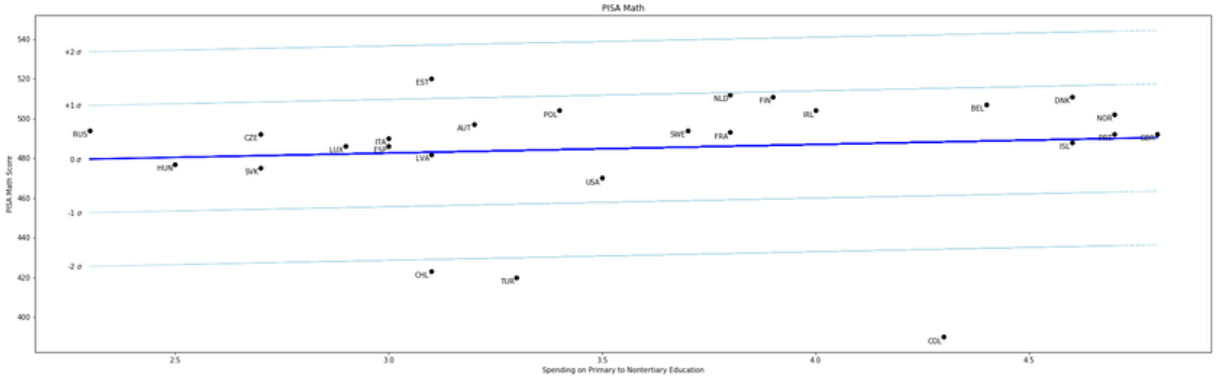
15. Education Spendings on GINI index
    (R-squared = 0.230) (Plot Available)

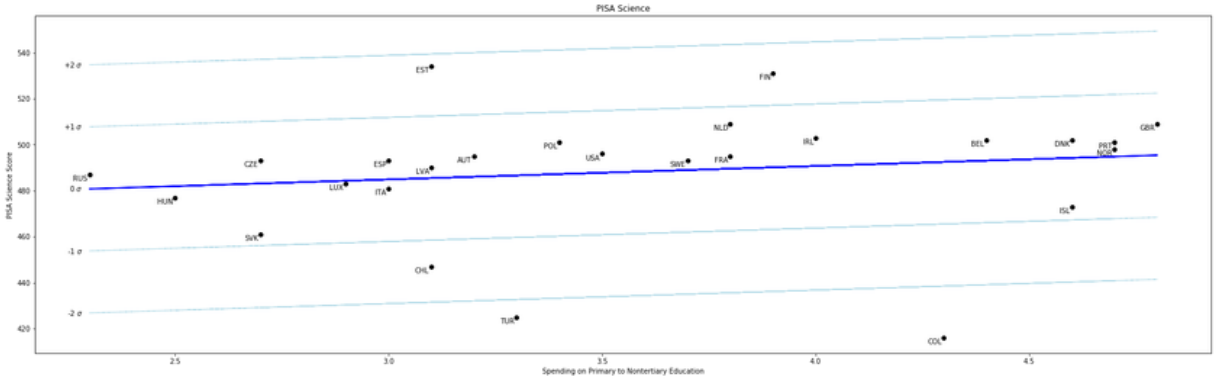
    Comment: Only education spendings explains less, however it is still significant.

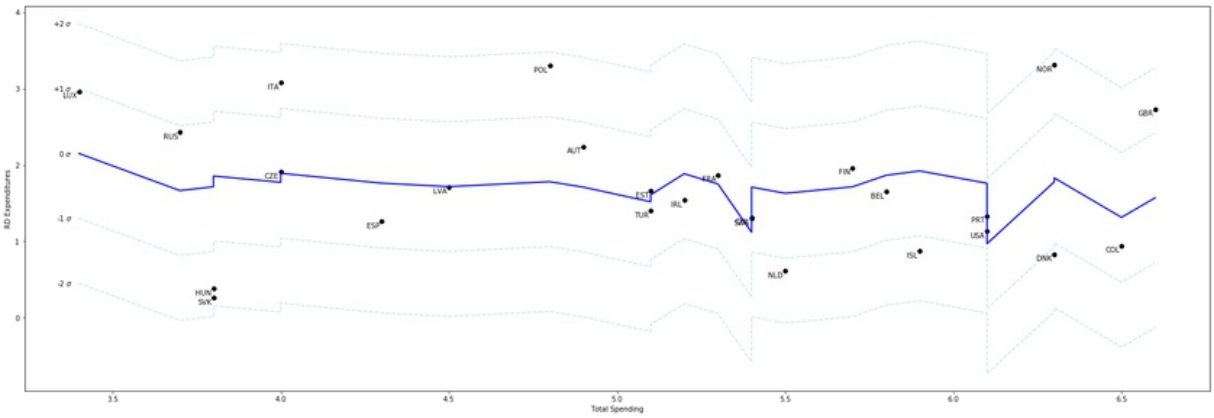
Code: LinearRegression.pdf

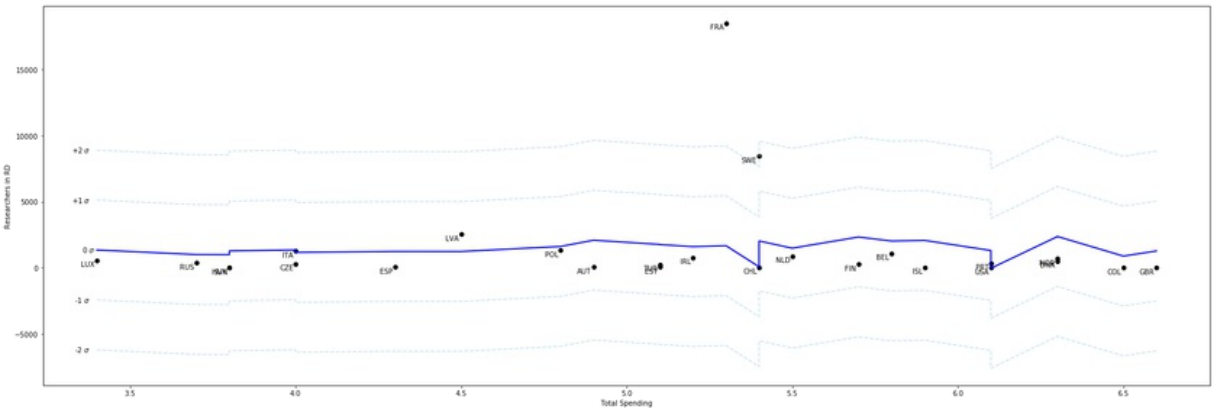
Plots:

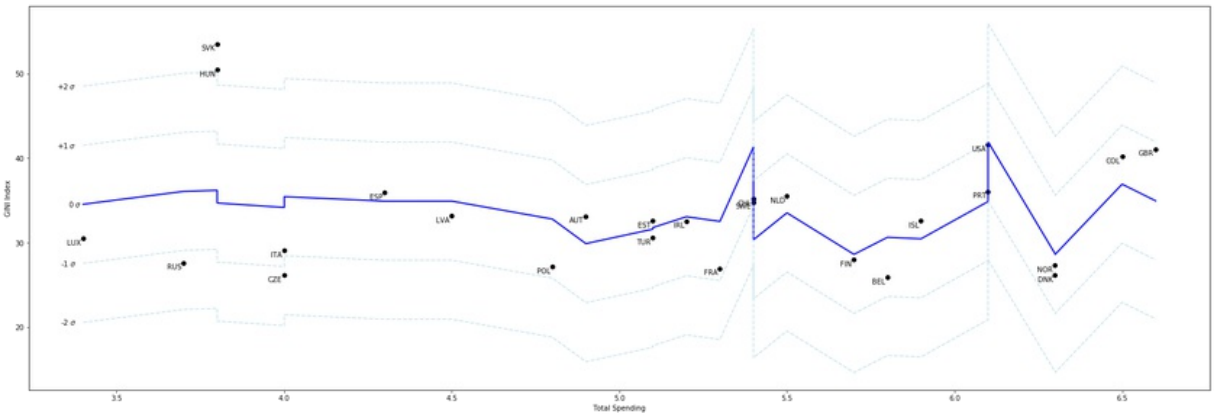












April 9, 2017 / Leave a comment

Minimized Data

I threw 2 data sets since they can be derived from others.

try : try_public + try_private
total: pryntry + try_public + try_private

Code: MinimizeData.pdf

Data: MinimizedDataFull.pdf

April 8, 2017 / Leave a comment

Search …

RECENT POSTS

- Reports
  May 19, 2017
- ZeroR Update
  May 19, 2017
- Equal Weight DTCs
  May 14, 2017
- Plan for last phase
  May 12, 2017
- Logistic Regressions for Percentages
  May 7, 2017
- Decision Trees for Percentages
  May 7, 2017
- Processing Data for Percentages
  May 7, 2017
- Confusion Matrices
  May 3, 2017
- Comparison of Different Models
  April 29, 2017
- Logistic Regressions
  April 29, 2017
- SVMs
  April 29, 2017
- Decision Trees
  April 29, 2017
- Linear Regressions
  April 29, 2017
- Getting and Processing Data
  April 29, 2017
- GitHub Integration
  April 26, 2017
- Part 1 Overview
  April 24, 2017
- Planning for the rest of the project
  April 24, 2017
- Regression (New)
  April 14, 2017
- Cluster
  April 13, 2017
- Regressions and Some Plots
  April 9, 2017
- Minimized Data

April 8, 2017

- Correlation Matrix & Scatter Plot Matrix
  March 24, 2017
- Filtered Missing Data
  March 24, 2017
- Merged Data
  March 23, 2017
- Code for Histograms
  March 18, 2017
- Histograms
  March 18, 2017
- Data Downloaded
  March 18, 2017
- Project Proposal (v3)
  March 18, 2017
- Project Proposal (v2)
  March 2, 2017
- Project Proposal
  March 1, 2017