STAT 420: Data Analysis Project

Apurva V. Hari, Alok K. Shukla

11/13/2016

Contents

Team

- Size : 2
- Details :

Name	NetID
Apurva V. Hari	vhari2
Alok K. Shukla	alokks2

Proposal

This Data Analysis Project is inspired by homework assignments from CMU class 36-401 (*Modern Regression*) and 36-402 (*Undergraduate Advanced Data Analysis*) .

Tentative Title

Who's Your Daddy? Is He Rich Like Me?
A survey of economic mobility across generations in contemporary USA.

Dataset

Background

The data come from a large study, based on tax records, which allowed researchers to link the income of adults to the income of their parents several decades previously. For privacy reasons, we don't have that individual-level data, but we do have aggregate statistics about economic mobility for several hundred communities, containing most of the American population, and covariate information about those communities.

Description

The data file `mobility.csv` has information on 741 communities. The variable we want to predict is economic mobility; the rest are predictor variables or covariates.

1. Mobility: The probability that a child born in 1980–1982 into the lowest quin- tile (20%) of household income will be in the top quintile at age 30. Individuals are assigned to the community they grew up in, not the one they were in as adults.
2. Population in 2000.

3. Is the community primarily urban or rural?
4. Black: percentage of individuals who marked black (and nothing else) on cen- sus forms.
5. Racial segregation: a measure of residential segregation by race.
6. Income segregation: Similarly but for income.
7. Segregation of poverty: Specifically a measure of residential segregation for those in the bottom quarter of the national income distribution.
8. Segregation of affluence: Residential segregation for those in the top qarter.
9. Commute: Fraction of workers with a commute of less than 15 minutes.
10. Mean income: Average income per capita in 2000.
11. Gini: A measure of income inequality, which would be 0 if all incomes were perfectly equal, and tends towards 100 as all the income is concentrated among the richest individuals (see Wikipedia, s.v. "Gini coefficient").
12. Share 1%: Share of the total income of a community going to its richest 1%.
13. Gini bottom 99%: Gini coefficient among the lower 99% of that community.
14. Fraction middle class: Fraction of parents whose income is between the na- tional 25th and 75th percentiles.
15. Local tax rate: Fraction of all income going to local taxes.
16. Local government spending: per capita.
17. Progressivity: Measure of how much state income tax rates increase with in- come.
18. EITC: Measure of how much the state contributed to the Earned Income Tax Credit (a sort of negative income tax for very low-paid wage earners).
19. School expenditures: Average spending per pupil in public schools.
20. Student/teacher ratio: Number of students in public schools divided by num- ber of teachers.
21. Test scores: Residuals from a linear regression of mean math and English test scores on household income per capita.
22. Highschooldropoutrate:Also,residualsfromalinearregressionofthedropout rate on per-capita income.
23. Colleges per capita
24. College tuition: in-state, for full-time students
25. College graduation rate: Again, residuals from a linear regression of the actual graduation rate on household income per capita.
26. Labor force participation: Fraction of adults in the workforce.
27. Manufacturing: Fraction of workers in manufacturing.
28. Chinese imports: Growth rate in imports from China per worker between 1990 and 2000.
29. Teenage labor: fraction of those age 14–16 who were in the labor force.
30. Migration in: Migration into the community from elsewhere, as a fraction of 2000 population.
31. Migration out: Ditto for migration into other communities.
32. Foreign: fraction of residents born outside the US.
33. Social capital: Index combining voter turnout, participation in the census, and participation in community organizations.
34. Religious: Share of the population claiming to belong to an organized religious body.
35. Violent crime: Arrests per person per year for violent crimes.
36. Singlemotherhood:Numberofsinglefemaleholdswithchildrendivided by the total number of households with children.
37. Divorced: Fraction of adults who are divorced.
38. Married: Ditto.
39. Longitude: Geographic coordinate for the center of the community
40. Latitude: Ditto
41. ID: A numerical code, identifying the community.
42. Name: the name of principal city or town.
43. State: the state of the principal city or town of the community.

Motivation

- Hands on experince with real life datasets.
- Practise with all techniques learnt in STAT420.
- Discover how applied statistics can help us answer socio-economic questions.

Data Snippet

Here is a snippet of data with only first 10 columns considered.

ID	Name	Mobility	State	Population	Urban	Black	Seg_racial	Seg_income	Seg_poverty
100	Johnson City	0.0622	TN	576081	1	0.021	0.090	0.035	0.030
200	Morristown	0.0537	TN	227816	1	0.020	0.093	0.026	0.028
301	Middlesborough	0.0726	TN	66708	0	0.015	0.064	0.024	0.015
302	Knoxville	0.0563	TN	727600	1	0.056	0.210	0.092	0.084
401	Winston-Salem	0.0448	NC	493180	1	0.174	0.262	0.072	0.061
402	Martinsville	0.0518	VA	92753	0	0.224	0.137	0.024	0.015

Credits

CMU, Statistics

https://www.stat.cmu.edu/~cshalizi/mreg/15/
http://www.stat.cmu.edu/~cshalizi/uADA/15/