Homework 1: Who's Your Daddy? Is He Rich Like Me?

36-402, Spring 2015

Due at 11:59 pm on Tuesday, 20 January 2015

GENERAL INSTRUCTIONS: You may submit either (1) a single PDF, containing all your written answers and figures, and a separate .R file with all the commands used to do your work, *or* (2) a single R Markdown file, integrating text, figures and code. Do not submit Word (.doc or .docx) files, since they will not be graded. (You can write in Word, just be sure to submit a PDF.)

When the assignment says "make a scatterplot of A against B", or "plot A against B", A goes on the vertical axis and B on the horizontal axis.

AGENDA: Getting back into practice with regression after the winter break; starting to unlearn some bad habits.

This assignment will look at economic mobility across generations in the contemporary USA. The data come from a large study[1], 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. We are interested in predicting economic mobility from the characteristics of communities.

The Data The data file `mobility.csv` has information on 741 communities[2]. 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 quintile (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.

[1] The solutions will say which. In the meanwhile, tracking it down would not actually help you very much.

[2] Technically, "commuting zones". These include cities and their suburbs and exurbs, but also many rural areas with integrated economies.

2. Population in 2000.

3. Is the community primarily urban or rural?

4. Black: percentage of individuals who marked black (and nothing else) on census 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 *national* 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 income.

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 number of teachers.

21. Test scores: *Residuals* from a linear regression of mean math and English test scores on household income per capita.

22. High school dropout rate: Also, *residuals* from a linear regression of the dropout 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. Single motherhood: Number of single female households with children divided 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.

Some of these variables are missing for some communities, and this may make a difference for some questions.

1. (5) Draw a map of mobility. That is, make a plot where the x and y coordinates are longitude and latitude, and mobility is indicated by color (possibly grey scale), by a third coordinate, or some other suitable device. Make sure your map is legible. Describe the geographic pattern in words.

2. (15) Make scatter plots of mobility against each of the following variables. Include on each plot a line for the simple or univariate regression, and give a table of the regression coefficients. Carefully explain the interpretation of each coefficient. (2 pts each) Do any of the results seem odd? (1 pt)

   (a) Population

   (b) Mean household income per capita

   (c) Racial segregation

   (d) Income share of the top 1%

   (e) Mean school expenditures per pupil

   (f) Violent crime rate

   (g) Fraction of workers with short commutes.

3. (15) Run a linear regression of mobility against all appropriate covariates.

   (a) (5) Report all regression coefficients and their standard errors to *reasonable* precision; you may use either a table or a figure as you prefer. Do not just paste in R's output.

   (b) (1) Explain why the ID variable must be excluded.

   (c) (4) Explain which other variables, if any, you excluded from the regression, and why. (If you think they can all be used, explain why.)

   (d) (5) Compare the coefficients you found in problem 2 to the coefficients for the same variables in this regression. Are they significantly different? Have any changed sign?

4. (10) *The wrong side of the tracks starts at Giant Eagle* Find Pittsburgh in the data set.

   (a) (1) What its actual mobility? What is its predicted mobility, according to the model?

   (b) (3) Holding all else fixed, what is the predicted mobility if the violent crime rate is doubled? If it is halved?

   (c) (3) Holding all else fixed, at what level of income segregation does the model predict that mobility will exceed 1.0?

   (d) (3) Holding all else fixed, what would the income share of the top 1% have to be for the model to predict that mobility will fall to 0.0?

   (We will see later in the course how to avoid the embarrassment of models that predict probabilities greater than 1 or less than 0.)

5. (10) *Free as in beer*

   (a) (1) The national mobility level is the average mobility across all communities, weighted by population. What is it?

   (b) (3) Suppose college were made free for everyone. Calculate the *change* in the predicted mobility for each community. Report the minimum, median, mean and maximum changes.

   (c) (1) Find the change to the predicted (not actual) national mobility level from making college free for everyone. *Hint:* consider a weighted average, or weighted sum, of your vector of answers from Problem 5b.

   (d) (3) Give a (rough) 95% confidence interval for the change in the predicted national mobility level.

   (e) (2) Explain at least one way in which this calculation is unrealistic.

6. (10) *Distinctions vs. differences*

   (a) (2) Make a table ranking the variables by the magnitude of the t statistic in the regression results (i.e., rank by $|t|$, not t).

   (b) (6) For each variable in the model, find the expected change in mobility from a one standard deviation change in that variable (assuming all else is fixed). Provide a table ranking variables by the magnitude of their impact.

   (c) (2) How similar is the ranking by impact to the ranking by t statistics?

7. (5) Make a map of the model's predicted mobility. How does it compare, qualitatively, to the map of actual mobility?

8. (5) *After making proper allowances*

   (a) (1) Make a map of the model's residuals.

   (b) (2) What are the five communities with the largest positive residuals? The five with the most negative residuals? (Can you mark these on the map?)

   (c) (2) One interpretation of these residuals is that they show communities where some factor not included in the model leads to higher (or lower) mobility than in otherwise-similar communities. Suggest at least one other interpretation. Could you test these ideas with this data set?

9. (5) *Expectations and reality*

   (a) (3) Make a scatterplot of actual mobility against predicted mobility. Is the relationship linear? Should it be, if the model is right? Is the relationship flat? Should it be, if the model is right?

   (b) (2) Make a scatterplot of the model's residuals against predicted mobility. Is the relationship linear? Should it be, if the model is right? Is the relationship flat? Should it be, if the model is right?

10. (20) *Model checking will continue until morale improves*

(a) (5) For *each* variable in the model, make a scatterplot of the model's residuals against the predictor variable. (You will have a lot of plots.)

(b) (5) Explain why, if the linear model is right, all the relationships you just plotted should be flat.

(c) (5) Explain why, if the usual assumptions for t tests and their p-values are right, each plot should have a roughly constant vertical spread of points as one moves from left to right.

(d) (5) Which residual plots look like they're flat with constant width? For the ones which don't look like this, describe how they differ.

Extra credit, 5 points: Add kernel smoothing lines to each of the residual plots. Comment.