Homework 7: Your Daddy's Rich (and Your Mama's Good-Looking)

36-313, Spring 2021

Due at 6 pm on Thursday, 21 October 2021

> **Agenda**: Examining levels, and spatial variation, in economic mobility across generations; examining the factors which predict higher or lower levels of mobility; examining predictions of what would happen under certain policy changes.

This assignment will look at economic mobility across from one generation to the next in the contemporary USA. 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 commuting zones[1], containing most of the American population, and covariate information about those commuting zones. We are interested in predicting economic mobility from the characteristics of commuting zones.

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

- *Mobility*: The fraction of child born in 1980–1982 into the lowest quintile (1/5) of household income who were in the top quintile at age 30. Individuals are assigned to the commuting zone they grew up in, not where they lived as adults.
- *Population* in 2000.
- Is the commuting zone primarily urban or rural?
- *Black*: percentage of individuals who marked black (and nothing else) on census forms.
- *Racial segregation*: a measure of residential segregation by race (higher=more segregated).
- *Income segregation*: Similarly but for income.
- *Segregation of poverty*: Specifically a measure of residential segregation for those in the bottom quarter of the national income distribution.
- *Segregation of affluence*: Residential segregation for those in the top quarter.
- *Commute*: Fraction of workers with a commute of less than 15 minutes.
- *Mean income*: Average income per person in 2000.
- *Gini*: The Gini index for the commuting zone, on a 0–100 scale.
- *Share 1%*: Share of the total income of a commuting zone going to its richest 1%.
- *Gini bottom 99%*: Gini index among the lower 99% of that commuting zone.
- *Fraction middle class*: Fraction of parents whose income is between the (national) 25th and 75th percentiles.
- *Local tax rate*: Fraction of all income going to local taxes.
- *Local government spending*: per capita.
- *Progressivity*: Measure of how much state income tax rates increase with income. _ *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).
- *School expenditures*: Average spending per pupil in public schools.
- *Student/teacher ratio*: Number of students in public schools divided by number of teachers.

[1] These are similar to the metropolitan statistical areas we saw in the Current Population Survey data in earlier homeworks, but the precise definitions and boundaries are different.

- *Test scores*: Residuals from a linear regression of mean math and English test scores on household income per capita.
- *High school dropout rate*: Also, residuals from a linear regression of the dropout rate on per-capita income.
- *Colleges* per capita
- *College tuition*: at the state public university, in-state, for full-time students
- *College graduation rate*: Again, *residuals* from a linear regression of the actual graduation rate on household income per capita.
- *Labor force participation*: Fraction of adults in the workforce.
- *Manufacturing*: Fraction of workers in manufacturing.
- *Chinese imports*: Growth rate in imports from China per worker between 1990 and 2000.
- *Teenage labor*: fraction of those age 14–16 who were in the labor force.
- *Migration in*: Migration into the commuting zone from elsewhere, as a fraction of 2000 population.
- *Migration out*: Ditto for migration into other commuting zones.
- *Foreign*: fraction of residents born outside the US.
- *Social capital*: Index combining voter turnout, participation in the census, and participation in commuting zone organizations.
- *Religious*: Share of the population claiming to belong to an organized religious body.
- *Violent crime*: Arrests per person per year for violent crimes.
- *Single motherhood*: Number of single female households with children divided by the total number of households with children.
- *Divorced*: Fraction of adults who are divorced.
- *Married*: Ditto.
- *Longitude*: Geographic coordinate for the center of the commuting zone
- *Latitude*: Ditto
- *ID*: A numerical code, identifying the commuting zone.
- *Name* of the commuting zone's main city or town.
- *State* the main city or town is located in.

Some of these variables are missing for some commuting zones.

General note: When you're asked to matter a scatterplot of A against B, or to plot A against B, A goes on the vertical axis and B on the horizontal axis.

1. *Summary statistics*

    a. (3) Draw a histogram of the mobility rates for the commuting zones. Describe the shape, in words.

    b. (2) What are the minimum, maximum, median and mean values of mobility? What's the standard deviation across commuting zones?

    c. (2) What is the mean value of mobility if we weight commuting zones by their population? What's the weighted standard deviation?

    d. (1) The weighted mean you found in Q1c should be different from the unweighted mean in Q1b. Does the sign of the difference indicate that larger commuting zones have higher or lower mobility rates than smaller ones?

2. *Mapping, and extremes*

    a. (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 something similar. Describe the geographic pattern of mobility in words.

    b. (5) Find the six commuting zones with the highest mobility rates, and the six commuting zones with the lowest rates. Report this in the form of a table (or two tables), showing the commuting zones' names, states, geographic coordinates, population, and mobility rates.

    c. (5) Re-draw the map to highlight (using color or shape or something similar) the twelve commuting

zones with extremely high or low mobility rates you found in Q2b. Distinguish the high and low mobility commuting zones.

d. (5) Do you notice anything the highest-mobility commuting zones have in common? That the lowest-mobility commuting zones have in common? That all these extreme commuting zones have in common? *Hint*: $\sqrt{\frac{p(1-p)}{n}}$.

3. *One big regression* In this question, you'll linearly regress mobility against all available and sensible covariates.

   a. (1) Why should the `ID` variable be excluded from the regression?

   b. (4) Why should `Name`, `State`, `Latitude` and `Longitude` all be excluded from the regression?

   c. (3) Which other variables do you need to leave out of the regression, and why? (If you think all the others can be used, explain.)

   d. (5) Report all regression coefficients and their standard errors to *reasonable* precision; use a table or a figure as you prefer.

4. *Oomph* One common way to gauge the importance of a predictor variable in a regression is to say how much we expect the outcome to change for a one-standard-deviation increase in that variable. In symbols, we want $\mathbb{E}\left[Y|X = x^{(+j)}\right] - \mathbb{E}\left[Y|X = x\right]$, where $x^{(+j)}$ is just like x, except that variable j has been increased by one standard deviation[2].

   a. (5) Explain why, for a linear model, $\mathbb{E}\left[Y|X = x^{(+j)}\right] - \mathbb{E}\left[Y|X = x\right] = \sigma_j \beta_j$, where σ_j is the standard deviation of variable j, and β_j is the coefficient on that variable.

   b. (3) Calculate the standard deviations across commuting zones for all the input variables your regression model in Q3. Present the results as a table. *Note*: Remember that the standard deviation for each variable isn't the same as the standard error in its coefficient.

   c. (5) Calculate the expected change in mobility rate from a one standard deviation increase in each of your predictor variables. Include a standard error for each calculation. Present the results as a table or graph.

5. *Interpreting the regression*

   a. (5) If local leaders would like to increase mobility, what variables should they try to increase? Which ones seem most important, based on what you found in Q3 and Q4?

   b. (5) If local leaders would like to increase mobility, what variables should they try to decrease? Which ones seem most important?

   c. (4) Are there variables that matter but which leaders cannot alter?

   d. (3) Are there variables that matter but which leaders *should not* try to alter?

6. *More mapping*

   a. (5) Make a map of *predicted* mobility rates, according to the linear regression model from Q3. Describe the geographic pattern in words.

   b. (4) Make a map of the *residuals* from the linear regression model in Q3. Describe the geographic pattern.

   c. (3) What are the six commuting zones with the largest positive residuals? The six with the most negative residuals? Report these in a table like that in Q2b, only with an extra column for the residuals.

[2] Using a one standard deviation change is common, but you also see people asking about what happens if the variable is moved from its 25th percentile to its 75th percentile, or from its 10th to its 90th. The common idea is to have the size of the change be somehow typical of the differences we see in the data.

d. (1) One interpretation of these residuals is that they show commuting zones where some factor not included in the model leads to higher (or lower) mobility than in otherwise-similar commuting zones. Your findings in part (c) should suggest another interpretation — what? *Hint*: $\sqrt{\frac{p(1-p)}{n}}$.

7. *Expectation vs. 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?

8. (1) *Timing* How long, roughly, did you spend on this assignment?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra credit: *Maybe not quite so big a regression*

a. (3) You should find something odd about the group of coefficients for the Gini index, the top 1% share, and the Gini index for the bottom 99%. What's going on here?

b. (3) You should also find something odd about the group of cefficients for income segregation, the segregation of poverty, and the segregation of affluence. What's going on here?

c. (3) How do your answers to Q4a and Q4b suggest changing the regression model (if at all)?