Homework 4: Adjust All the Things

36-313, Fall 2022

Due at 6 pm on Thursday, 29 September 2022

> **Agenda**: Measuring between-group inequality for various kinds of group; practice adjusting for other variables; practice arguing about what to adjust for.

We are continuing to work with the ASEC 2020 data from last time, but this time we'll be exploring issues about explaining, adjusting, or accounting for inequalities between groups.

1. *Racial categories and income inequality* For purposes of this problem set, we'll continue to divide the population into black, white, Asian, and others[1], as in homework 3 or lecture 7.
    a. (3) What fraction of those surveyed belong to each of these four categories?
    b. (4) What is the mean household income for each race?
    c. (3) What fraction of the total variance in income is between racial groups?
2. *Food stamps and income inequality* The variable `FOODSTMP` records whether or not the person surveyed received food stamps[2].
    a. (3) What fraction of those surveyed received food stamps?
    b. (4) Find the mean household income for those who did and did not receive foodstamps.
    c. (3) What fraction of the total variance in income is between these two groups?
3. *Education and income inequality* Previously, we've simply divided survey recipients into those with and with a bachelor's degree, but the `EDUC` variable records more detailed educational levels. (Report results by names, not numeric codes.)
    a. (3) What fraction of those surveyed are at each of the different educational levels?
    b. (4) What is the mean income for each educational level?
    c. (3) What fraction of total variance is between educational levels?
4. *Geography and income inequality* The Census Bureau divides the country up into a large number of "metropolitan statistical areas", based on commuting and business patterns. The `METAREA` variable records which metropolitan area for each person in the survey, using a large set of numerical codes. Find the mean income for each metropolitan area, but do not print it out (yet). If you do this by running a regression, be sure that the computer treats `METAREA` as categorical, not numeric, variable.
    a. (4) What are the six metropolitan areas with the highest mean incomes? (Include the incomes as well, in this and all similar questions.) Give the names, not the code numbers.
    b. (3) What are the six metropolitan areas with the lowest mean incomes?
    c. (5) Using a histogram (or something similar), plot the *distribution* of mean incomes across metropolitan areas. Comment on the shape of the distribution.
    d. (3) What fraction of the total variance in income is between metropolitan areas?
5. *Occupation and income inequality* The variable `OCC` records what occupation each survey respondent worked in at the time of the survey (or their main occupation, if they had more than one). The codes are available at [https://cps.ipums.org/cps/codes/occ_2020_codes.shtml]. In particular, note that the code `0000` indicates "not in the workforce". Find the mean income level for each occupation, but don't

[1] I am not asking you to consider Hispanic status, not because it's not important but in order to keep the data-wrangling to a minimum.

[2] Formally, the Supplemental Nutrition Assistance Program (SNAP). This is a federal benefit available to families with incomes below certain thresholds, which essentially gives them money every month to buy food, and only food. (The name comes from the first days of the program, when people would get actual paper tokens, the "stamps", which they'd hand to grocers, etc., instead of cash, and the grocers would redeem the stamps with the government for actual money; these days SNAP issues a payment card that gets processed like any other debit card.)

display it, yet. If you use a regression, make sure `OCC` is treated as a categorial and not a numerical variable.

   a. (4) What are the six occupations with the highest mean incomes?
   b. (3) What are the six occupations with the lowest mean incomes?
   c. (5) Using a histogram (or something similar), plot the distribution of mean incomes across occupations. Comment on the shape of the distribution.
   d. (3) What fraction of total variance in income is between occupations?

6. *One big regression* Run a linear regression of household income on race, education, metropolitan area and occupation (but not foodstamps). This should have 798 coefficients, so please don't print them out.
   a. (5) Are the racial contrasts bigger in magnitude than you found in Q1, smaller, or mixed? Have any of them changed sign?
   b. (5) Are the educational contrasts bigger in magnitude than you found in Q3, smaller, or mixed? Have any of them changed sign?
   c. (4) What fraction of the total variance in income is between the categories in this model?
7. *Explanations and adjustments*
   a. (5) Suppose we're primarily interested in racial inequality in income. Does it make sense to adjust for education, geography and occupation in the way the model in Q6 does? Explain.
   b. (5) Continuing from (c), would it also make sense to adjust for receiving food stamps?
   c. (5) You've calculated the fraction of variance that's between group variance for racial categories, foodstamps, educational levels, metropolitan areas, occupations, and the composition or intersection of race, education, area and occupation. The fraction of variance that's between groups is often called the "proportion of variance explained". In which of these cases (if any) does it make sense to say that the models *explain* income inequality? Why?
8. *Timing* (1) 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.