Welcome to our research team! You've come at a busy time; we are preparing a paper on economic mobility and the intergenerational elasticity of income. Here are a few tasks we'd like you to complete, to the best of your ability. While we prefer that you use Stata, you are free to use whichever statistical software you are most comfortable with. The presentation date is fast approaching, so please complete the tasks within the allotted time as indicated in the email. Please send us your results in a pdf, along with the code file which you ran to get your results.

Questions 2—4 can each be answered in a single paragraph (question 1 doesn't require text). Use population weighting where appropriate. Make graphs look sufficiently professional to include in an academic presentation. Your code should be carefully commented so that we can read it easily.

Question 1

In our lab we often represent income with a person's percentile rank in the national income distribution. The CSV file *national_collapse.csv* lists the average income percentile of children, given the income percentile of their parents, their gender and their race.

Aggregate the data across racial groups. Present the resultant data in a scatter plot, with parent income on the x-axis and a separate series for each gender. Add fitted lines for each gender to the graph and report the corresponding slope coefficients.

Question 2

Economic mobility varies across space. The CSV file *tract_collapse.csv* lists a neighborhood-specific measure of economic mobility: the average income percentile of children whose parents were in the 25th percentile, among children who grew up in a particular Census tract. Separate statistics are included for Black children and White children, along with an estimate that includes children of all racial groups. For each tract separately, we estimate the relationship between parent income and the adult income of children who spent at least one year of childhood in that tract and use that relationship to predict income for children with parents at the 25th percentile. We do this for pooled race and gender as well as separately by race and release statistics if there are at least 20 children in the sub category.

For now, use only the estimates which combine all racial groups. Aggregate the data to the commuting zone level using the county-to-commuting zone crosswalk *cz_county_crosswalk.csv*. For each commuting zone, compute how economic mobility differs from the national mean. Display the five best and the five worst commuting zones on a horizontal bar graph with the difference from the mean on the x-axis and the commuting zone on the y-axis. Plot the five best commuting zones in blue and the five worst commuting zones in red.

What do you notice about the resulting groups of best and worst commuting zones?

Question 3

We are interested in how economic mobility relates to average incomes. The CSV file *tract_covariates.csv* contains the average income among all households in each tract in the year 2000. Merge this data to the tract-level dataset used in the previous question, and correlate household income with our measure of economic opportunity for Black children, White children and all children. Briefly justify your choice of weights. Contrast the three correlations and suggest two explanations as to why they differ. Discuss how your explanations could be tested (perhaps using other data or in other contexts).

Question 4

Consider a data generating process $y_i = \mu_{t(i)} + e_i$, where y_i is the adult income of child i who grew up in tract $t(i)$, and the tract-level latent mean $\mu_{t(i)}$ is defined to be uncorrelated with the individual-level error term e_i.

a) Let $\overline{y_t}$ be the sample mean of incomes of children who grew up in tract t. Let x_t be the level of education spending in the tract. Will the correlation between $\overline{y_t}$ and x_t generally be larger or smaller than the correlation between μ_t and x_t? Prove your answer using covariance algebra.

b) Say you are provided with s_t, the standard error of $\bar{y}_t$. Assume s_t doesn't itself have any error. Given your answer in part a) propose a correction that could be used to find the correlation between μ_t and x_t and show how you can use the standard error to implement that correction.

c) Now say you observe $\bar{z}_t$, the average college attendance of the same set of children who grew up in tract *t.* Explain why it would be inappropriate to use the correction you proposed in part b) to find the correlation between μ_t and $\bar{z}_t$.

d) Now assume you are not provided with s_t. Instead, the children who grew up in each tract are randomly divided into two even-sized subsamples a and b, and you are provided the average income of children in these subsamples. Explain how you could use these to find the correlation between μ_t and x_t.