# First Day on the Job

Files included (in .csv format):

- `cz_level_data` - Estimates of intergenerational mobility for children at the 25th percentile and a set of covariates by commuting zone (this is a specific definition for agglomerations in the United States)
- `cz_names` - Names for all commuting zones in the United States

Welcome to our research team! You've come at a busy time; we are getting ready to present a paper on the intergenerational elasticity of income and how this varies by place. 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 whatever 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 answers and results in a pdf, and also send us your code file (the file that you ran to get your results). Your code should be carefully commented so that we can read it easily, and your explanations to any questions we ask should be concise. Use population weighting where appropriate.

## 1 The Intergenerational Elasticity of Income

We are interested in the intergenerational elasticity of income, and we know that for children who never move during childhood ("permanent residents"), their earnings rank in a given commuting zone (this is a specific definition for agglomerations in the United States) is given by

$$y_i = \alpha_{cs} + \psi_{cs} p_i + \epsilon_i,$$

where $y_i$ denotes child i's rank in the income distribution relative to all other children within their birth cohort, $\alpha_{cs}$ denotes the constant for commuting zone $c$ and birth cohort $s$, $\psi_{cs}$ is the parameter of interest, $p_i$ is the rank of the child's parents and $\epsilon_i$ is the error term. From this regression, we can produce estimates of the expected rank of a child in a given commuting zone $c$ and birth cohort $s$ with parent income rank $p$. This will be given by

$$\bar{y}_{pcs} = \hat{\alpha}_{cs} + \hat{\psi}_{cs} p$$

We are interested in how this metric varies across commuting zones in the United States for children with parents at the 25th percentile of the income distribution (variable name "perm_res_rank_p25"). Use the provided data to construct a new variable: For each CZ, compute the difference of perm_res_rank_p25 from the national mean. Then display the five best and the five worst CZs on a horizontal bar graph (in ascending or descending order) with the difference from the mean on the x-axis and the CZ on the y-axis. Plot the five best CZs in blue and the five worst CZs in red, and include the name of the commuting zone and the respective state in the y-axis labels (note that you will have to merge on the names of the commuting zones as these are stored in a separate data set). Format the figure as professionally as you can.

## 2 The Intergenerational Elasticity of Income II

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

## 3 State-Level Data

Apart from the CZ-level data, we are also interested in data at the state level. Use the appropriate commands to aggregate the data to the state level. Then use the resulting dataset to create a map of the United States, coloring states with lower expected ranks of children with parents at the 25th percentile in a darker shade. [Hint: In Stata, this can be done using the package *maptile*. In R, you can use the packages *maps* and *ggplot2*.]

## 4 Estimates Based on Movers

So far, we have focused on the expected ranks of children who have grown up in just one commuting zone ("permanent residents"). Now we will also look at children who have moved during their childhood ("movers") and the estimated effect on earnings for these children of spending an additional year in the commuting zone they have moved to. Using permanent residents' outcomes, $\bar{y}_{pds}$, to predict moving children's outcomes, $y_i$, we can write these as

$$y_i = \alpha + \beta_m \bar{y}_{pds} + \theta_i,$$

where $\beta_m$ captures the mean impact of spending year $m$ of one's childhood onwards in an area where permanent residents have 1 percentile better mean outcomes.

Note that we may not be estimating $\beta_m$ consistently in the equation above. Show that, assuming that the timing of moving to commuting zone $d$ is orthogonal to $\theta_i$, we can consistently estimate the one year exposure effect $\gamma_m = \beta_m - \beta_{m-1}$. [Prove this using conditional expectation functions, variances, and covariances if you can.]

## 5 Correlations

Going back to the dataset used in question 1, the variable "total_exposure_effect_p25" captures the causal effects of spending 20 years from birth in a given commuting zone for children at the 25th percentile. Standardize the following covariates: fraction of African American residents, poverty share, racial segregation, fraction of single-parent households, dropout rate, school expenditures per student, the gini coefficient for the bottom 99 percent of the income distribution, and the median house value. Then run individual regressions of the exposure variable on each of the standardized covariates, restricting the sample to observations where neither the exposure variable nor permanent residents' outcomes are missing. Allow for autocorrelation in the error terms at the state level. Repeat this procedure for the outcomes of permanent residents at the 25th percentile, and put the resulting coefficients and standard errors in a table with two columns (one for each dependent variable). Format this table as professionally as you can. How do the regression coefficients compare? What does this suggest about the outcomes of permanent residents?