High-Dimensional Confounding

The Job Corps is the largest U.S. labor market program targeting disadvantaged youths. It provides academic, vocational, and social training, as well as health care counseling and job search assistance, for an average duration of eight to nine months. Mathematica Policy Research carried out a randomized experiment with the Job Corps. About 60% of the experimental participants were randomly selected to receive an offer to participate in the Job Corps. Out of these, 73% joined the Job Corps program and actually started to participate after an average duration of 1.4 months. The remaining experimental participants were assigned to the control group. Control group members were not eligible for the Job Corps programs for three years following randomization (approximately 1% participated anyway).

We estimate the effects of actually participating in the Job Corps on earnings fours years after the randomisation. The experimental Job Corps data `job_corps.csv` contains 10,516 observations.

Variable list:

- `EARNY`: Earnings per week in Year 4
- `assignment`: Offer to join Job Corps
- `participation`: Actual participation in Job Corps
- `female`: Dummy for females
- `age_1`: Aged between 16-17 years
- `age_2`: Aged between 18-19 years
- `age_3`: Aged between 20-24 ye
- `ed0_6`: 0-6 months education program in last year
- `ed6_12`: 6-12 months education program in last yea
- `hs_ged`: High school or GED credential
- `white`: Dummy for white
- `black`: Dummy for African-American
- `hisp`: Dummy for Hispanics
- `oth_eth`: Dummy for other ethnicity
- `haschld`: Dummy for parents
- `livespou`: Lives with spouse or partner

- `everwork`: Ever had job for two weeks or more
- `yr_work`: Worked in year prior to random assignment
- `currjob`: Has job at random assignment
- `job0_3`: Below 3 months employed in last year
- `job3_9`: 3-9 months employed in last year
- `job9_12`: 9-12 months employed in last year
- `earn1`: Yearly earnings less than $1,000
- `earn2`: Yearly earnings $1,000-5,000
- `earn3`: Yearly earnings $5,000-10,000
- `earn4`: Yearly earnings above $10,000
- `badhlth`: Dummy for bad healt
- `welf_kid`: Family on welfare when growing up
- `got_fs`: Received food stamps in last year
- `publich`: Public or rent-subsidized housing
- `got_afdc`: Received AFDC in last year
- `harduse`: Used hard drugs in last year
- `potuse`: Smoked marijuana in last year
- `evarrst`: Ever arrested dummy
- `pmsa`: Lives in PMSA
- `msa`: Lives in MSA

Load the data `job_corps.csv`. Install the packages `hdm`, `glmnet`, `tidyverse`, `AER`, `lmtest`, `sandwich`, and `fBasics`.

Exercise 1: Conventional Methods

1. Report the descriptive statistics of all variables.
2. Estimate the unconditional effect of Job Corps participation on earnings (using OLS). Can we interpret this effect causally? What are potential sources of bias?
3. Use an instrumental variable approach to estimate the effect of Job Corps participation on earnings. Use the randomised offer to participate in the Job Corps as instrument for participation. How do the effects change compared to the OLS results? For which population are the effects identified?

Exercise 2: Double Selection Procedure

1. Use a Lasso to predict earnings. Store the selected variables.
2. Use a Lasso to predict participation in Job Corps. Store the selected variables.
3. Take the union of all variables selected as control variables in an OLS regression of earnings on Job Corps participation.
4. Use the `rlassoEffect` command to replicate the result.

Exercise 3: Double Machine Learning

1. Partition the sample in two equally sized parts.
2. Predict the earnings under non-participation in Job-Corps in both partitions.
3. Predict Job Corps participation in both partitions.
4. Calculate the efficient score using a cross-fitting procedure.
5. Estimate the average treatment effects on the treated. Are the effects significant?