ECON 121, Applied Econometrics and Data Analysis   **Prof. Tom Vogl**

Fall 2025   **University of California San Diego**

PROBLEM SET 4: EFFECTS OF HEAD START

DUE BY 11:59 PM PST ON THURSDAY 11/20

Head Start is an early-childhood development program run by the U.S. federal government. It provides health, nutrition, and education services to children from disadvantaged backgrounds. The Markdown template will load a sample of of children of NLSY '79 participants, some of whom participated in Head Start. All children in the sample have at least one sibling also in the sample. The variables are ordered as follows:

- *head_start - sibdiff* relate to head start participation.

- *mom_id - lnbw* were determined prior to Head Start participation. (Note: the PPVT is an early-childhood cognitive test. The other variable names are self-explanatory.)

- *comp_score_5to6 - comp_score_11to14* are test scores in childhood.

- *repeat - fphealth* are outcomes in the teenage years and young adulthood.

For various reasons, some variables have missing data. Decide how to deal with missing values on your own.

You are encouraged to work in a group of up to 4 members. You may write code together, but you must write verbal answers yourself. Please use the Markdown template for your code. Then produce a single PDF with code, results, and writing, and upload to Gradescope. You may not load any additional R packages beyond those already written into the template. If you use any functions or syntax that we have not used in class, explain what they do and why you are using them.

1. List your group members.

2. Summarize the data. What can you say about the backgrounds of children who participated in Head Start relative to those who did not?

3. Using OLS, estimate the association between Head Start participation and age 5-6 test scores. Make sure you compute standard errors correctly. If we assume Head Start participation is exogenous, what can we conclude about the effects of Head Start on test scores? Interpret the magnitude of the estimated coefficient in terms of the standard deviation of test scores. Is it reasonable to assume that Head Start participation is exogenous? Do you think there are likely to be omitted variables at the family level? If so, how do you think they might bias the estimated coefficient?

4. Now focus on "between mother" variation in Head Start participation and age 5-6 test scores. Create a new data frame in which each observation is a mother, and the variables are the family mean of Head

Start participation and the family mean of the age 5-6 test score. Using OLS, estimate the association between mean Head Start participation and mean test scores across mothers. How does the estimated coefficient compare with the one from question (3)?

5. Returning to the original data frame, estimate the association between Head Start participation and age 5-6 test scores in a model with mother fixed effects. What do the results imply about the effects of Head Start on test scores? Interpret the magnitude of the estimated effect in terms of the standard deviation of test scores. If the fixed effect results are different from questions (3) and (4), explain why. Which estimate most likely reflects the effect of Head Start participation on test scores, and why?

6. Include pre-Head Start variables as covariates in the regression with mother fixed effects. Explain your choice of covariates. Does the inclusion of covariates change the estimated coefficient on Head Start? What do you conclude about the robustness of the fixed effect estimate of the effect of Head Start?

7. Some advocates for early-childhood education suggest that the effects of programs like Head Start are long-lasting. Carry out fixed effects analyses of test scores at later ages. Does Head Start participation have similar effects on test scores in later childhood, or do the effects fade out with age? To make the test scores comparable across ages, you can standardize them by subracting the mean and dividing them by the standard deviation.

8. Estimate fixed effects models of the effect of Head Start on longer-term outcomes besides test scores. Many of these outcomes are binary, but you may use linear models. Interpret your results.

9. Focusing on one longer-term outcome (your choice), test whether the effect of Head Start participation on the outcome varies by race/ethnicity. Interpret your results.

10. Democrats advocate for expanding federal funding for early-childhood education programs, while Republicans argue for cuts. Based on your results, which position seems better supported by evidence? Would you feel comfortable using your results to predict the effects of such an expansion? Why or why not?