PROBLEM SET 2

**Due on Monday, February 26, 2024**

I - INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. Download this Word document file into your computer
2. Insert all your answers into this Word document. Guidance [here](https://www.dropbox.com/s/ox9fhmbpvy2viw5/How%20to%20incorporate%20handwritten%20work%2C%20Stata%20output%2C%20and%20screenshot%20images.pdf?dl=0) on how to insert non-Word objects such as handwritten work or screenshot images in your answers.
3. **Once your document is complete, please save it as a PDF**. This is important to make sure all your work is preserved in the process of submission to Canvas.
4. Please submit an electronic copy of the PDF and your **replicable Stata or R script** to the Canvas assignment page.

II - IDENTIFICATION

1. Your information

|  |  |
| --- | --- |
| Your Last Name: |  |
| Your First Name: |  |

(2) Group Members (please list the classmates you worked with on this problem set):

|  |
| --- |
|  |

1. Compliance with Harvard Kennedy School Academic Code[[1]](#footnote-1) (mark with an X below)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Yes** | **No** | | I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code |  |  | |

For this problem set, we will be examining the methods used in the following paper:

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, *106*(4), 979-1014.

# Conceptual Questions (40 points + 1 extra point)

1. Clearly state the primary research question that the author is trying to answer. Why should policymakers care about this question? (2 points)

|  |
| --- |
|  |

1. The authors used an instrumental variable approach because they believed a naïve regression specification (regressing earnings on education) would be insufficient. What are two possible confounders (omitted variables) that could bias the results from this regression? Explain the mechanism by which each omitted variable could bias the results and use the omitted variable bias formula to argue whether it would lead to an understatement or overstatement of the true effect. (3 points)

|  |
| --- |
|  |

1. What is/are the instrument(s) used by the authors in this study, and what are the authors instrumenting for? (2 points)

|  |
| --- |
|  |

1. Generally, what conditions must an instrument satisfy to be considered valid?
   1. Explain these conditions in broad terms and in the specific context of the instrument(s) used in the paper. (3 points)

|  |
| --- |
|  |

* 1. Explain these characteristics using random variables and potential outcomes. (2 points)

|  |
| --- |
|  |

1. Do you believe that the instrument(s) in the paper is/are truly exogenous? Why or why not? If so, provide a brief argument for this assumption. If not, provide an alternate mechanism through which the instrument(s) might affect the outcome variable, which suggests the exogeneity assumption may be violated. (2 points)

|  |
| --- |
|  |

1. To assess whether the instrument is relevant, we can examine whether the instrument (quarter of birth) predicts the instrumentalized variable (compulsory schooling).
   1. Explain how Table I is constructed, and give some intuition for the authors’ choices. (2 points)

|  |
| --- |
|  |

* 1. Interpret the coefficient of the first quarter for the outcome variables “Total years of education” and “High school graduate” for the 1930-1939 cohort. (2 points)

|  |
| --- |
|  |

* 1. Why do the authors estimate the coefficients displayed in the bottom part of Table 1 (“College graduate”, “Completed master’s degree”, “Completed doctoral degree”)? How do these results support the validity of their instrument? Which assumption of the IV model are they addressing here? (3 points)

|  |
| --- |
|  |

1. Consider Table III and Table IV. Provide a general formula and a basic intuition for the Wald estimator. How does it compare to the OLS estimate? What is the advantage of using TSLS, instead of the Wald estimator? (4 points)

|  |
| --- |
|  |

1. How would you construct a reduced form table? Why might you want to report reduced form estimates? What figure in the paper fulfills this purpose? (3 points)

|  |
| --- |
|  |

1. Subsequent papers have found that the instrument (quarter-of-birth) is weak for some specifications in the paper.
   1. What is the intuition for why weak instruments are problematic? (1 point)

|  |
| --- |
|  |

* 1. Read the following, explain the intuition, and explain the implications for weak instruments. (2 points)

A screenshot of a cell phone

Description automatically generated

|  |
| --- |
|  |

* 1. Optional: if you know what the bootstrap does, why does bootstrapping *not* solve the weak identification issue? (1 extra point)

|  |
| --- |
|  |

1. If you were to write the paper today, how would you detect weak instruments, and what statistic would you use for inference?

*Hint: you may want to refer to Andrews, Stock and Sun (2019).*

* 1. How is the effective F-statistic constructed? (1 point)
  2. How are Anderson-Rubin confidence sets constructed? (1 point)

|  |
| --- |
|  |

# The Local Average Treatment Effect

1. Explain the monotonicity assumption in the context of this study. What is required regarding the relationship between variables for monotonicity to be met, and is it reasonable to assume that defiers do not exist? In your explanation, be sure to touch on what it means to be a defier in this study. (3 points)

|  |
| --- |
|  |

1. Interpret the IV estimates in Table IV with appropriate units in the context of the study’s research question, treating them as a local average treatment effect. In your interpretation, clarify the population for which this local average treatment effect is identified (i.e., who are the compliers?). (2 points)

|  |
| --- |
|  |

1. In 3-5 sentences, discuss how these results might inform policy outside of this setting. In your discussion, be sure to comment on the challenges of generalizing instrumental variable findings. (2 points)

|  |
| --- |
|  |

# Data Analysis Questions (20 points)

The enclosed is a subsample from Angrist and Krueger’s dataset. Specifically, for men born between 1930 and 1939, it includes the following information from the 1980 Census:

* LWKLYWGE: log of weekly earnings
* EDUC: years of completed education
* YOB: year of birth
* QOB: quarter of birth
* Age, marriage status (1=married), race (1=black), urban dummy (SMSA, 1= center city)
* 8 region of residence dummies (NEWENG, MIDATL, ENOCENT, WNOCENT, SOATL, ESOCENT, WSOCENT, MT)

1. Figure I can be thought of as a “graphical first-stage”, as it shows the mean of completed years of education by quarter-of-birth for each year of birth between 1930 and 1939. Replicate Figure I, and highlight those born in the first quarter (for each year between 1930 and 1939) in your figure. (2 points)

*Hint: you may want to create year-of-birth and quarter-of-birth dummies. They will also be useful for the following questions.*

|  |
| --- |
|  |

1. Table I shows the relationship between quarter-of-birth and educational outcomes. Replicate the first row of Table I, i.e., find the coefficients of the first, second, and third quarter-of-birth dummies on total years of education. (2 points)

|  |
| --- |
|  |

1. Create a reduced form table that illustrates the relationship between quarter-of-birth and weekly earnings. In other words, regress log weekly earnings on the quarter-of-birth dummies (our instruments). Include year fixed effects. (2 points)

|  |
| --- |
|  |

1. Table III reports OLS and Wald estimates of returns of education. Replicate both estimates (in the last two rows) for men born between 1930-1939 (Panel B). *Hint: See footnote 13 in Angrist and Krueger (1991) for details on how they calculate the Wald estimate. Note that if you want to use the function felm, since there are no covariates, you will need to include 1 as a covariate (i.e., y ~ 1 | 0 | x ~ z).* (2 points)

|  |
| --- |
|  |

1. Table V reports different specifications of the TSLS for men born between 1930-1939. Run TSLS regressions replicating Column 2 and Column 6. For Column 2, instrument for education with a full set of quarter-of-birth times year-of-birth dummies, and include year fixed effects. For Column 6, instrument for education with the same set of quarter-of-birth times year-of-birth dummies, and include regional fixed effects, year fixed effects, race, urban, and married status dummies. (4 points)

|  |
| --- |
|  |

1. Now repeat the first TSLS regression in Question 18 (Column 2, without additional controls and only year fixed effects), but instead of using a built-in IV function, regress education directly on the instruments and then use predicted education to estimate the wage return of education. (Use lm if you choose to use R, since felm does not support predict). Do your results match your results from the previous question? (3 points)

|  |
| --- |
|  |

1. For this question, define treatment Xi as completing high school (12 or more years of education), and set the instrument as binary, with Zi equal to 1 if born in the fourth quarter, and 0 otherwise. The sample includes men born between 1930-1939.
   1. What is the share of the complier population? (1 point)

|  |
| --- |
|  |

* 1. What is the average untreated outcome (log of weekly earnings) for never-takers? (1 point)

|  |
| --- |
|  |

* 1. What is the average treated outcome (log of weekly earnings) for always-takers? (1 point)

|  |
| --- |
|  |

* 1. Is there selection into treatment? State the assumptions necessary for your conclusion. (2 points)

|  |
| --- |
|  |

# IVs in Your Own Work (8 points)

1. Think about a social relationship that would be best studied using an IV approach. Briefly state the research question and the main variables of interest in non-technical terms. (4 points)

|  |
| --- |
|  |

1. Write out the empirical specification you would use and explain the equation. (2 points)

|  |
| --- |
|  |

1. If you clustered your standard errors or included fixed effects, explain why these methods reduced the likelihood of bias in your results (and if applicable, in which direction). If you did not, explain why these methods were not appropriate in your setting. (2 points)

|  |
| --- |
|  |

1. We abide by the Harvard Kennedy School Academic [code](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another’s work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus. [↑](#footnote-ref-1)