PROBLEM SET 2

Due on Tuesday March 7, 2023 10:00 am.

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 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 R script** to the Canvas assignment page.

II - IDENTIFICATION

(1) Your information

Your Last Name:	*Chaturvedi*
Your First Name:	*Shreya*

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

Manisha Jha, Neha Verma

(3) Compliance with Harvard Kennedy School Academic Code[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	X	

[1] We abide by the Harvard Kennedy School Academic code 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.

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 (26 points)

1. Read the paper. In a couple of sentences, explain the primary research question that the authors are trying to answer. What is the key logic behind the authors' answer? In a separate sentence, explain why policymakers should care about that question. Use non-technical jargon so that someone without statistical training could understand. (2 points)

This paper examines whether compulsory school attendance affects individuals' educational attainment and earnings in the long run. The authors use the school start age policy and compulsory schooling requirements to support their argument that compulsory schooling laws increase educational attainment, leading to higher earnings.

The study's findings show that an additional year of compulsory schooling leads to a 7.5% increase in earnings for those compelled to attend school. The authors use the season of birth as an instrument to estimate the causal effect of compulsory schooling on earnings, which provides policy implications for the returns to education puzzle.

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

Two possible omitted variables that could bias the coefficient estimate for education upwards were identified: ability and socioeconomic background. The authors argued that not accounting for ability, which is difficult to measure, could lead to an overstatement of the true effect of education on earnings. Similarly, not including socioeconomic background, which can be dependent on factors such as race, class, and parental educational attainment, could also lead to an overestimation.

The omitted variable bias formula was used to argue whether the bias caused by these variables would lead to an understatement or overstatement of the true effect. In both cases, the bias would be positive, which means that individuals who attain higher levels of education would have higher earnings. Therefore, the effect of education on earnings would be overestimated if the confounding variables were not accounted for.

Since the selection bias term is positive, $E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] > TOT$

3. What is/are the instrument(s) used by the authors in this study, and what are those instruments instrumenting for? (1 point)

The quarter of birth, which is distributed randomly among students, is used as an instrument in this study to create exogenous variation in education level obtained. This is due to the combination of school age policies and compulsory schooling laws, which compels students born in certain months to attend school longer than students born in other months. The authors use this instrument to address the endogenous relationship between schooling years and earnings and to overcome any omitted variable bias issues that may arise when estimating the causal effect of compulsory schooling on earnings.

4. Generally, what characteristics must an instrument have for it to be considered valid?
   a. Explain these characteristics in words, and specific terms of the instrument(s) used in the paper. (4 points)

For an instrument to be considered valid, it must meet certain characteristics:
Firstly, it must be relevant and actually be correlated with the outcome being measured, such as how quarter of birth determines the number of years of schooling received.
Secondly, it must satisfy the exclusion restriction, meaning it only affects the outcome through the treatment, such as how quarter of birth only affects earnings through attending more school.
Thirdly, the instrument must satisfy the independence assumption and be randomly assigned, so it's not correlated with personal attributes that could bias results.
In this study, the quarter of birth is the instrument used and satisfies these characteristics by having a non-zero covariance with years of schooling and affecting earnings only through schooling and being as good as randomly assigned.

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

Relevance: cov(Z_i, X_i) != 0
Exclusion restriction: cov(Z_i, u_i) = 0
Independence: cov(Z_i, u_i) = 0

5. 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 for how the exogeneity assumption may be violated. (4 points)

While the instrument at the outset seems exogenous, as it is very difficult to establish a coherent causal link between the quarter of birth and factors such as parental income levels. However, there may be certain cultural or socioeconomic factors that could influence the likelihood of a child being born in a particular month within certain populations, which could in turn impact their parents' income levels. If this ends up being the case in our sample of study, it would violate the exogeneity assumption.

6. To assess whether the instrument is relevant, we can examine whether the instrument (quarter of birth) predicts the instrumentalized variable (compulsory schooling). In the following parts a – c, provide interpretations with concrete units, in a way a policymaker can understand.
   a. Explain how Table 1 is constructed, and give some intuition for the authors' choices. (2 points)

Table 1 reports the effects of different birth quarters on education indicators such as years of education and high school graduation rates for men born in the 1930s and 1940s. The authors remove the time trend to focus on seasonal variation and include post-secondary education to show no pattern in educational achievement by season of birth. The F-test column indicates statistical significance of the differences between birth quarters.

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

In the 1930s cohort of men, the first quarter of birth is associated with completing an average of 0.124 years less of schooling compared to the last quarter of birth. Furthermore, the first quarter of birth is associated with a 1.9 percentage point

decrease in the likelihood of graduating from high school compared to the last quarter of birth.

c. Why do the authors estimate coefficients for the bottom part of Table 1 ("College graduate", "Completed master's degree", "Completed doctoral degree")? How do the results support the validity of their instrument? What assumption of the IV model are they addressing here? (2 points)

The authors included post-secondary education in their analysis to examine whether seasonal differences in education were caused solely by compulsory schooling laws. The results show no seasonal variation in educational attainment among those with post-secondary education, which supports the theory that compulsory schooling laws are the driving force behind the seasonal variation among those in high school. This validates their instrument and addresses the exclusion restriction assumption in their IV model. The authors estimated coefficients for college, master's, and doctoral degrees to further test if differences in years of schooling by birth quarter were caused by compulsory schooling laws. The absence of statistically significant differences between birth quarters for post-secondary education supports the validity of their instrument.

7. 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? (3 points)

The Wald estimator, also known as the IV estimator, measures the causal effect of years of schooling on earnings by using birth quarter as an instrument.
Wald estimator (β) = Reduced Form/First Stage = $cov(Y_i, Z_i) / cov(X_i, Z_i)$.
It measures the strength of the relationship between the instrument and the endogenous variable, and it is not statistically significantly different from the OLS estimate.
However, the advantage of using TSLS is that it purges years of schooling of the variation that is related to the error term, allowing for a consistent and unbiased estimate of the causal effect, even in the presence of unobserved confounders.
The Wald estimator and TSLS provide different ways of estimating the causal effect of years of schooling on earnings, and they both have their advantages depending on the assumptions and data available.

8. How would you construct a reduced form table? What would be the purpose thereof? What figure in the paper fulfills this purpose? (3 points)

The purpose of a reduced form table is to estimate the effect of the instrument (birth quarter) on the outcome of interest (earnings), which would be used as the numerator to calculate the Wald estimator. Figure V in the paper fulfills this purpose by showing the relationship between log weekly earnings and quarter and year of birth, revealing a sawtooth pattern. The reduced form table can be constructed by regressing ln (earnings) by the quarter of birth, to estimate the effect of birth quarter on earnings. The coefficients associated with ln (weekly wage) in Table III give us the reduced form.

9. What would it mean for the instrument to be weak in this context? What would be the result thereof? (2 points)

When the instrument is considered weak in this context, it means that the instrument, quarter-of-birth, only explains a small proportion of the variation in schooling. If this is the case, then relevance is compromised, and bias may occur because quarter-of-birth is no longer a reliable instrument to establish the causal link between schooling and income. This may lead to imprecise estimates and/or biased inferences, primarily because of large standard errors, making it difficult to detect statistically significant effects even when they exist. A weak instrument occurs when there is a relevance issue in the first stage, indicated by a weak relationship between birth quarter and years of schooling, leading to an undefined Wald/IV estimate.

The Local Average Treatment Effect (10 points)

10. 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 does it means to be a defier in this study. (4 points)

Monotonicity assumption in this study requires the birth quarter to weakly impact treatment status in the same way/direction for everyone in the data, which means there should be no defiers. To meet the monotonicity assumption, an increase in the instrument should always lead to an increase or no change in the likelihood of attending school. The authors provide evidence to support the validity of the assumption by showing that the instrument has a positive effect on the likelihood of attending school in all specifications. Defiers in this study would be individuals who comply with compulsory schooling laws only if they do not apply to them, but do not comply if the laws do apply to them. It is hard to determine if there are any defiers in this study, but it is reasonable to assume that there are none. If defiers exist, they would violate the monotonicity assumption and affect the reliability of the causal inference.

11. 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?). (3 points)

The IV estimates in Table IV show that an additional year of schooling for men born between 1920-1929 leads to an increase in weekly wages by 7.69% (2), 13.1% (4), 6.69% (6), and 10.07% (8), after controlling for various factors. These estimates are based on exogenous variation in the quarter someone is born, which compels those born earlier in the year to have fewer years of schooling. The local average treatment effect is identified for compliers, who are those born later in the year and have more years of schooling compared to those born earlier in the year who have fewer years of schooling.

12. With 3-5 sentences, discuss the how you believe these results inform policy outside of the specified context. In your discussion, be sure to touch on the problems specific to generalizing from instrumental variable findings and the particular scope conditions of this paper's findings. (3 points)

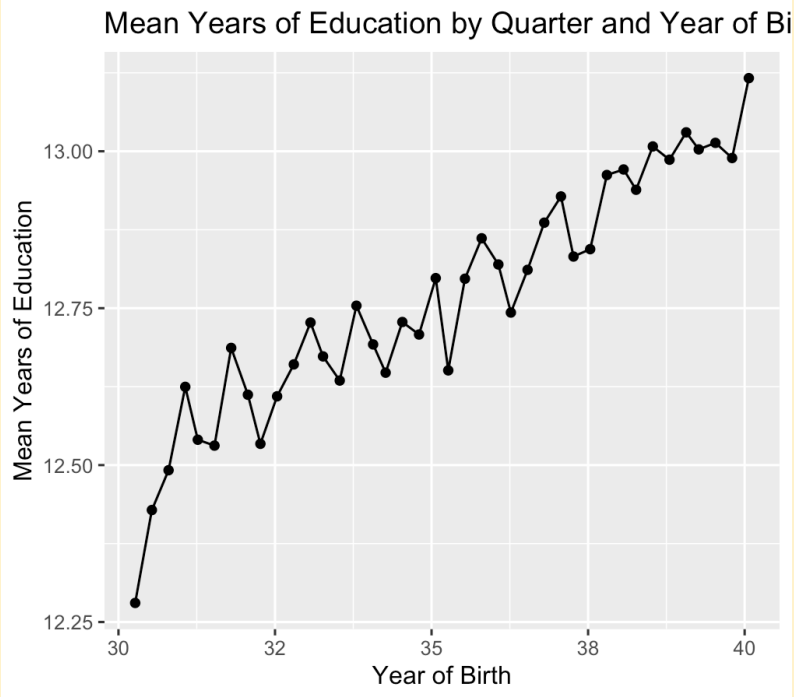
This study has important implications for policymakers when interpreting results and making decisions based on instrumental variable findings. While IVs can be useful for estimating causal effects, it's important to ensure that the population being studied is relevant to the policy being considered, and to be aware of potential issues with the instrument used. Furthermore, the study's finding that the difference between the causal effect using an IV estimate and the OLS estimate for returns to education is not statistically different suggests that omitted variable bias may not be a major issue in some contexts.

Data Analysis (16 points + 3 extra 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)

13. 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 1940. Replicate Figure I. You can highlight the mean of years of education of those born in the first quarter (for each year between 1930 and 1940). (4 points)



14. 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. Your estimates do not need to be exactly the same as in the paper, but roughly of the same magnitude. (4 points)
Hint: unlike the authors, we can remove the first two and the last two quarters in our subsample. Our estimates will hence be different.

The Effect of Quarter of Birth on Total Years of Education

	Model 1
Birth cohort	1930-1939
Mean Years of Education	12.78
(Intercept)	0.053
	(0.012)
QOB1	−0.119
	(0.017)
QOB2	−0.081
	(0.017)

15. **[Optional]** Create a reduced form table. In other words, you want to regress the log of weekly earnings on the quarter of birth dummies (our instrument). You may want to include year fixed effect. (3 points)

16. Table III reports OLS and Wald estimates of returns of education. Replicate both estimates for men born 1930-1939 (Panel B). (4 points)

OLS and Wald estimates of returns of education

	OLS	WALD ESTIMATE
(Intercept)	4.995	4.597
	(0.004)	(0.306)
EDUC	0.071	
	(0.0003)	
EDUC(fit)		0.102
		(0.024)

17. Table V reports different specifications of the TSLS for men born 1930-1939. Run TSLS similar as in Column 2 and Column 6. First, instrument education with a set of quarter-of-birth x year-of-birth dummies, and add year fix effects as control. Why do we want to include year fix effects?
Second, similarly to Column 6, instrument education with the same interaction dummies, and add regional fix effects, year fix effects, race, married status, and an urban dummy. (4 points)

TSLS ESTIMATES WITH FIXED EFFECTS

	(2) TSLS	(6) TSLS
EDUC_COEF	0.089	0.081
	(0.017)	(0.017)
RACE		−0.230
		(0.027)
SMSA		−0.158
		(0.018)
MARRIED		0.244
		(0.005)

IVs in Your Own Research

18. 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. (2 points)

Relationship between increased access to public transportation reduce unemployment rates for low-income individuals:
Variables of interest: Access to public transportation (proximity to nearest bus or train station) and unemployment rate among low-income individuals.
OLS regression may not accurately estimate the relationship between access to public transportation and unemployment because there may be unobserved factors that affect both variables. To address this, an IV approach could use changes in public transportation funding as an instrument to determine the causal effect of transportation on unemployment for low-income individuals. By doing so, we can account for the potential endogeneity of the relationship.

19. Write out the empirical specification you would use and explain the equation. (4 points)

$Unemployment_i = \gamma_0 + \gamma_1 Instrument_i + \delta_1 Transportation_i + \varepsilon_i$ *where:*

- *$Instrument_i$ is the instrumental variable (changes in public transportation funding)*
- *γ_1 is the coefficient of interest, representing the causal effect of changes in public transportation funding on access to public transportation*
- *δ_1 is the coefficient of interest, representing the causal effect of access to public transportation on unemployment*
- *γ_0 is the intercept term*
- *ε_i is the error term*

20. 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)

Further Hints for Data Analysis

Question 13

- Create year of birth and quarter of birth dummies- this will also be useful later.
- Find the average number of years of education by quarter of birth for each year.
- It is easier to plot if you create a "continuous variable" made of year + quarter of birth.

Question 14

- You will need to 'detrend' the years of education variable by subtracting a moving average, as described in the paper on p. 985.
    - You can use a for loop to do this.
- The outcome variable is education minus the moving average you created.
- The quarter of birth dummies you created before will be useful here.

Question 16

- For the Wald estimator, you can use the function `ivreg` from the `AER` package or `felm` from the `lfe` package.

Question 17

- Be careful how you include instruments versus other covariates in your formula.
    - If you're using `felm`, remember the syntax is:
      y ~ x1 + x2 | f1 + f2 | (Q|W ~ x3+x4) | clu1 + clu2
      where x3 and x4 are instruments for Q and W. x1 and x2 would be covariates. clu1 and clu2 are to cluster standard errors – don't have to worry about these in this case. f1 and f2 are fixed effects. You can also include dummies as covariates as in x1 and x2 instead.
    - If you're using `ivreg`, you would have to repeat the "non-instrument" covariates before and after the | separator, meaning:
      y ~ endogenous variable + x1 + x2 | instrument(s) + x1 + x2