**Questions week 5**

**Please download from CANVAS the file “ak91.csv”.**

**The file contains information the wages people earned and the years of schooling they experienced, plus a few other variables.**

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| wage | Salary in thousands of USD |
| Schooling | Years of schooling |
| qob | Quarter of birth |
| yob | Year of birth |

**The goal of this exercise will be use an instrumental variable (IV) strategy to analyse the returns to schooling using quarters of birth (QOB) as instruments for years of schooling in the US. *Background: In the United States, the grade level a child starts school at is typically determined by their birth date. There was a time when the cut-off was the end of December. This meant that if a child was born on or before December 31, they were enrolled in the first grade, while those with birthdays on January 1 or later began in kindergarten. Consequently, two children born a day apart could end up in different grades.***

***Assuming that these children remain in school for the necessary amount of time to earn a high school diploma, the specific starting grade won't impact whether they graduate. It only affects when they graduate. But it's important to note that for much of the twentieth century, compulsory education laws in the U.S. mandated school attendance until the age of 16. Once a child turned 16, they were legally permitted to leave school. The image below illustrates this point (S=Schooling):***

A picture containing text, line, receipt, diagram

Description automatically generated

*Source: Scott Cunningham (Causal Inference – the Mixtape)*

1. **Draw a causal diagram for our scenario. Discuss based on your diagram why we may face a situation in which we have unobserved confounders if we are trying to answer the question: How does another year of schooling affect your salary?**
2. **Let’s start with some descriptive analysis:**
   1. **Should we log-transform our outcome variable (wage)?**
   2. **Check the frequency table for ‘qob’ for our sample to check for balance.**
   3. **Plot the schooling year average (per person) against the quarters of birth (qob) across time for our data.**
   4. **Check the mean of wages (according to your choice in a) across all treatment levels (quarters of birth = qob) as well as for quarter 4 vs quarters 1,2,3 only (we need to create a dummy for this).**
   5. **Check the mean of schooling across all treatment levels (quarters of birth = qob) as well as for quarter 4 vs quarters 1,2,3 only (we need to create a dummy for this).**
3. **Check whether ‘qob’ is a strong enough instrument for schooling (*tip: you need to use an F-test*).**
4. **Let’s now perform some regression analysis.**
   1. **Perform a naïve estimate of the impact of schooling on wage, in line with your choice from B a) with and without controlling for ‘year of birth’ (yob). What is the effect of one more year of schooling on wages?**
   2. **Let’s now perform 2SLS using ‘qob’ as instrument (with and without YOB as control). What do you find now?**
   3. **Let’s estimate a proxy for the ITT. Let’s regress the wage on the dummy for the fourth quarter of birth vs. rest (in line with your choice of b).**
   4. **Check whether you can trust the standard errors of your 2SLS from b) and c).**
5. **Which estimand does the IV estimator recover for this data set in your view? Why?**
6. **Consider the following tables of theoretical estimates from a 2SLS regression: We run online advertisements that encourage people to exercise more. You want to use exposure to the advertisements as an instrument for how much you exercise, and then will look at the effect of exercise on blood pressure. The theoretical coefficients for three people are shown below in the table. Your 2SLS equations are:**

* exercise\_hours = α0 + γ ad\_exposure
* blood\_pressure = α1 + ẞ exercise\_hours\_hat

|  |  |  |
| --- | --- | --- |
| **Name** | **Effect of ads on exercise\_hours (**γ) | **Effect of exercise\_hours\_hat on blood\_pressure (ẞ)** |
| Jakeila | **0.5** | **-2** |
| Kyle | **0.25** | **-8** |
| Li | **0.0** | **-10** |

* 1. **What is the LATE for this sample of people in the table?**
  2. **What is the ATE for this sample of people in the table (if there was full compliance)?**
  3. **Which assumptions are required to obtain LATE in our 2SLS regression?**
  4. **Which estimand do you think does the IV estimator yield? LATE or ATT? Why?**