**Causal Inference Practice 2**

1. Virginia's community college system is considering a transition toward more on-line classes in the future but wants to assess the potential impact on students.

You are brought in as a consultant to help them. They provide you with the following dataset of past student performance that includes data from two semesters. One prior to Covid-19 and one during Covid-19 when all class (regardless of what a student signed up for) were moved online.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student ID** | **Semester** | **Class** | **Scheduled** | **Mode** | **Final Grade** |
| 1 | 2020-1 | Calculus | Online | Online | 75 |
| 2 | 2020-1 | Calculus | Online | Online | 80 |
| 3 | 2020-1 | Calculus | Online | Online | 85 |
| 4 | 2020-1 | Calculus | In-Person | In-Person | 85 |
| 5 | 2020-1 | Calculus | In-Person | In-Person | 90 |
| 6 | 2020-1 | Calculus | In-Person | In-Person | 95 |
| 7 | 2021-1 | Calculus | Online | Online | 60 |
| 8 | 2021-1 | Calculus | Online | Online | 65 |
| 9 | 2021-1 | Calculus | Online | Online | 85 |
| 10 | 2021-1 | Calculus | In-Person | Online | 50 |
| 11 | 2021-1 | Calculus | In-Person | Online | 65 |
| 12 | 2021-1 | Calculus | In-Person | Online | 95 |

Given the situation and (only) this dataset….

1. What empirical strategy could you implement to try to estimate the causal effect of taking a class online? Write out the regression specification for this strategy in mathematical notation.

1. What is your estimate of the average causal effect of taking class online?

1. What assumption are you making to generate your counterfactual? Quantify and explain the counterfactual in this case.

1. The federal government is trying to assess the impact of its early childhood education program for children from low income families (Head Start) on children's grades in elementary school.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Family ID (f)** | **Family Name** | **Child ID (i)** | **Child Name** | **Head Start Attendance**  **(HS)** | **Grade 3**  **Test Score**  **(Grade)** |
| 1 | Smith | 1 | Alex | Yes | 95 |
| 1 | Smith | 2 | Bo | No | 90 |
| 2 | Brown | 3 | Cassie | Yes | 80 |
| 2 | Brown | 4 | Devon | Yes | 90 |
| 3 | Garcia | 5 | Edgar | Yes | 85 |
| 3 | Garcia | 6 | Frank | No | 70 |
| 4 | Jones | 7 | Jenny | No | 100 |
| 4 | Jones | 8 | Kaz | No | 90 |

1. Find the estimate of that you would find by running the following regression:

1. Is your estimate of likely to be an unbiased estimate of the causal effect of Head Start on Grade 3 test scores? Explain.

1. Devise an alternate regression specification (given the data above) that would likely remove some of the bias in your estimate of in (a).

1. Will all of the observations in the data above contribute to your estimate of the effect of HS in your regression in (c)? If not, which ones?

1. Find your new estimate of the average causal effect of Head Start.
2. Using the above dataset (hs\_family.dta) and the i. functionality for the reg function in STATA, replicate your answers to a) and e).
   1. Interpret each coefficient.
   2. Why are there only 3 family coefficient estimates?
3. Implement fixed effects regression manually by:
   1. Creating a new variable with the mean grade by family (egen command)
   2. Creating a new variable with the mean HS attendance by family (egen command)
   3. Create a de-meaned grade and HS attendance variable (gen command)
   4. Regress the de-meaned grade on the de-meaned HS attendance (reg command)
4. Several states in the Northeast implemented expanded school breakfast programs in the early 1990s in an effort to improve test scores. Consider the data below and accompanying stata dataset (gendd\_panel.dta):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **State** | **State Name** | **Year** | **Program** | **Test Score** |
| 1 | PA | 1990 | 0 | 0 |
| 1 | PA | 1991 | 0 | 10 |
| 1 | PA | 1992 | 1 | 40 |
| 1 | PA | 1993 | 1 | 50 |
| 2 | NJ | 1990 | 0 | 10 |
| 2 | NJ | 1991 | 1 | 30 |
| 2 | NJ | 1992 | 1 | 50 |
| 2 | NJ | 1993 | 1 | 60 |
| 3 | NY | 1990 | 0 | 30 |
| 3 | NY | 1991 | 0 | 40 |
| 3 | NY | 1992 | 0 | 60 |
| 3 | NY | 1993 | 0 | 70 |
| 4 | MA | 1990 | 1 | 60 |
| 4 | MA | 1991 | 1 | 70 |
| 4 | MA | 1992 | 1 | 90 |
| 4 | MA | 1993 | 1 | 100 |

1. Devise a regression specification for estimating the average causal effect of the program.
2. Using gendd\_panel.dta and the i. functionality for the reg function in STATA estimate this regression.
   1. Interpret each coefficient.
   2. Why are there only 3 state coefficient estimates?
3. Implement the generalized difference-in-difference strategy manually by:
   1. Creating a new variable with the mean test score by state (egen command)
   2. Creating a new variable with the mean program by state (egen command)
   3. Create a state de-meaned score and program variable (gen command)
   4. Review the state de-meaned variables (browse command), where is there within-state variation that will be leveraged?
   5. Using the state de-meaned variables, repeat a) through c) removing the means by year from these variables.
   6. Regress de-meaned score on de-meaned program variable. Compare to b).
4. How do MA and NY factor into your estimation?
5. Every year, the federal government provides $68 billion in Earned Income Tax Credits (EITC) to 28 million households. The amount of money that households receive as part of this program is determined by their pre-tax income and the number of children in the household.

In 1995, there was a major change to the program: the credit formula was changed to give a substantially larger credit to 2+ children households (there was no change for 1 child households).

You obtain the following dataset from the Department of Defense school system (DODEA). They are interested in understanding the *effect* of receiving this larger tax credit on student test scores so they determine how best to target their resources.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Child Name** | **Year** | **No. of Children in Household** | **Pre-Tax Income** | **Test Score** |
| Andre | 1994 | 1 | 4,000 | 40 |
| Beth | 1995 | 1 | 4,000 | 35 |
| Cory | 1994 | 2 | 4,000 | 50 |
| Devon | 1995 | 2 | 4,000 | 50 |
| Esther | 1994 | 1 | 10,000 | 75 |
| Faye | 1995 | 1 | 10,000 | 70 |
| Garth | 1994 | 2 | 10,000 | 80 |
| Heather | 1995 | 2 | 10,000 | 90 |

1. In words, briefly discuss the “best” (i.e. least restrictive) assumption that you could make to construct a counterfactual for the test score outcomes of children that are “treated” with a larger tax credit (given this situation and the data you have available above). Be precise!
2. Given the assumption you discuss in (a), calculate the counterfactual for Devon. Show your work.
3. Propose a regression specification to implement your intuition in part (a). Be sure to define all variables and underline the coefficient of interest (i.e. the coefficient that captures the causal effect of additional family income on test scores).