**Causal Inference Practice 1**

1. Governor Cuomo asks for your opinion regarding the effect of attending pre-K on children’s later adult income *(note: a child’s pre-K attendance is voluntary and the choice to attend is made by the child’s parents)*. He provides you with the following dataset from New York:



1. Adapting this situation to the notation of the Potential Outcomes (Rubin) Model, **calculate** $E\left[D\_{i}=1\right]$ and $E\left[D\_{i}=0\right]$.
2. One of the governor’s aides shows him a chart with $E\left[D\_{i}=1\right]$ and $E\left[D\_{i}=0\right]$ and tells the governor that the difference represents the causal effect of Pre-K. Is this correct? Respond to this claim in two ways (be precise!):
3. Using the notation of the Potential Outcomes (Rubin) model (i.e. mathematical notation).
4. Using **plain** English (1-2 sentences max).
5. Another aide shows the governor the estimate of $β\_{1}$ from the following regression and tells him that this represents the causal effect of Pre-K:

$$Y\_{i}=α+β\_{1}PK\_{i}+β\_{2}HighInc\_{i}+u\_{i},$$

 where $Y\_{i}$ is the income at age 30 of child i, $PK\_{i}$ is an indicator equal to 1 if child i attended pre-K, $HighInc\_{i}$ is an indicator equal to 1 if child i’s family was “high income”, and $u\_{i}$ is an error term.

1. What is the comparison implied by this estimate of $β\_{1}$? In other words, what is the estimated counterfactual for children who attended Pre-K? Explain briefly and **be precise**.
2. In 1-2 sentences maximum, argue whether you believe $β\_{1}$ is an accurate estimate of the *average causal effect* of Pre-K attendance on this population. **Be as specific as possible** (for example, by providing a specific scenario in support your argument).
3. 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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Class** | **Mode** | **Final Grade** |
| 1 | Alex | Calculus | Online | 60 |
| 2 | Bo | Calculus | Online | 75 |
| 3 | Cassie | Calculus | In-Person | 58 |
| 4 | Devon | Calculus | In-Person | 72 |
| 5 | Edgar | Calculus | In-Person | 80 |
| 6 | Frank | Calculus | In-Person | 85 |
| 7 | Greta | Psychology | Online | 70 |
| 8 | Henry | Psychology | Online | 70 |
| 9 | Isaiah | Psychology | Online | 85 |
| 10 | Jenny | Psychology | Online | 90 |
| 11 | Kaz | Psychology | In-Person | 85 |
| 12 | Lawrence | Psychology | In-Person | 100 |

1. If you were to run the following regression, $y\_{i}=α+βOnline\_{i}+u\_{i}$ , where $Online\_{i}$ is an indicator variable (=1 if online, =0 otherwise) and $u\_{i}$ is an error term…

i)What is the counterfactual for student's taking online classes that this regression constructs?

ii) What would your estimate of $β$ be?

 iii) What assumption would be required for this estimate to be the average causal estimate of the effect of taking a class online?

1. If you were to run the following regression, $y\_{i}=α+βOnline\_{i}+γCalculus\_{i}+u\_{i}$ , where $Calculus\_{i}$ is an indicator variable (=1 if Calculus class, =0 otherwise)…

i) What is the counterfactual for student's taking online classes that this regression constructs?

ii) What would your estimate of $β$ be in this case?

 iii) What assumption would be required for this estimate to be the average causal estimate of the effect of taking a class online?

iv) What is the omitted variable bias in your estimate of $β$ in regression (1)? In plain english, what accounts for this bias?

1. At Fort Campbell, soldiers that are exiting the army have the option to enter training program (TAP) aimed at helping them transition to the civilian labor force. The Pentagon is reviewing the program and wants to assess its impact on unemployment duration after exiting the Army with the dataset below.



1. The initial report presents the effect of TAP as the estimate of $β\_{1}$ in the following regression:

$$Y\_{i}=α+β\_{1}TAP\_{i}+u\_{i},$$

where $TAP\_{i}$ is an indicator (dummy) variable equal to 1 if a soldier attended TAP (and 0 otherwise) and $Y\_{i}$ is months unemployed after exiting the Army. Using the dataset above, **calculate** this estimate of $β\_{1}$. Show your work (your answer should be a number).

1. You decide that it is important to control for soldier ability and run the following regression:

$$Y\_{i}=α+δ\_{1}TAP\_{i}+δ\_{2}HighAbility\_{i}+ ϵ\_{i},$$

where $HighAbility\_{i}$ is an indicator (dummy) variable equal to 1 if a soldier is high ability (and 0 otherwise). Using the dataset above, **calculate** this estimate of $δ\_{1}$. Show your work (your answer should be a number).

1. Quantify the bias in the initial report’s estimate arising from the failure to control for soldier ability in the initial report. Show your work (your answer should be a number).
2. **Explain** **why** the initial report did or did not produce as accurate an estimate of the effect of the program. Be specific and use plain English. (1-2 sentences max)
3. You are tasked with identifying the causal effect of weekly personal training on the AIOCT times of cadets. Before leaving, your predecessor produced an estimate of $β\_{1}$ from the following regression specification:

$$T\_{j}=β\_{0}+β\_{1}Train\_{j}+u\_{j},$$

where $T\_{j}$ is the AIOCT time of cadet j, $Train\_{j}$ is a equal to 1 if cadet j receives training (=0 otherwise), and $u\_{j}$ is an error term.

* 1. What would have to be true for your predecessor’s estimate of $β\_{1}$ to be the causal effect of the program?
	2. If DACs tend to assign heavier cadets to personal training what would that mean for estimate of $β\_{1}$? How could you change the specification above to improve your estimate?
	3. Now consider if no one receives personal training in 2013, 2014, and 2015, and then only English and Engineering majors receive personal training in 2016 and 2017. What would that mean for estimate of $β\_{1}$? How could you change the specification above to improve your estimate?