Problem Set 6

Professor: Christopher Lucas

Due Monday, March 2

You are required to use either `knitr` or `Markdown`, these are fully compatible with `R` and LaTeX.

Please (1) bring a printed version of your answers to class and (2) upload an electronic version on Canvas. Hard copies of your solutions should be handed in at the start of class and electronic copies are due before class begins at 4:00 pm. Late submissions will not be accepted. Please, comment our code and show your work in questions that require calculations.

Please set your seed using set.seed(02139) when the question requires simulations.

1 Fictitious Observational Study

The table below summarizes outcomes from a fictitious observational study assessing whether receiving assistance from one's elected representative after making a request for assistance increases political self-efficacy. Data come from a survey of 800 constituents who made requests to their representatives. The survey measured political self-efficacy using a thermometer-style scale ranging from 0 to 100. We'll denote the outcome Y_i, for constituents $i = 1, \ldots, N$, and the treatment $D_i \in \{0, 1\}$. Our quantity of interest is the average effect of receiving a representative's assistance on a constituent's self-efficacy, $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, the ATE.

D_i	$\mathbb{E}[Y_i]$
1	73
0	56

The survey also measured constituent income, coded in our data as a discrete covariate $X_i \in \{1, 2, 3\}$, indicating low, middle, and high income. In the sample, 50% of constituents are low-income, 30% are middle-income, and 20% are high-income. Frequencies by income level are displayed in the following table.

Freq.	$X_i = 1$	$X_i = 2$	$X_i = 3$	Total
$D_i = 1$	212	163	125	500
$D_i = 0$	188	77	35	300
Total	400	240	160	800

Throughout this problem, assume that the observed data represent the entire population of interest and the variables are measured without error, so that we can ignore sampling variability. (In real applied settings, you would want to incorporate sampling uncertainty and report confidence intervals, etc., as part of your result.)

(a) Write an expression for τ using the expressions in the following table. Then create a similar table that replaces the expressions in each cell with actual values, using only the observed data from this study. Leave the original expression in any cell you cannot populate with a value from the observed data.

d	$\Pr(D_i = d)$	$\mathbb{E}[Y_i(1)\|D_i = d]$	$\mathbb{E}[Y_i(0)\|D_i = d]$
1	$\Pr(D_i = 1)$	$\mathbb{E}[Y_i(1)\|D_i = 1]$	$\mathbb{E}[Y_i(0)\|D_i = 1]$
0	$\Pr(D_i = 0)$	$\mathbb{E}[Y_i(1)\|D_i = 0]$	$\mathbb{E}[Y_i(0)\|D_i = 0]$

(b) Suppose you have some reason to be concerned about the identifying assumption of conditional ignorability. Calculate sharp bounds for τ without making any assumptions at all. Explain in simple, concise language what these bounds represent. You may find it helpful to use the table you produced in part (a) and fill in the values that produce the upper and lower no-assumptions sharp bounds.

(c) One possible threat to identification is that constituents with higher self-efficacy are more likely to follow up on their requests, and in turn more likely to receive assistance from their representatives (assume follow-up requests are not captured in the survey). If this were the case, treated constituents would tend to have greater self-efficacy than untreated units, whether or not they received treatment. State this assumption formally and use it to calculate new bounds.

(d) Now consider that elected representatives have limited time and resources and must make strategic decisions about which requests to prioritize. Representatives want to maximize the probability of winning re-election, and they know there are greater payoffs to responding to requests from high income people, who more likely to both turn out and make large campaign donations. Assume representatives can accurately guess a constituent's income level based on factors like the constituent's address and the type of request. As a result of this incentive structure, the probability of response to requests from high-income constituents is unrelated the number of times they follow up (and therefore unrelated to pre-existing levels of self-efficacy). For low- and middle-income constituents, however, positive self-selection may still hold. Express this assumption formally. Would the resulting bounds be more or less credible than those under your assumption in (c)? Explain.

(e) Use the assumption described in (d) and the following table of conditional expectations to calculate new bounds.

D_i	$\mathbb{E}[Y_i]$	$\mathbb{E}[Y_i\|X_i = 1]$	$\mathbb{E}[Y_i\|X_i = 2]$	$\mathbb{E}[Y_i\|X_i = 3]$
1	73	69	73	83
0	56	51	55	70

2 The Consequences of Child Soldiering

In this problem you will analyze the data in `child_soldiering.csv`, from the Blattman and Annan (2010) article "The Consequences of Child Soldiering."[1] The authors are interested in the impact of abduction by the Lord's Resistance Army on political, economic, and psychological outcomes. The data come from a survey of male youth in war-afflicted regions of Uganda. We will focus on the effect of abduction, which appears in the data as `abd`, on years of education, `educ`. Other variables in the data are:

- `C.ach`—`C.pal`: sub-district identifiers
- `age`: respondent's age in years
- `fthr.ed`: father's education (years)
- `mthr.ed`: mother's education (years)
- `orphan96`: indicator for whether parents died before 1997
- `fthr.frm`: indicator for whether father is a farmer
- `hh.size96`: household size in 1996

(a) Estimate the ATE of abduction on years of education with OLS. Fit the model

$$\mathbf{y} = \hat{\alpha} + \mathbf{d}\hat{\beta} + \mathbf{X}\hat{\gamma} + \hat{\mathbf{u}}$$

Where $\hat{\alpha}$ is an intercept, $\hat{\beta}$ is the OLS estimator for the ATE, $\mathbf{d}$ is an $N \times 1$ vector indicating treatment status for units $i \ldots N$, $\hat{\gamma}$ is a 4×1 vector of coefficients, and $\mathbf{X}$ is an $N \times 4$ matrix of possible observed confounders: age, father's education, mother's education, and residency in Acholibur (`C.ach`); and $\hat{\mathbf{u}}$ is a vector of residuals. Report your results in a regression table.

What *identification assumptions* are necessary to interpret the ATE estimate produced from this regression model as the causal effect of abduction on years of education? What additional assumptions are required for the OLS estimator $\hat{\beta}$ to be unbiased and consistent for the ATE?

(b) Thinking about the context of the experiment (see `http://www.mitpressjournals.org/doi/pdfplus/10.1162/REST_a_00036`the paper for more details), come up with a possible unobserved binary confounder that predicts both education and abduction. Describe your binary confounder in words, then draw a DAG to represent the relationships between abduction (D), education (Y) and your potential confounder (U).

(c) Now we will consider what the bias created by omitting our unobserved confounder would have to be to cut the ATE in half.

[1] Reading the article is not necessary for completing the problem, but can be found at `http://www.chrisblattman.com/documents/research/2010.Consequences.RESTAT.pdf`.

i) Write down an analytical expression for the bias created by omitting U, assuming that the average effect of U on Y is the same for treated and untreated units. Is this assumption plausible for the confounder you identified in part (b)?

ii) For now, assume it is plausible. Create a contour plot of the bias that would reduce the magnitude of the ATE by half, where the x-axis is δ (the difference in average U_i between treatment conditions) and the y-axis is γ (the effect of U on Y). Refer to slide 13 of the Observational Studies Part III: Nonparametric Bounds and Sensitivity Analysis lecture for more precise definitions of these sensitivity parameters (δ and γ).

(d) Now we can see the values of δ and γ that would reduce our ATE by half. However, in order to get a sense of whether it is plausible that our unobserved confounder could create this amount of bias, we need to compare it to something. To do this, we will compare the hypothetical degree of confounding based on the values of our sensitivity parameters with the actual degree of confounding created by some observed covariates. For each observed covariate, the x-axis value δ represents the relationship between that covariate and abduction, while the y-axis value γ represents the relationship between that covariate and the outcome, both conditional on other covariates.

Start by adding the binary covariate residency in Acholibur (`C.ach`) to the plot as a benchmark. Next, add the covariates `age` and `fthr.ed`. To make these last two covariates easier to visualize, assume they have a negative relationship with our outcome, education, but of the same magnitude as the observed relationship.

Label all the points on your plot. Discuss how predictive the unobserved confounder would need to be in relation to the benchmarks to reduce the ATE by half. What do your results suggest about threats to inference from an unobserved confounder?