Introduction to Causal Inference

Problem Set 3

Professor: Teppei Yamamoto

Due Monday, July 18 (at beginning of class)

Only the required problems are due on the above date. The optional problems will not directly count toward your grade, though you are encouraged to complete them as your time permits. If you choose not to work on the optional problems, use them for your self-study during the summer vacation.

Required Problems

Problem 1

After months of deep thought, you decide that your dissertation will focus on the relationship between X and Y. In preparation for the meeting with your committee, you read all the relevant literature, conduct field interviews, and finally write down a directed acyclic graph that you believe captures all of the relevant variables, and their inter-relationships. You proudly present Figure 1 to your committee.

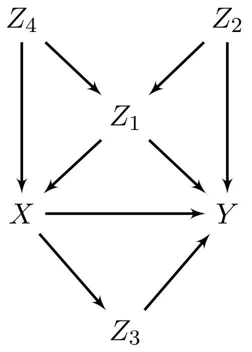

Figure 1: Your Dissertation In A Graph

(a) Your committee chair, who is not used to DAGs, asks you to explain the purpose of your DAG. As a researcher, what does it buy you, and what does it not buy you?

(b) Explain, in your own words, Pearl's back-door criterion.

(c) For Figure 1, enumerate all paths from X to Y.

(d) In the path $\{X \leftarrow Z_4 \rightarrow Z_1 \leftarrow Z_2 \rightarrow Y\}$, what type of node is Z_1? Does conditioning on Z_1 block or unblock this path from X to Y?

(e) Now consider path $\{X \leftarrow Z_1 \rightarrow Y\}$. In this path, what type of node is Z_1? Does conditioning on Z_1 here block or unblock this path?

(f) Previous research in your area has exclusively conditioned on Z_1, claiming that this is necessary and sufficient to identify the relationship between X and Y. Given Figure 1, does conditioning on Z_1 satisfy the back-door criterion for identifying the effect of X on Y? Why or why not?

(g) Based on your DAG, enumerate the minimum conditioning sets that satisfy the back-door criterion for identifying the effect of X on Y.

(h) **(Optional)** Consider the DAG in Figure 2. To identify the effect of X on Y, should we include M in our conditioning set? Why or why not? What is this an example of?

Now, use `R` to demonstrate your answer by simulation. Repeatedly simulate data from a data generating process consistent with the DAG in Figure 2. You may choose particular parameters, values, or distributions for each variable as you so wish, as long as the causal relationships between the variables match those encoded in the DAG. Then, for each simulated sample from the DGP, estimate the ATE of X on Y both with and without conditioning on M. Constrast your results graphically, and interpret whether they support your answer above.

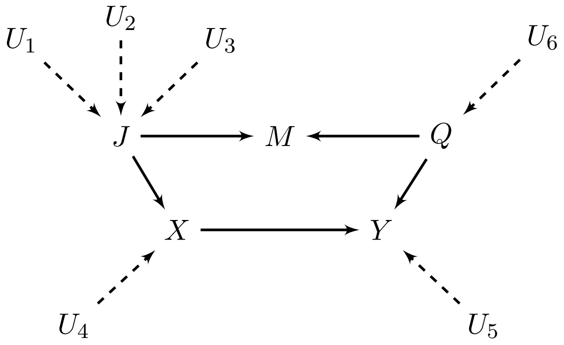

Figure 2: Directed Acyclic Reindeer

Problem 2

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;

[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.

the data are 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.akw`, etc.: 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) Check covariate balance in the unmatched dataset using the `Matching` package's `MatchBalance()` function, for all covariates. Your output should be in the form of a balance table. Make sure to present statistical tests of the equality of means and equality of distributions. Based on your table, which of the observed covariates seem to be the most important factors in selection into abduction? Using the (naïve) difference-in-means estimator, estimate the ATE of abduction on education without adjusting for any of the covariates, and report your estimate along with a standard error.

(b) Consider the authors' description of abduction:

> Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks. Most were abducted after 1996 and from one of the Acholi districts of Gulu, Kitgum, and Pader.
>
> Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release.

Given this description, choose some covariates on which to match, and then do so using the `Match()` function. Be sure to carefully check the options available to you in this function. For now, use only one match for each treated unit, use the Mahalanobis distance metric to determine weights, and do not (yet) use exact matching. Apply the matching estimator to estimate the average effect of abduction on education among those who were abducted, i.e., the ATT. Report your estimate and standard error, as well as balance statistics for the matched data.

(c) Re-estimate the ATT using exact matching on `C.ach, C.akw, C.ata, C.oro, C.paj`, and `age`. Report your estimate, its standard error, and produce a balance table as before. (**Optionally**, also show a balance plot that compares the differences between treated and control for this match and the match in (b), similar to the plot shown in slide 40 of the lecture slides.) In general, do your results differ from previous results? Why or why not?

(d) Re-estimate the ATT as in (b), but using one-to-many matching. That is, for each treated unit, choose $M = 2$ and $M = 10$ control matches, where M is the number of matches for each treated unit. Compare the results with the estimates obtained in (b). Do they differ? Why? What trade-offs are we making by increasing the number of matches?

(e) Now, instead of matching on covariates, we'll use propensity scores to estimate the ATT.

i) Using logit, regress `abd` on `C.ach, C.akw, C.ata, C.oro, C.paj`, and `age`. For each unit, generate a propensity score as the predicted probability of taking treatment. On one plot, overlay the density curves of the propensity scores for the treated and untreated units. What does this plot reveal?

ii) Using one-to-one matching on propensity scores, estimate the ATT, its standard error, and plot the densities of the propensity scores for the treated and untreated units. Assess balance using a balance table, as before.

iii) Now estimate the ATT using inverse-probability-weighting (IPW).

iv) (**Optional**) Use the bootstrap to estimate the SE of your IPW estimate, as well as a 95% bootstrap confidence interval. Be sure to account for uncertainty in the estimation of propensity scores as well. Do your results accord with your previous findings? Why or why not?

Optional Problems

Problem A

The *curse of dimensionality* makes it difficult to work in a situation where there are many pre-treatment covariates to condition on. Suppose we have covariates X_k for $k = 1, \dots, P$, where P is the number of pre-treatment covariates (i.e., the dimensionality of the covariate space). Then x_i is a vector of covariate values for observation i, $x_i = [x_{i1}, \dots, x_{iP}]^T$.

(a) Write an expression that gives the Euclidean distance between observations i and j in terms of their covariates x_i and x_j, respectively.

(b) Create a dataset X with 500 observations and 20 covariates, where each covariate should be independently drawn from $\mathcal{N}(0, 1)$. Now, consider a new observation with covariates $x^\star$ all equal to zero. Find the new observation's nearest neighbor in the dataset you created, using your expression for Euclidean distance and only the first covariate. That is, find the point i in the dataset whose first covariate x_{1i} is closest to $x_1^\star$. Record the Euclidean distance between x_{1i} and $x_1^\star$. Repeat this process, each time adding an additional covariate: next using two covariates, then three, and so on through all 20. Use your results to plot the Euclidean distance to the new observation's nearest neighbor as a function of dimensionality.

(c) What do your results demonstrate about matching?

(d) We'll now examine how the curse of dimensionality affects matching on discrete covariates. Create another dataset with 500 observations, in which a treatment D_i is allocated so that the first 250 units in the data are treated and the rest are control. Draw each of the 20 covariates independently from the multinomial distribution with 4 categories, with probabilities $p = [0.1, 0.2, 0.3, 0.4]$ for the treatment group and $p = [0.25, 0.25, 0.25, 0.25]$ for the control group. Now, match each treated unit to an untreated unit with exactly the same x_{i1}, with replacement. Note the proportion of treated units that can be matched, and repeat as in (b) with increasing dimensionality up to 20. Plot the proportion of unmatched treated units as a function of dimensionality. What do you conclude? How would your plot change with increasing units in your data?

(e) Now we'll examine the curse of dimensionality as it relates to distance metric-based matching. Using the same dataset you created in (d), match each treated unit to an untreated unit using nearest neighbor matching with the Mahalanobis distance (see Part II, Slide 3 in the Observational Studies lecture slides for the definition). Again start by finding matches and calculating the distance between them based on a single covariate, then two, then three, until you reach $P = 20$. Plot the mean Mahalanobis distance between nearest neighbor matches across all treated units as a function of P. How do your results differ from those in part (d)?

(Note: Please write your own code to calculate the Mahalanobis distance between observations — do not use the canned function `mahalanobis()`. Also notice that our pre-treatment covariates are categorical, meaning that we will need to first transform them into appropriate sets of dummy variables before calculating distances.)

Problem B

In this problem we revisit Blattman and Annan's data on child soldiers in Uganda (`child_soldiering.csv`), which we used in Problem 2.

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

$$Y_i = \hat{\alpha} + \hat{\beta} D_i + \hat{\gamma} \mathrm{X} + u_i$$

Where $\hat{\alpha}$ is an intercept, $\hat{\beta}$ is the OLS estimator for the ATE, and $\hat{\gamma}$ is a length-four vector of coefficients for X, which is an $N \times 4$ matrix of possible observed confounders: age, father's education, mother's education, and residency in Acholibur (`C.ach`); residuals appear as u_i. If we want to interpret the ATE estimate produced from this regression model as the causal effect of abduction on years of education what is our identification assumption? 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 the paper for more details), come up with a possible binary confounder that predicts both education and abduction. Describe your binary confounder, in words, and then draw a DAG to represent the relationships between abduction (D), education (Y) and your potential confounder (U).

(c) Consider your confounding variable, how plausible is the assumption that D_i and U_i do not interact, in other words that the average effect of U_i on Y_i is constant between treatment groups? For now, let's say this assumption is plausible and let's examine this binary confounder that predicts both education and abduction.

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 or imbalance in U) and the y-axis is γ (the effect of U_i on Y_i). 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, and even likely, that our unobserved confounder could create this amount of bias we need to compare it to something. Compare the hypothetical degree of confounding based on the values of our sensitivity parameters with the observed degree of confounding based on observed covariates.

Add the covariate residency in Acholibur (`C.ach`) to the plot as a benchmark. Assume that the covariates (`age`) and (`fthr.ed`) have a negative relationship (but the same magnitude) with our outcome, education. Add these observed covariates to the plot as well where the x-axis value is the change in propensity scores when moving from the first to third quartiles of these two variables.

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.