Exercise: Matching

Let's simulate some fake data and see whether we are able to recover the correct treatment effect using matching methods.

1. First, let's generate some confounder variables for 100 people.

(a) The variable 'age' should be drawn randomly from the normal distribution with mean 40 and standard deviation 7.

(b) The variable 'gender' should be drawn randomly from the binomial distribution with a 0.5 probability of being male or female.

(c) The variable 'income' should be drawn randomly from the normal distribution with mean 500 and standard deviation 50.

(d) The variable 'education' should be randomly drawn from one of four numerical categories with equal probability: 0 (None), 1 (Primary), 2 (Secondary), 3 (Tertiary).

2. Our outcome is going to be attitudes to redistribution. Use the expressions below to simulate potential outcomes, with a treatment effect of 5.

$$y_0 = N(20, 5) + \frac{age}{4} - 5 * gender + \frac{income}{50} - 3 * education$$

$$y_1 = y_0 + 5$$

3. Treatment D is receiving a government social program, but treatment is **not** randomly assigned in any way. Instead, treatment depends on age, gender, income and education. Imagine we know the treatment assignment mechanism so that binary (1/0) treatment is determined by the following expression:

$$D = \begin{cases} 1 \text{ if } 2 * gender + \frac{age}{8} + \frac{income}{50} + 2 * education + N(0, 3) > 19 \\ 0 \text{ else} \end{cases}$$

4. Calculate observed outcomes based on potential outcomes and treatment.

5. As always, as a benchmark, let's run the 'naive' regression of the outcome on the treatment with no controls. Why is the result different from our assumed treatment effect? Be specific.

6. Our first task is to try and do a 'manual' matching example - to try and 'match' one treated unit with one control unit so that the *only* thing that is different about them is their treatment status. Take the first treated unit in your dataset. What are its values of gender, age, income and education? Manually, by trial-and-error (not using any package or pre-prepared function), identify the most similar *control* unit. How different are your matched pair on these four variables?

7. Compare the outcome between your matched treated unit and control unit. Is this consistent with our assumed treatment effect? Why is it similar? Why is it different?

8. Matching repeats this process for multiple units and then finds the average difference in outcomes between the treated and control units. Use the *matchit* package to conduct 'nearest neighbour' (the default) matching method on your dataset for the two categorical confounder variables: gender and education. What is the result of the matching procedure - how many units were matched?

9. Use *match.data* to extract the matched dataset and calculate the average difference in means between the treated and control groups. How does the result compare to the naive regression in Q5?

10. To understand how matching changed our dataset, check the *summary* information about your matched data. On which variables did balance improve? Did balance deteriorate on any variables?

11. Matching *ONLY* makes a difference if we throw away some data - the data for which we cannot find good matches. The more data we throw away, the better matched/balanced is our remaining data. Conduct your nearest neighbour matching procedure again, but this time use the *exact* parameter to also require that matched treated and control units have exactly the same gender and education. How many units are matched now? Has balanced improved or deteriorated on any variables? What is the average difference in mean outcomes between treated and control groups?

12. An alternative way of limiting the number of matches is to specify a maximum distance measure beyond which paired units are dropped. Run your matching procedure again, specifying a *caliper* of 0.1 (or try other values if this doesn't work). How many units are matched now? Has balanced improved? What is the average difference in mean outcomes between treated and control groups?

13. One problem with this nearest neighbour matching procedure is that it is 'dumb', matching one pair, and then another, even if the distance between all paired units would be lower if the matches were switched around. Try using the 'optimal' and 'genetic' methods of *matchit* to improve your analysis. Has balanced improved? What is the average difference in mean outcomes between treated and control groups?

14. Try conducting matching with the Coarsened Exact Matching (`cem`) methodology. This turns continuous variables into categorical variables and then uses exact matching.

15. Finally, let's calculate the propensity score (the probability each unit was treated) and match on this new propensity score. First run a logit regression of treatment on your four confounding variables, save the fitted values from this regression, and match on these fitted values (the probability each unit was treated) using nearest-neighbour matching and a caliper of 0.1 of a standard deviation.

16. The risk of using matching is that we have so many options that we can keep trying until we find a 'big' effect. So we should always be guided by a clear, measurable goal: improving balance. Which of the matching methods you used above maximize balance on the four confounding variables?