Problem Set 2
MGMT 737

Spring 2025

You should have gotten this homework assignment from the Github classroom environment (https://classroom.github.com/classrooms/192971645-yale-mgmt-737-spring-2025-classroom). In submitting your problem set, you should have two files in your Github repository:

1. `homework2-code.R`, which contains your code,
2. `homework2-writeup.pdf`, which contains your writeup

You may have other files / folders if necessary when there are images or auxiliary files. My very strong preference is for you to write this up in R. If this is not possible, you can use Python. Please let me know if you're planning on this. (This is not a coding preference, but mainly a grading issue.)

1. Propensity Scores. This analysis will use the dataset from Problem 1 as well as the PSID dataset from Dehijia and Wahba, `lalonde_psid.csv`. These datasets have identical variables. The new dataset is a sample of observations from the Panel Survey of Income Dynamics that can be used as alternative control observations. Importantly, these observations were not in the initial randomization.

   (a) Using `age`, `education`, `hispanic`, `black`, `married`, `RE74` and `RE75`, construct a propensity score using the *treated* group in `lalonde_nsw.csv` and the control sample of `lalonde_psid.csv`. Use a logit regression model to do so (you may use a canned routine to run the regression). Report the average p-score for the treated and control samples, and plot the propensity score densities for the treatment and control groups.

   → Label the calculated value as `treated_pscore` and `control_pscore` in your code.

   (b) Using your p-score estimates, calculate the IPW and SIPW estimate for control and treated mean of the outcome, and the average treatment effect. Contrast these estimates to the control mean of the outcome from the NSW sample, and the treatment effect from last week's problem set.

   → Label the calculated value as `control_mean_ipw`, `treated_mean_ipw`, `ate_ipw`, `control_mean_sipw`, `treated_mean_sipw`,and `ate_sipw`.

   (c) Compare the ATE in the previous question to the treatment effect estimated using a linear regression using the PSID and NSW treatment sample, with `age`, `education`, `hispanic`, `black`, `married`, `RE74` and `RE75` as controls.

   → Label the calculated coefficient as `coef_lin_reg`

   (d) Now revisit your estimates from part a and b, and following Crump et al. (2009), discard all units with estimated propensities outside the range of $[0.1, 0.9]$. Reestimate the IPW and SIPW estimator of the ATE from part b using this trimmed sample.

   → Label the calculated values as `ate_ipw_trimmed`, and `ate_sipw_trimmed`.

   (e) Finally, calculate the IPW and SIPW estimates for the ATE using this trimmed sample for Black and non-Black individuals. Compare this estimate to the ATE for Black and non-Black individuals using the full randomized sample (last week's design).

   → Label the calculated values as `ate_ipw_black`, `ate_sipw_black`, `ate_ipw_nonblack`, and `ate_sipw_nonblack`.

   (f) We now consider what our results look like if we just consider difference-in-differences. Estimate a linear regression model (using canned software package like `lm` in R is fine – be sure to cluster on the individual) with two specifications:

      i. The difference in `RE78` less `RE75` on the left hand side against a treatment indicator.
      → Label the calculated coefficient as `coef_lin_reg_diff`.

ii. `RE78` on the left hand side against a treatment indicator and `RE75`.
→ Label the calculated coefficient as `coef_lin_reg_lagged`.

Report the estimates and compare to your previous estimates. Can you think of other specifications that might work?