Problem Set 9: Machine Learning MGMT 737

1. **Chernozhukov et al. (2022)** This problem will implement the heterogeneity analysis from Chernozhukov et al. (2022) using their publicly available pacakge. I will walk you through an example from an RCT in Atkin, Khandewal and Osman (2017) using the method.

   (a) We will be replicating first Column 1 from Table 4 (the ITT). The dataset is called `analysis.dta`. You can download it here. First, load the file `jpal_analysis.dta`. There are 191 observations. First, replicate the main specfciation, which is a regression of `ever_export` on `treatment` and `ever_export_b` and includes strata fixed effects (`strata`). You should replicate the column 1 coefficient from Table 4.

   (b) Now for simplicity, we're going to omit the strata FE and the baseline covariates. Re-estimate the coefficient. How much does it change? Why do you think that is?

   (c) Now we will use a set of baseline covariates `ever_export_b`, `log_weight_sidda_b`, `qual_corners_b`, `qual_waviness_b`, `qual_weight_b,` `qual_touch_b`, `qual_sidda_packed_b`. There are a lot of other covariates but due to the design, they're not always osberved for all firms, and we do not want to induce additional selection issues. These are available for 190 of the 191 firms. Interact these covariates with the treatment variable and re-estimate the model. Do you find evidence of heterogeneity in the treatment? Would you trust these estimates? Why or why not?

   (d) Now, we will use the `GenericML` package from Chernozhukov et al. (2022) to estimate the treatment effect heterogeneity. You can install the package by running `install.packages("GenericML")`. Discussion of the package is available here. This is a bit of a complex package, and so your first task is to read through the Example on the documentation page.
   You can now do this problem set in two ways: First, you can figure out how to implement it yourself. See the last question for the goal; Second, I can walk you through it. If you want help, see the following problems. IMPORTANT NOTE: the "lasso" learner breaks the code in this example, so you should not use it in the learner set – I used the other defaults proposed in the README example.

   (e) Next, you should look at the `Homework9_problem1_solution_skeleton.R` I have provided. I want you to check the following aspects and how I have changed them from the example: (1) the learners (2) the Z covariates (3) the propensity score model (4) the treatment effect model.

   (f) Try running this code. You should find a 90% confidence interval for beta_2 of (0, 1.355).

   (g) Report the BLP estimates for beta_2 and the GATES estimates for gamma1, gamma4, and gamma.4 - gamma.1. How do you interepret these? Finally, calculate the CLAN for the `ever_export_b` covariate. Did firms that had previously exported have higher or lower treatment effects, on average?

2. **Lee Bounds** We now consider the simple application of Lee Bounds in the the NSW application from the first homework. Recall that the dataset is `lalonde_nsw.csv`. The outcome variable is `re78`(real earnings in 1978). The treatment indicator is `treat`. The remaining variables are potential covariates. Crucially, we are now going to study *wage receipts*, which is `re78` for individuals who have positive income. So, define a binary indicator `employed` which is 1 if `re78` is positive and 0 otherwise, and define `wage` as `re78` if `employed` is 1 and missing otherwise.

   For purposes of context, recall from Lalonde (1986) that the NSW training program guaranteed a job for 9-18 months, and that this randomization occurred over a multi-year period (1976-77) – this is discussed in Dehijia and Wahba's JASA article. We have far fewer data points than Lee (2009)'s application but we're going to do our best!

   (a) Estimate the effect of the treatment on the wage receipt, `wage`, and compare it to the effect on income (`re78`). Is it higher or lower? Why do you think that is?

(b) Estimate the effect of the treatment on the probability of employment. Is this significantly different from zero? (You may use statistical packages)

(c) Given the NSW program guaranteed a job, would you expect that the treatment ffect on wages is biased upwards or downwards? Why?

(d) Now, calculate p_0 from Lee (2009), which is

$$p_0 = \frac{Pr(\texttt{employed} = 1|\texttt{treat} = 1) - Pr(\texttt{employed} = 1|\texttt{treat} = 0)}{Pr(\texttt{employed} = 1|\texttt{treat} = 1)} \tag{1}$$

(e) Now, calculate and report the bounds on the average treatment effect on wages using the formula from Lee (2009).

(f) Do the same exercise, but use log(`wage`) as the outcome variable. (this is what Lee (2009) does in the paper). (N.B. since we're doing wages we can take logs without issues!) How do these bounds compare to the main results on wages from Lee (2009)'s Table 5?

(g) Now we want to use `nodegree` to sharpen our bounds. To do so, we'll reestimate Equation (1) for each covariate. This will give us p_{x1} and p_{x0}. Now, you find the percentiles for the outcome conditional on treatment, employment *and* `nodegree` (i.e. $F_{Y|T=1,S=1,X=1}$ and $F_{Y|T=1,D=1,X=1}$), and use these to construct bounds for each covariate. Report these bounds for *log wages.* (See Proposition 1.b from the paper for more details)

(h) Finally, you can aggregate these two bounds up using the density of the covariates for the employed untreated group to get a final bound. Compare to your results above.