Problemset (2),
Foundations of Machine learning,
HT 2023

Maximilian Kasy

In this problem, you are asked to implement some simulations and estimators in R. Your code should run from start to end in one execution, producing all the output. Output and discussion of findings should be integrated in a report generated in R-Markdown (or Quarto). Figures and tables should be clearly labeled and interpretable. The findings should be discussed in the context of the theoretical results that we derived in class.

We will implement calibrated simulations to evaluate the double/debiased estimator of the average treatment effect discussed in **?**.

For this question, we will use data from three files: `nsw_treated.csv`, `nsw_control.csv`, and `nsw_psid.csv`.[1] The National Supported Work (NSW) Demonstration was a subsidized work program. The datasets `nsw_treated.csv` and `nsw_control.csv` contain an experimental sample from a randomized evaluation of the NSW program: `nsw_treated.csv` contains the experimental treatment group and `nsw_control.csv` contains the experimental control group. The dataset `nsw_psid.csv` contains a non-experimental control group, which was obtained from the Population Survey of Income Dynamics (PSID). In all of the three datasets, the variables are defined as follows:

nsw	=1 for NSW participants, =0 otherwise
age	age in years
educ	years of education
black	=1 if African American, =0 otherwise
hispanic	=1 if Hispanic, =0 otherwise
married	=1 if married, =0 otherwise
re74	real (inflation adjusted) earnings for 1974
re75	real (inflation adjusted) earnings for 1975
re78	real (inflation adjusted) earnings for 1978
u74	=1 if unemployed in 1974, =0 otherwise
u75	=1 if unemployed in 1975, =0 otherwise
u78	=1 if unemployed in 1978, =0 otherwise

[1] You can download these files from `https://maxkasy.github.io/home/files/teaching/ML_Oxford_2023/nsw_treated.csv`, `https://maxkasy.github.io/home/files/teaching/ML_Oxford_2023/nsw_control.csv`, and `https://maxkasy.github.io/home/files/teaching/ML_Oxford_2023/nsw_psid.csv`.

1. As a first step, we set up the calibrated simulation of data.

   (a) Bind the three data-sets together, and fit random forest models to the combined data-set, separately for the treated units and the control units, for the prediction of $Y_i =$ *u78*.[2] Discard the non-experimental sample for the remainder of this problem.

   (b) For each observation in the experimental sample impute counterfactual values $\hat{Y}_i^1$ and $\hat{Y}_i^0$, corresponding to setting D_i to 1 or 0. We will hold these imputed values constant for the rest of the exercise. Calculate the average α of $\hat{Y}_i^1 - \hat{Y}_i^0$. We will take this average as our "true" average treatment effect for our subsequent evaluations of bias. Impute a predicted value $\hat{Y}_i$ to each observation in the experimental sample.

   (c) Write a function which takes no arguments and returns a vector of simulated outcomes for the data, where for each observation, Y_i is drawn independently from the $Ber(\hat{Y}_i)$ distribution.

2. Next, we will implement 6 types of estimators for the average treatment effect. These estimators are

   (a) the regression (or "naive plugin") estimator,

$$\hat{\alpha} = \tfrac{1}{n} \sum_i [\hat{m}(X_i, 1) - \hat{m}(X_i, 0)],$$

   where $\hat{m}(x, d)$ is an estimator of $m(x, d) = E[Y|X = x, D = d]$,

   (b) the inverse probability weighting estimator,

$$\hat{\alpha} = \tfrac{1}{n} \sum_i \left[\left(\frac{D_i}{\hat{p}(X_i)} - \frac{1 - D_i}{1 - \hat{p}(X_i)} \right) \cdot Y \right],$$

   where $\hat{p}(x)$ is an estimator of the propensity score $p(x) = E[D|X = x]$, and

   (c) the double-robust estimator, using the orthogonal score discussed in class,

$$\begin{aligned} \hat{\alpha} = \tfrac{1}{n} \sum_i \Bigg(& [\hat{m}(X_i, 1) - \hat{m}(X_i, 0)] \\ & + \left(\frac{D_i}{\hat{p}(X_i)} - \frac{1 - D_i}{1 - \hat{p}(X_i)} \right) \cdot Y \\ & - \left(\frac{D_i}{\hat{p}(X_i)} - \frac{1 - D_i}{1 - \hat{p}(X_i)} \right) \cdot \hat{m}(X_i, D_i) \Bigg). \end{aligned}$$

   You might wish to truncate $\hat{p}(X_i)$ so that it is bounded away from 0 and 1.

   Each of these can be implemented using (A) the full data, or (B) the sample splitting and averaging approach we discussed in class.

   Lastly, each of these can be implemented using different supervised learning methods, to estimate outcome regressions and propensity scores. For this homework, we will just

[2] You might for instance use the *ranger* package in R to do so.

consider linear OLS, Logit, and Random Forest regressions, but you might wish to play around with other estimators as well. Write a function which takes as its argument a data-set, and an option specifying a supervised learning method, and returns estimates for each of the six types of estimators.

3. Now we will set up a simulation combining the calibrated data and these estimators. In particular, write a function that takes as its argument the number of replications R, as well as a supervised learning method, and returns, for each of the 6 estimators, a list of R estimates.

   To do so, loop over replications (using parallel computing, e.g. the *future* package), and for each iteration simulate a draw of outcomes using the function written in step (a), which then serves as input for the function written in step (b).

4. For each estimation method and each supervised learning method considered, calculate the mean, the median, the variance, and the mean squared error, across replications.

   Produce figures and tables showing the distribution of each estimator, relative to the "true" effect α, as well as the mean bias, median bias, standard error, and root mean squared error, for each estimation method. Discuss you findings.

References

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

Wager, S. and Xu, K. (2021). Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855*.