# Problemset (2),
# Foundations of Machine learning,
# HT 2022

Maximilian Kasy

In this problemset you are asked to implement some simulations and estimators in R. Please make sure that your solutions satisfy the following conditions:

- The code has to run from start to end on the grader's machine, producing all the output.
- Output and discussion of findings have to be integrated in a report generated in R-Markdown.
- Figures and tables have to be clearly labeled and interpretable.
- The findings need to be discussed in the context of the theoretical results that we derived in class.

1. In this problem, you are asked to implement and compare several methods for supervised learning on some real world data.
   (a) Go to `https://archive-beta.ics.uci.edu/` and download the "Adult" dataset, as well as another dataset of your choice. For the following, drop the weighting variable.
   (b) Split the data into a training sample (80% of observations) and a hold-out sample (20% of observations). Set the latter aside for later use.
   (c) Implement different ways to predict the outcome, including
      i. Unregularized OLS or logit regression.
      ii. The same regression, but with ridge and lasso penalties.

iii. Random forest.

iv. Neural net.

For each of these, make sure to appropriately tune hyper-parameters using cross-validation.

There are many implementations of these methods; you might find some guidance in `https://bradleyboehmke.github.io/HOML/`.

(d) Evaluate each of your predictive models using the hold-out data you set aside initially. Discuss their relative performance, for both of your datasets.

2. In this problem, we will implement calibrated simulations to evaluate the double/debiased estimator of the average treatment effect discussed in Chernozhukov et al. (2018).

(a) As a first step, we set up the calibrated simulation of data. To do so, take the "Adult" data-set from problem 1, generate a dummy $D_i$ for "some college or more," and then drop the education variables as well as the weighting variable.

Fit a random forest model for the prediction of adult income above 50k to these data, and impute a predicted value $\hat{Y}_i \in [0, 1]$ (i.e., a probability) to each observation. We will hold these predicted values constant for the rest of the exercise.

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.

Lastly, for each observation also impute counterfactual values $\hat{Y}_i^1$ and $\hat{Y}_i^0$, corresponding to setting $D_i$ to 1 or 0, and calculate the average 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.

(b) Next, we will implement 6 types of estimators for the average treatment effect. These estimators are (i) the regression (or "naive plugin") estimator, (ii) the inverse probability weighting estimator, and (iii) the double-robust estimator, using the orthogonal score discussed in class. 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 (as in problem 1), to estimate outcome regressions and propensity scores. 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.

(c) 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, the mean as well as the variance across replications.

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). For each estimation method and each supervised learning method considered, report the squared bias as well as the variance across replications.

(d) Assemble your results in one big table, and discuss your findings.

## References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.