# Lab 6 - Machine Learning for Social Science

**To be handed in no later than October 15th, 10:00.** The submission should include code, relevant output, as well as answers to questions. We recommend the use of RMarkdown to create the report.

## Part 1: Meta-learners for job training evaluation

The dataset `"job_training_updated.csv"` contains information about 12,000 individuals who either participated or did not participate in a job training program, including:

- `training`: Binary indicator of whether individual participated in training (treatment)
- `earnings`: Post-training annual earnings in thousands of dollars (outcome)
- `age`: Age of individual
- `education`: Years of education
- `prior_earnings`: Earnings before training program
- `employment_history`: Years of prior employment
- `urban`: Binary indicator of urban residence

1. Fit regular OLS regression using `lm()`, including all non-treatment and non-outcome variables as control variables. Interpret the coefficient for the treatment variable as the average treatment effect. Considering what we talked about in the lecture, what properties of the data would lead you to believe your estimate is biased? Motivate.

2. Next, you shall estimate an orthogonal learner, using decision trees as the method for predicting both the treatment and the outcome. Please follow the following steps:

   a. Train a decision tree model using `rpart()` to predict `training` from all confounders using the full dataset. For classification trees, use `method="class"` and for the control parameters use: `cp=0, minbucket=5, maxdepth=30` (i.e., `control=rpart::rpart.control(cp=0,minbucket=5, maxdepth=30)`).

   b. Train a decision tree model using `rpart()` to predict `earnings` from all confounders using the full dataset. For regression trees, use `method="anova"` and the same control parameters: `cp=0, minbucket=5, maxdepth=30` (i.e., `control=rpart::rpart.control(cp=0,minbucket=5, maxdepth=30)`).

   c. Make predictions of treatment (using model from a, with `type="prob"`) and outcome (using model from b) for all observations.

   d. Calculate residuals for all observations: `X_tilde = X - X_hat`, `Y_tilde = Y - Y_hat`.

   e. Estimate the ATE by regressing `Y_tilde` on `X_tilde` using `lm()`.

   f. Report the ATE. How does it compare to your OLS estimate in #1?

   g. Which of the two methods do you trust more? Can you think of any aspect of the implementation of the orthogonal learner which could bias its estimate?

3. Given your conclusions in #2, do you think either of the following two changes to the setup of the orthogonal learner could improve the ATE estimate? (i) switching from a decision tree to a random forest, (ii) add cross-fitting. Motivate.

4. Now you shall implement the two updates discussed in #3. Please do the following:

   a. Divide your data into 5 folds (hint: you can use `createFolds()` from the `caret` package).
   b. Create a for-loop which in each iteration $i$ does the following:
      i. Train a random forest model using `randomForest()` (with `ntree=200` and `mtry=2`) predicting `training` from confounders on data in folds $\neq i$.
      ii. Train a random forest model using `randomForest()` (with `ntree=200` and `mtry=2`) predicting `earnings` from confounders on data in folds $\neq i$.
      iii. Use models from (i) and (ii) to predict treatment (with `type="prob"`) and outcome for observations in fold $i$.
      iv. Calculate residuals `X_tilde` and `Y_tilde` for observations in fold $i$.
      v. Store residuals from fold $i$.
   c. Combine dataset of residuals and regress `Y_tilde` on `X_tilde` using `lm()`.
   d. Report the estimated ATE. Do you trust this estimate more than those in #2, and if so why (or why not)?

5. Suppose we learn that the true average treatment effect is 5.5 thousand dollars. Report which method came closest, and discuss what this says about the properties of the data—in particular the relation between the confounders and the treatment and outcome.

---

## Part 2: Heterogeneity I

The dataset `"scholarship.csv"` contains information about 15,000 students who either received or did not receive a college scholarship, including:

- `scholarship`: Binary indicator of scholarship receipt (treatment)
- `completed`: Binary indicator of degree completion within 6 years (outcome)
- `gpa`: High school GPA (scale 0-4)
- `parental_income`: Parental income in thousands of dollars
- `first_generation`: Binary indicator of first-generation college student status
- `sat_score`: SAT score (scale 400-1600)
- `distance_to_college`: Distance from home to college in miles
- `financial_need`: Measure of financial need (scale 0-100)

1. Suppose your co-author, who has done a careful literature review, has found support for two of the variables, `first_generation` and `financial_need`, having a moderating effect. What you shall do first is examine whether you find evidence of this in your data. Implement a standard linear regression using `lm()` (or `glm()` if you prefer logistic regression) with the treatment variable as well as all other input variables (presumed confounders) included, with `first_generation` and `financial_need` interacted with the treatment variable. Report your findings: do you find evidence supporting your colleague's conclusion from the literature?

2. Considering what we discussed in the lecture, what is one limitation of this standard approach to effect heterogeneity? What are properties of the data (or state of the field) that could make this limitation more or less problematic?

3. Next, you shall consider an alternative approach to effect heterogeneity, using causal trees. At a high level, describe what is the key difference in assumption we make when using causal trees compared to the traditional approach?

4. Perform a causal tree analysis by doing the following:

    a. Estimate the causal tree using the function `causalTree()` from the `htetree` package, specifying the formula as in #1 (except drop the interactions and leave out the treatment variable; the latter is specified separately). Use the following parameters: `split.Rule="CT"`, `cv.option="CT"`, `split.Honest=TRUE`, `split.Bucket=TRUE`, `minsize=60`, `cp=0`, `bucketNum=40`.
    b. Visualize the tree using `rpart.plot()` and describe the combination of splits which identify the population with (a) the largest treatment effect and (b) the smallest.
    c. Suppose our dataset is a standard observational dataset common to the social sciences, e.g., a survey dataset of a random sample of the population. Given this information, what could be a potential threat to the validity of our causal tree results?

5. Given potential concerns of selection bias, you shall next examine the potential imbalance of treated and untreated observations inside different leaves. To do so, please follow these steps (Hint: various code-chunks are provided that may be helpful):

    a. Estimate a propensity score model—a standard logistic regression model using `glm()` with `family=binomial()`—predicting the treatment variable based on the confounders. Specify `type="response"` in the `predict()` function.

```
# Tips: add the predicted propensity score to your main dataset
dt$ps_hat <- as.numeric(predict(ps_model, type = "response"))
```

b. Calculate the mean and standard deviation of the propensity scores within each leaf and treatment group combination. (Hint: use `$where` to extract leaf assignments)

```
# Tips: add the leaf identified to your main dataset
dt$leaf <- factor(causal_tree_object$where)
```

```
# Assuming your dt contains:
#    - scholarship,
#    - propensity score,
#    - leaf id,
# ...You can calculate means, sds as follows:
leaf_group_long <- dt[, .(
  n       = .N,
  mean_ps = mean(ps_hat),
  sd_ps   = sd(ps_hat)
), by = .(leaf, scholarship)]
# > step 2
leaf_group_stats <- data.table::dcast(
  leaf_group_long,
  leaf ~ scholarship,
  value.var = c("n", "mean_ps", "sd_ps"),
  fill = NA_real_
)
```

c. Based on the mean and standard deviation, calculate the standardized difference in means measure within each leaf. What do these indicate about your results in #4?

```r
# Assuming that you have an R object "leaf_group_stats"
# that looks like one above, you can calculate SMD as follows:
leaf_balance <- copy(leaf_group_stats)[
  , `:=`(
      SMD_ps = {
        denom <- sqrt((sd_ps_1^2 + sd_ps_0^2)/2)
        ifelse(is.finite(denom) & denom > 0,
               abs(mean_ps_1 - mean_ps_0) / denom, NA_real_)
      }
    )
][order(-SMD_ps)]
```

6. Given your findings in the previous task, you shall next do a causal tree analysis wherein you incorporate inverse probability weighting. To do so, please do the following:

   a. Refit the causal tree using `causalTree()` with same specifications as in #4, and add the `weights` argument set to `1/p` for treated units and `1/(1-p)` for control units, where `p` is the predicted propensity score (see code chunk below for how you could do this). This incorporates IPW into the tree.

```r
# This assumes you have stored the propensity scores in dt
# - includes pmax() to cap very high/small values
dt$w_ipw <- ifelse(test = dt$scholarship==1,
                   yes = 1/pmax(dt$ps_hat, 0.02),
                   no = 1/pmax(1-dt$ps_hat, 0.02))
```

   b. Assess the balance for this tree in the same way you did in #5 (but you can skip the first step which estimates the propensity score model). Did the balance improve in comparison to #4?

   c. Visualize the tree and provide an interpretation of its structure, highlighting what you think is interesting in it. Are the conclusions you draw from this tree different from those in #4? What does this suggest about the findings in #4?

   d. To get a sense of the subgroups contained within each leaf (of interest), please describe its average properties in terms of all the input variables (except treatment).

   e. How do these results map onto your findings in the first analysis in #1? Do you find that the variables suggested are most important indeed are so? What would you say to your co-author?

7. (BONUS) Lastly, to get a sense of the sensitivity of the tree structure to the particular data we estimate it on, please estimate (and visualize) two different trees with different `seed.numbers` assigned. Do they depart meaningfully from each other? Speculate about why you think this difference (or lack thereof) exists.

---

## Part 3: Heterogeneity II

In this part, you will continue working with the scholarship dataset from Part 2.

1. In this last part, you shall continue with the exploration of heterogeneous treatment effects. But instead of standard OLS and causal trees, you shall use causal forests. Before doing so, please answer the question: why is it not necessary to include inverse probability weighting in causal forest, like we did for trees?

2. Now, run a causal forest analysis following these steps:

   a. Estimate the causal forest using the function `causal_forest()` from the `grf` package, specifying the input arguments as follows: X (matrix of covariates), Y (outcome), W (treatment), `num.trees=2000`, and `honesty=TRUE`. Set a seed for reproducibility. Report the average treatment effect using `average_treatment_effect()`.

   b. Examine which variables were most important to account for the heterogeneity in treatment effect by using the function `variable_importance()`. Make a bar chart and interpret. Does this result line up with your findings using causal tree—do the most important variables here overlap with those showing up in the best causal tree?

   c. For the two variables you identified as most important, please examine how the effects vary along these dimensions. To do so, divide into quintiles of these variables (if continuous) and calculate average treatment effects for each subcategory separately. Plot how the treatment effect varies across quintiles and interpret. Does this result provide additional information to what you could infer from the causal tree?