# Causal Machine Learning: Homework 1

***Due date*: September 22, 11:59PM**

Your name:

## Bias-Variance Tradeoff

1. In your own words, explain the bias–variance tradeoff when using a *sample* to approximate the *population* Conditional Expectation Function (CEF), and clarify where the variance in prediction comes from.

2. In your own words, explain how *Random Forests* address the bias–variance tradeoff, and why balancing this tradeoff is important.

3. In your own words, explain how *Boosted Trees* differ from *Bagging* and *Random Forests* in how they address the bias–variance tradeoff.

## Frisch–Waugh–Lovell Theorem

In lecture, we showed that the multivariate OLS regression of $Y_i$ against the treatment variable $D_i$ and control covariates $W_i$

$$Y_i = \beta_1 D_i + \beta_2' W_i + \epsilon_i$$

can be recovered in a bivariate regression,

$$\tilde{Y}_i = \beta_1 \tilde{D}_i + \mu_i$$

$$\beta_1 = \frac{Cov(\tilde{Y}_i, \tilde{D}_i)}{Var(\tilde{D}_i)}$$

where $\tilde{D}_i$ and $\tilde{Y}_i$ are defined as residuals from their respective regressions against $W_i$.

$$\tilde{D}_i = D_i - \gamma_{DW}' W_i, \quad \tilde{Y}_i = Y_i - \gamma_{YW}' W_i$$

$$\gamma_{DW} = \arg\min_{\gamma} E\left[(D_i - \gamma' W_i)^2\right], \quad \gamma_{YW} = \arg\min_{\gamma} E\left[(Y_i - \gamma' W_i)^2\right]$$

1. Show that $\beta_1$ can be similarly recovered by using $Y_i$ rather than residualized $\tilde{Y}_i$

$$\beta_1 = \frac{Cov(Y_i, \tilde{D}_i)}{Var(\tilde{D}_i)}$$

2. Symmetrically, can one use unresidualized $D_i$ in the numerator to express $\beta_1$?

$$\beta_1 \stackrel{?}{=} \frac{Cov(\tilde{Y}_i, D_i)}{Var(\tilde{D}_i)}$$

3. In a high-dimensional setting where $p/n \not\to 0$, the OLS estimator of $\beta_1$ is no longer well-behaved, and its sampling variability can be severely inflated. We introduced Double LASSO in lecture to address this high-dimensional problem.

(a) Explain in words the relationship between the *penalty level* $\lambda$ in Double LASSO and the theoretical variance of your estimated $\beta_1$.

(b) After running Double LASSO, we obtain residualized $D_i$ and $Y_i$ by partialling out high-dimensional controls. In the final bivariate regression of residualized $Y_i$ on residualized $D_i$, can we rely on the standard sandwich formula for the standard error of $\hat{\beta}_1$? Why or why not?

## Test of Convergence Hypothesis

In lecture, we provided an empirical example of partialling-out with Double LASSO to estimate the regression coefficient $\beta_1$ in the high-dimensional linear regression model. Specifically, we estimated how the rates at which economies of different countries grow ($Y_i$) are related to the initial wealth levels in each country ($D_i$) controlling for country's institutional, educational, and other similar characteristics ($W$). The relationship is captured by $\beta_1$, the *speed of convergence/divergence*, which measures the speed at which poor countries catch up ($\beta_1 < 0$) or fall behind ($\beta_1 > 0$) rich countries, after controlling for $W_i$.

Indeed, residualized $D_i$ and $Y_i$ can also be estimated using Machine Learning (ML) methods beyond LASSO, as demonstrated in lecture. In this question, you will estimate residualized $D_i$ and $Y_i$ using the ML methods introduced in class. For this exercise, we will use the dataset `GrowthData`, which is included in the `hdm` package.

```r
library(hdm)
## function to get data
getdata <- function(...) {
  e <- new.env()
  name <- data(..., envir = e)[1]
  e[[name]]
}

## now load your data calling getdata()
growth <- getdata(GrowthData)
```

```r
## create the outcome variable y, treatment d, and control covariates W
y <- growth$Outcome
W <- growth[-which(colnames(growth) %in% c("Outcome", "intercept", "gdpsh465"))]
D <- growth$gdpsh465
```

1. Using the package `randomForest` and its default settings of `ntree` and `nodesize`, report the estimated $\beta_1$. Use `set.seed(1)`. What can you conclude from your estimation?

2. Now change the setting and let `ntree = 10` and `nodesize = 1`. Use `set.seed(1)`. Re-estimate $\beta_1$ and report your findings. In your own words, explain the possible reasons of changes in your findings.

```r
set.seed(1)
library(randomForest)
```

3. We may also use a Neural Network model to estimate the residualized $D_i$ and $Y_i$. Set the random seed with `set.seed(1)`. Configure the hyperparameters as follows:

- Build a neural network with one hidden layer containing 200 neurons.
- Train the network using Stochastic Gradient Descent (SGD) with a `learning rate` of 0.05, 200 `epochs`, and a `batch size` of 10.
- Implement early stopping based on performance on a 20% validation set, with a patience of 10 `epochs`.

Based on your estimation results $\beta_1$ under this model, what can you conclude about the relationship between country's initial wealth and growth rate?

```r
set.seed(1)
library(keras3)
```