Fall 2019, BIOS 740: Computing Assignment 1 (C1)

In this assignment, you will use the AIDS Clinical Trials Group Study 175 data (`ACTG175`) located in R package `speff2trial`. This dataset is a randomized clinical trial for HIV type 1 infected patients with 4 treatment arms. Arm 0 was Zidovudine (ZDV) monotherapy 600mg daily. Arm 1 was combination therapy of ZDV 600mg daily plus Didanosine (DDI) 400mg daily. Arm 2 was combination therapy of ZDV 600mg daily plus Zalcitabine (also known as Dideoxycytidine or DDC) 2.25mg daily. Arm 3 was DDI monotherapy 400mg daily. It has been discovered that the combination therapies (arms 1 and 2) are often more effective than monotherapies (arms 0 and 3) but it is not clear which combination therapy is better. Thus for this assignment, we want to compare the two treatments (arms 1 and 2) and exclude the other two treatments (arms 0 and 3) from the data.

The outcome of interest is `cd496`, which is CD4 cell count (cells/mm^3) at 96 weeks, where larger counts means better immunological status. We limit our analysis to those who have their CD4 count recorded at 96 weeks (i.e., remove observations with missing `cd496`). For simplicity, we will not use all of the X-variables in the original data so please exclude the following X-variables from the feature space: `oprior, z30, zprior, preanti, str2, treat, offtrt, cd420, r, cd820, cens, days`. For more information about the `ACTG175` dataset, please refer to the documentation of `speff2trial` (hyper-linked). This dataset has no missing data in the covariates so you don't need to perform any missing data analysis, but keep in mind this would not usually be the case for real world applications.

The goal of this assignment is to find the optimal treatment plan for individuals under the single-decision time setting. You will need to estimate the optimal individualized treatment regime $\hat{d}^{opt} \in \mathcal{D}$ and estimate its corresponding value function

$$\hat{V}(\hat{d}^{opt}) = \frac{\sum_{i=1}^{nte} Y_i 1\{A_i = \hat{d}^{opt}(X_i)\}/P(A_i|X_i)}{\sum_{i=1}^{nte} 1\{A_i = \hat{d}^{opt}(X_i)\}/P(A_i|X_i)}$$

where nte stands for the size of test set (see k-fold cross validation below for more information about training and test splits). Consider the following three precision medicine methods:

1. Random forests (Breiman 2001)
2. Doubly robust augmented inverse probability weighted estimator (AIPWE, Zhang et al 2012)
3. Residual weighted learning with linear kernel (RWL, Zhou et al 2017)

Here are the guidelines for the report:

- Perform any data preprocessing you find necessary.
- Apply all three approaches separately to the HIV data and justify any specific model configuration. Use k-fold cross-validation (CV) once to create training and testing sets. It is not required to perform repeated k-fold CV for this assignment but it is often recommended for real world research.
- Present results in terms of estimated value functions and their standard errors (SE) in one table. You will have one estimated value and one SE for each model. For each model, the estimated value function is the mean of $\hat{V}$'s across all k test folds and the SE is the standard deviation of the $\hat{V}$'s across all test folds.

- Compare results of each model and elaborate your observations and interesting findings.
- Briefly explain the pros (at least 2) and cons (at least 2) of each of the three aforementioned approaches. Identify whether the approach belongs to the regression-based type, the classification-based type, or a hybrid of the two.

This assignment should be submitted as a well-written technical report with sufficient explanation in the same style as Methods and Results sections in peer-reviewed journals. It should be programmed in R using the following packages: `randomForest` (hyper-linked) for RF and `DynTxRegime` (hyper-linked) for AIPWE and RWL. More specifically, AIPWE will need the function `optimalSeq()` and RWL will need `rwl()`. To help you finish this assignment, we recommend that you look up their documentations and tutorials as well as the original paper of the models listed above. You will find this tutorial of `DynTxRegime` here (hyper-linked) very useful.

This assignment is **due before class on September 15**. Please turn in report together with your code (as an appendix) as one stapled hard copy. Your code should have an appropriate amount of comments in between. The report needs to be typed up and no longer than 5 pages (including results but not including code) and the code part should be no longer than 3 pages. PDF files generated from RMarkdown or LaTeX are preferred. Email submission is only allowed if notified and approved by the course instructor in advance. The quality of the report is judged by (1) completion, (2) statistical correctness, (3) code presentation, and (4) explanation and report presentation.

References

1. Breiman L (2001), Random Forests, Machine Learning 45(1), 5-32.
2. Zhang B, Tsiatis AA, Laber EB, and Davidian M (2012). A robust method for estimating optimal treatment regimes. Biometrics 68:1010–1018.
3. Zhou X, Mayer-Hamblett N, Khan U, and Kosorok MR (2017). Residual weighted learning for estimating individualized treatment rules. Journal of the American Statistical Association 112:169–187.

Hints

- For step 1 "data preprocessing", you can consider tricks like data transformation (e.g. log or standardization) and/or feature selection.
- Different models might require the treatment variable to be coded differently. Sometimes A needs to be 0/1 but sometimes it needs to be ± 1. Some functions automatically convert 0/1's to ± 1's.
- To help you get started on coding:

```
library(tidyverse)
library(speff2trial)
data(ACTG175)
dat = ACTG175 %>%
filter(!is.na(cd496), arms %in% 1:2) %>%
```

```
select(-pidnum, -str2, -offtrt, -cd420, -r, -cd820, -cens, -days,
 -arms, -treat, -oprior, -z30, -zprior, -preanti)
```

- Random forests: For each training set, the training should be stratified by their actual treatment separately but you need to predict on the entire test set. After repeating this for all k folds, each subject will have two counterfactual predictions (under two treatments). The optimal treatment for each subject would be the treatment that has higher predicted potential outcome.

- AIPWE: you will need a regime function (indexed by η) that defines the class of treatment regimes. Let's follow the Zhang et al 2012 paper and restrict our options to a linear class. Please refer to the paper for more information. The domain/bound for η can be specified to be -1 and 1.

- The RWL model has built in functions to look for optimal treatments so specify `cvfolds` to be something *not* `0L`. Note that this CV is different from the k-fold CV in the guidelines as this CV is nested within the model training to tune hyperparameters whereas the guideline CV is the outer validation to assess $\hat{V}$ and generate standard errors for $\hat{V}$.

- In general, you will need to specify some model choices (e.g., `moMain`, `moCont`, and `moPropen`) so please make sure you explain why you choose to do so. This assignment is fairly open but we need to know your justification in detail.