TAKE-HOME FINAL EXAM (APSTA-GE.2012) assigned: 12/7/18

Professor Jennifer Hill due: 12/18/18

A **hard copy** of the exam must be submitted by **noon** on **Tuesday, December 18th 2018.** In addition to an online submission, a hard copy must be submitted, to **a box** on the shelves right next to my office (2nd floor of Kimball Hall, **Room 210, 246 Greene St**.).

This exam has a slightly complicated structure. *Please attend closely to the details*. Also please **do all work independently.** You are welcome to email me questions regarding points of confusion but you may not solicit answers from me or from others via Piazza or other means.

Each of the questions below has a very similar structure. The goal in each case is to explore the implications of violating assumptions required by a given causal approach. In each case you are asked to generate multiple datasets. Then you will be asked to implement (a version of) the approach and comment on the bias of the estimator. Given differences in the difficulty level of each simulation exercise and implementation, the write up for each method is worth a different number of points. You need to answer a subset of questions totalling either **10 or 11 points**.

Requirements for each method (recall you will only answer a subset of the methods on this list):

ALL DATASETS SHOULD HAVE 1000 OBSERVATIONS

1) **Propensity scores** (6 points). All worlds should have 5 covariates, one binary treatment variable, and potential outcomes for one variable. The covariates should have a dependence structure. The response surface for World A can be linear and should satisfy all other assumptions as well. At least one of the response surfaces (that is, E[Y(0) | X] or E[Y(1) | X] in World B should be non-linear (with R^2 less than .75) but the other assumptions should hold. World C can use the same response surface as World B but should violate one of the other assumptions. You should estimate a causal effect in each world both with linear regression and either propensity score matching or IPTW. You must present and discuss balance diagnostics and overlap plots for each of the three worlds.

2) **Instrumental variables** (4 points). All worlds should have one binary instrument, one binary treatment variable, and potential outcomes for one variable. You will need to generate potential outcomes both for the outcome variable and the treatment variable (just as in Assignment 5). In World A all assumptions are satisfied. In World B one key assumption is violated. In World C a different key assumption is validated. Define the typical estimand in IV analyses. Estimate a causal effect for this estimand using TSLS.

3) **Regression discontinuity** (3 points). All worlds should have a forcing variable (X), a binary treatment variable, and potential outcomes for a single outcome variable. Assume a sharp discontinuity. World A should satisfy all assumptions. World B should violate a parametric assumption. World C should violate a structural assumption. Estimate a

causal effect for the effect at the threshold using a linear regression with an interaction term with the treatment group.

4) **Regression discontinuity + IV** (7 points) [cannot be combined with (2) or (3) above]. All worlds should have a forcing variable (X), a binary *assignment* variable (Z), a binary treatment receipt variable (D), potential outcomes for the treatment variable (that is participation in the program or receipt of treatment), and potential outcomes for a single outcome variable. In World A all of the assumptions will hold. In World B one of the RD assumptions is violated. In World C one of the IV assumptions is violated. Estimate the causal effect using two-stage least squares.

5) **Difference in differences** (2 points). Both worlds should have observations from both pre-treatment and post-treatment time periods and observations from both exposed and unexposed groups. Each should have at least one covariate and exactly one binary treatment variable. World A satisfies all assumptions. World B violates one assumption.

6) **Fixed Effects** (3 points). All worlds should have 100 participants with 10 observations per person. Each world should have 3 covariates, one binary treatment and potential outcomes for one outcome variable. Assume that this is panel data so that the 10 observations are measured for each person at the same ten equally spaced time points (e.g. every year for 10 years). In World A all of the assumptions hold. In World B at least one of the covariates is a variable that (a) is affected by the treatment and (b) affects the outcome. Use a fixed effects strategy to estimate the effect of the treatment on the outcome.

7) **Bayesian Additive Regression Trees** (4 points). All worlds should have 5 covariates, one binary treatment, and potential outcomes for one variable. The covariates should have a dependence structure. World A can be linear and should satisfy all other assumptions as well. At least one of the response surfaces in World B should be non-linear (with R^2 less than .75) but the other assumptions should hold. World C can use the same response surface as World B but should violate one of the other assumptions. Estimate a causal effect using BART.

Write up: What to turn in for ***each*** submission

The answer for each question should be in the form of an RMarkdown file labeled with your name and the question number, in the following format: LastF_Q#.Rmd. For instance if I was submitted a file for the first question I would name it: HillJ_Q1.Rmd.

Format of the response for **each** submission (corresponding to a method above). You need to submit an **RMarkdown file** that includes the following well-labeled sections:

1) Description of a hypothetical real life scenario that maps to the variables described (e.g. see some of the examples from your homework assignments). *(about 2 paragraphs)*
2) Description of the data generating process. Write out the formal models.
3) R code for data generating process. Provide the R code used to generate the data.

4) Methods and Estimand
    - *a)* Provide a description of the method(s) used *and the estimand*. If the method being addressed is a propensity score approach then also describe the role of the balance and overlap diagnostics. *(3-4 paragraphs)*
    - b) Provide the code used to estimate the results.
5) Discuss the assumptions required for the method to yield a valid causal estimate for the estimand. *(approximately 1/2 page)*
6) Provide results (estimate and s.e. or confidence interval) for each method and world in an attractive display (table or figure). Briefly discuss and provide a causal interpretation for **one** estimate. *(a few paragraphs)*
7) Discuss the bias of each method and tie that to what you did to violate the assumptions in each world. *(about 2 paragraphs)*
8) Conclude with an overview of the lessons you have learned from your simulations. *(about 2-3 paragraphs)*