**640 Final Project Assignment**

**3/28/2023**

**Scope**: The final project can be some new methodological ideas, or a real application (cannot be a simple application of a standard method to a benchmark public data), or ideally both, related to the material covered in the class. A few potential topics in case you don’t have any are given below, but a self-identified project is much more desirable.

**Team composition**: Each team can be 1-3 people.

**Initial sign up**: Please decide your team and tentative topic by **April 14 (Thursday)** and put the brief information to the google sign-up sheet:

<https://docs.google.com/spreadsheets/d/1CXBtOKBF9GzUZ61dJRJP62wpZdYNeEQ8JI8V7ICzHAg/edit?usp=sharing>

**Final report and video presentation**: Each team only need to submit one report and one video presentation of the project. The report is 5 pages maximum, excluding tables and figures, which go to the appendix. Each project should have an accompanying reproducible document that includes your code that produces all the analysis results upon one click. For example, R markdown is such a tool. The link to your video and the reproducible code should be included in your final report. The report (and video) is due on 4/30 (Sunday) 11pm, and will be submitted to the Duke Box folder “STA 640 final project 2023S.” Each team creates their own subfolder on Box.

**Example topics**

1. **Continuous treatment.** We have focused on binary treatment throughout the class, and briefly mentioned multiple (aka multi-valued) treatments in the Chapter of propensity score, but didn’t discuss in detail continuous treatment. Consider how to conduct causal inference within the potential outcome framework for continuous treatments. What are the estimands? How to identify them? What are the assumptions? There are several different directions: (1) propensity score for continuous treatment (e.g. [Hirano and Imbens, 2004](https://www.math.mcgill.ca/dstephens/PSMMA/Articles/HIrano-Imbens-2004.pdf)); (2) incremental treatment effects, where one redefines local treatment effect instead of the global causal dose-response function as the estimand. ([Rothenhäusler and Yu, 2019](https://arxiv.org/abs/1907.13258))
2. **Do fixed/random effects account for unmeasured confounding?** A common saying in econometrics is that fixed/random effects in analyzing panel data absorb unmeasured confounding. In other words, with repeated measurements of the same unit, a model with unit-specific fixed/random effect can bypass unmeasured confounding. Some discussion can be found in Angrist and Pischke (2009, Mostly Harmless Econometrics, Chapter 5). A recent bold example is Dee et al. (2023, Nature Communications). But this claim is a mystery to me, why and in what sense? [Hazlett and Wainstein (2020, Political Analysis](https://www.cambridge.org/core/journals/political-analysis/article/understanding-choosing-and-unifying-multilevel-and-fixed-effect-approaches/8101D49CFD3B129F5753FC878F416980)) has some useful discussion, but it also does not clarify the mystery to me. Hints: some simulations would shed light on this claim.
3. **Instrumental variables with clustered data**. In medical research, a common IV is the cluster-specific proportion of treated (cluster-specific marginal propensity). If we do make the standard IV assumptions: exclusion restriction and non-zero correlation between IV and treatment. Then we have the following questions:
   1. Whether we should use 2SLS with OLS in both stages or 2SLS with random effects model in both stages? **Hint**: I think here we would need to (i) clearly define the estimand, (ii) make assumptions about the true data generating process and treatment assignment. The conclusions can differ.
   2. 2SRI (two-stage residual inclusion): when the treatment is binary, a popular method is the 2SRI, which uses logistic model in the first stage and then uses the residual from stage 1 as an additional predictor in stage too. 2SRI is equivalent to 2SLS if both stages use OLS (see the classic Terza et al. 2008). I was wondering whether the 2SRI estimator is consistent if data is clustered. **Hint**: design a simulation study.
   3. Compare 2SLS with cluster-specific propensity being IV and a propensity score estimator in estimating the ATE. **Hint:** I see it highly depends on the setup. Some preliminary simulation studies are helpful.
4. **Double robust estimator**: does it really help in finite sample? In theory, DR (also known as augmented IPW, AIPW) is the best (semiparametric efficient) estimator with both PS and outcome model is correct, better than IPW and outcome regression estimator in terms of variance. But I had a small simulation study that showed the opposite with sample size of 100 (see appendix). I think one can do some simulations to check this.
5. **Endogenous stratification in randomized experiments**: In medical research, there is an increasing practice of endogenous stratification, advocated by David Kent of Tufts University. The idea is to fit a model (e.g. OLS) that uses baseline characteristics to predicts the potential outcome of an unit in the absence of the treatment, i.e. Y(0), and then stratify the units. This is on the contrary to exogenous stratification, i.e. stratify units based on their pre-treatment covariates. Abadie, Chingos, and West (2018, Endogenous Stratification in Randomized Experiments) had a paper examining this issue and pointed out that sample splitting is crucial to avoid bias arising from such endogenous stratification. Unfortunately, such practice is still common in medicine. In this project, you can use simulations or analytic derivation to further explore this issue. For example, Abadie et al. (2018) used leave-one-out CV for their estimator, you can try things like 10-fold CV.