Complex Experimental Design

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We learned how to use kernel regression to estimate the CATE.
- For any local parameter that is smooth around x, we can estimate its value using information from the neighbors.
- Within the neighborhood, we can use regression models to increase the accuracy (local regression).
- Bandwidth selection is critical in kernel regression since it is about the bias-variance trade-off.
- We can find the optimal bandwidth through cross-validation.

Block randomization

- Controlling covariates in Lin's regression leads to efficiency gains.
- This is an **ex post** adjustment.
- Another way to achieve this is to control for them **ex ante**.
- Suppose we observe $\mathbf{X}_i$ before randomizing the treatment, and dimensionality of the covariates space is not too high.
- Consider the case with two binary covariates, $old_i \in \{0, 1\}$ and $college_i \in \{0, 1\}$.
- We can divide the sample into blocks based on the values of covariates, and randomize within each block:

$$(old_i, college_i) = \begin{cases} (0,0), \\ (0,1), \\ (1,0), \\ (1,1). \end{cases}$$

Block randomization

- Within each block, we may implement either the Bernoulli trial or complete randomization.
- The probability of being treated may vary across blocks:

$$p_i = \begin{cases} 0.2 \text{ if } (old_i, college_i) = (0,0), \\ 0.3 \text{ if } (old_i, college_i) = (0,1), \\ 0.8 \text{ if } (old_i, college_i) = (1,0), \\ 0.4 \text{ if } (old_i, college_i) = (1,1). \end{cases}$$

- It might be desirable to treat more units in groups where the CATE is larger.
- Now, the probability of being treated is a function of $\mathbf{X}_i$: $p_i = P(D_i = 1|\mathbf{X}_i) = g(\mathbf{X}_i)$.
- These treatment assignment mechanisms are individualistic, probabilistic, and unconfounded.
- They are known as block randomization or stratified experiments.

Block randomization: assumption

- Such a design implies the following two assumptions:

$$D_i \perp \{Y_i(0), Y_i(1)\} | \mathbf{X}_i,$$
$$0 < g(\mathbf{X}_i) < 1.$$

- Suppose $g(\mathbf{X}_i)$ is not a constant and $\mathbf{X}_i$ affect the value of Y_i.
- $\mathbf{X}_i$ also affect the value of D_i through $g(\mathbf{X}_i)$.
- If old individuals are treated with a higher probability, there will be more of them in the treatment group.
- Now, $\mathbf{X}_i$ are confounders rather than just moderators.
- The difference-in-means estimator is no longer consistent.

Block randomization: estimation

- We need to account for the difference in the probability of being treated.
- $g(\mathbf{X}_i)$ is a constant within each block.
- We can first estimate the CATE in each block and then take the average over the estimates, weighted by the proportion of each block.
- Suppose we have two groups, the old and the young, with sizes N_O and N_Y.
- The number of treated units in the two groups are N_{1O} and N_{1Y}, respectively.
- We should first estimate τ_O, τ_Y, σ_O^2, and σ_Y^2 as before, and obtain

$$\hat{\tau} = \frac{N_O}{N}\hat{\tau}_O + \frac{N_Y}{N}\hat{\tau}_Y$$

$$\widehat{Var}[\hat{\tau}] = \frac{N_O^2}{N^2}\hat{\sigma}_O^2 + \frac{N_Y^2}{N^2}\hat{\sigma}_Y^2.$$

Block randomization: estimation

- Or, we apply the HT or HA estimator with varying probabilities:

$$\hat{\tau}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i Y_i}{g(\mathbf{X}_i)} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - D_i) Y_i}{1 - g(\mathbf{X}_i)}.$$

$$\hat{\tau}_{HA} = \frac{\sum_{i=1}^{N} D_i Y_i / g(\mathbf{X}_i)}{\sum_{i=1}^{N} D_i / g(\mathbf{X}_i)} - \frac{\sum_{i=1}^{N} (1 - D_i) Y_i / (1 - g(\mathbf{X}_i))}{\sum_{i=1}^{N} (1 - D_i) / (1 - g(\mathbf{X}_i))}.$$

- These estimators are consistent and asymptotically normal.
- Variance estimation can be obtained by adjusting the original formula.
- Therefore, we have two (equivalent) approaches to estimate the ATE.
- But they are based on slightly different ideas.

Block randomization: estimation

- We may still rely on the regression estimator under block randomization.
- The first estimator is equivalent to running Lin's regression with block indicators as covariates.
- In the previous example, we have

$$Y_i = \mu + \tau D_i + \beta(old_i - \overline{old}) + \delta D_i * (old_i - \overline{old}) + \varepsilon_i.$$

- The second estimator (Hajek) is equivalent to the WLS estimator, as illustrated before.
- Remember that the weight for unit i equals

$$W_i = \frac{D_i}{g(\mathbf{X}_i)} + \frac{1 - D_i}{1 - g(\mathbf{X}_i)}$$

- Again, we should apply the HC2 variance estimator.

Block randomization: simulation

```
## The true ATE is 3.233

## The estimate from the unweighted HA estimator is 3.634

## The estimate from the weighted HA estimator is 3.012

## The Lin's regression estimates is 3.012

## The WLS estimates is 3.012
```

Block randomization: discussion

- We should block on covariates which have a strong prediction power for the outcome.
- Sometimes we use existing strata like schools or villages.
- Blocking is ensured to reduce the variance of your estimator, since
$$Var[E[\hat{\tau}|\mathbf{X}_i]] \leqslant Var[\hat{\tau}].$$
- We can combine block randomization with regression adjustment, e.g., applying Lin's regression within each block.
- There exists a tradeoff between the balance in observable covariates and the balance in unobservable ones (Harshaw et al. 2019).
- To improve how well our experiment predicts the reality, balancing all the variables may not be optimal.

Cluster-randomized experiments

- Sometimes it is impossible or too costly to assign the treatment at the unit level.
- Instead, we randomize at a higher level, such as villages, schools, clinics, etc.
- Each unit at this higher level is called a cluster, denoted as $\{\mathcal{C}_c\}_{c=1}^{C}$.
- Every unit in the same cluster receives the same treatment.
- $E[D_i D_j] = E[D_c]$ in a cluster-randomized experiment for i and j belonging to cluster c.
- The covariance between i and j from the same cluster is no longer zero.
- A cluster is different from a stratum or block!
- We can still rely on the estimators we have learned.
- But the standard errors have to be adjusted (clustered).

Clustered standard errors

- With the regression estimator, our variance estimator takes the sandwich form:

$$\widehat{Var}\left[\hat{\beta}\right] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\hat{\varepsilon}\hat{\varepsilon}'\mathbf{X}')(\mathbf{X}'\mathbf{X})^{-1},$$

  where $\mathbf{X}\hat{\varepsilon}\hat{\varepsilon}'\mathbf{X}' = \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \mathbf{X}_i \mathbf{X}_i'$.
- All the off-diagonal elements in $\hat{\Sigma} = \hat{\varepsilon}\hat{\varepsilon}'$ equal to zero.
- This is true only when the units are independent to each other (as $E[\varepsilon_i \varepsilon_j] = 0$).
- In cluster-randomized experiments, dependence within clusters leads to the fact that $E[\varepsilon_i \varepsilon_j] \neq 0$ if i and j belong to the same cluster.
- Remember that $\varepsilon_i = Y_i(0) - \bar{Y}(0) + (\tau_i - \tau)D_i$.

Clustered standard errors

- The sandwich variance estimator is still valid.
- But we need to calculate the off-diagonal elements in $\hat{\Sigma}$:

$$\mathbf{X}\hat{\varepsilon}\hat{\varepsilon}'\mathbf{X}' = \sum_{c=1}^{C} \mathbf{X}_c\hat{\varepsilon}_c\hat{\varepsilon}_c'\mathbf{X}_c'$$

$$= \sum_{i=1}^{N} \hat{\varepsilon}_i^2\mathbf{X}_i\mathbf{X}_i' + \sum_{i=1}^{N} \sum_{j:\{i,j\in\mathcal{C}_c\}} \hat{\varepsilon}_i\hat{\varepsilon}_j\mathbf{X}_i\mathbf{X}_j',$$

  where $\hat{\varepsilon}_c$ represents the regression residuals for units from cluster c, and $\mathbf{X}_c$ represents the covariates of these units.
- The robust standard error may overestimate or underestimate the true variance.
- It hinges on the correlation between units within the same cluster.

Clustered standard errors

- The model-based approach justifies clustered standard errors with unobservable group attributes.
- But this is not testable and leads to confusions in practice.
- Abadie et al. (2017): we should cluster standard errors when there is clustering in either sampling or treatment assignment.
- Focus on clustering in the treatment rather than the outcome.
- There is no need to always cluster at the highest level.
- Regression estimator + standard errors clustered at the level of assignment or sampling.
- FRT or bootstrap should be implemented at the cluster level.

Asymptotics in cluster-randomized experiments

- Because units in the same cluster are dependent with each other, the effective sample size is smaller than N.
- As N grows, C should grow while $\frac{N}{C}$ remains stable.
- Then, the asymptotic distribution will be normal and the convergence rate to be $\sqrt{C}$ (Su and Ding 2021).
- In practice, we need a large number of clusters.
- There should be no cluster that is much larger than the others in size.

Covariate adjustment with clustering

- When clusters have the same size, an equivalent approach is to calculate the average outcome $\bar{Y}_c$ for each cluster and analyze the experiment at this aggregated level.
- Covariates at the individual level should also be aggregated.
- We can then use Lin's regression with covariates at both the individual and the cluster level.
- It enhances efficiency, and we only need to use the robust variance estimator.
- Su and Ding (2021) show that it leads to bias if clusters differ in sizes.
- We should use the weighted average outcome (and covariates) in each cluster:

$$\bar{Y}_c^* = \frac{C}{N} \sum_{i \in \mathcal{C}_c} Y_i.$$

Cluster-randomized experiments: application

```
## The SATE is 2.969836

## The OLS estimate is 2.99

## The OLS estimate using aggregated outcome is 2.99

## The true variance of the OLS estimate is 0.165

## The true variance of the OLS estimate
##  using aggregated outcome is 0.165

## The estimated variance of the OLS estimate is 0.054

## The estimated variance of the OLS estimate
##  using aggregated outcome is 0.232

## The clustered variance of the OLS estimate is 0.239
```

References I

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" National Bureau of Economic Research.

Harshaw, Christopher, Fredrik Sävje, Daniel Spielman, and Peng Zhang. 2019. "Balancing Covariates in Randomized Experiments Using the Gram-Schmidt Walk." *arXiv Preprint arXiv:1911.03071*.

Su, Fangzhou, and Peng Ding. 2021. "Model-Assisted Analyses of Cluster-Randomized Experiments." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (5): 994–1015.