Observational Studies

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We can increase the efficiency of estimating the SATE via block randomization.
- The probability of being treated can be different across blocks.
- If so, covariates that determine the blocks become confounders.
- We should either estimate the CATEs across the blocks and aggregate them, or adjust the probability of being treated for each unit.
- When an entire cluster is assigned to the treatment status, units within the cluster are dependent on each other.
- We need to account for the uncertainty using clustered standard error.

From experiments to observational studies

- Causal identification hinges on randomization of the treatment.
- The method of difference is not feasible due to the curse of dimensionality.
- If the treatment assignment is (conditionally) randomized, all the confounders are balanced in expectation.
- The same idea applies to observational studies.
- The key is to derive credible identification assumptions based on our substantive knowledge.
- The identification assumptions should assert that the treatment is independent to the potential outcomes conditional on certain variables.
- Their validity depends on our understanding of the treatment assignment mechanism.

Uniqueness of observational studies

- In observational studies, we don't know the exact treatment assignment mechanism.
- It could be a real experiment implemented by a third party.
- E.g., the Brazilian government conducts auditing on a randomly selected set of municipalities.
- We know there is something at random, but not the probability of being treated for each unit.
- Therefore, we have two tasks in observational studies:
  - clarify the source of randomness (the identification assumption), and
  - estimate the treatment assignment mechanism (based on structural restrictions).
- The first should be supported by our substantive knowledge (design of the study).
- The second is a statistical problem.
- We should separate these two tasks when evaluating or conducting a study.

From block randomization to observational studies

- In observational studies, it is usually difficult to argue that

$$D_i \perp \{Y_i(0), Y_i(1)\}.$$

- It could be satisfied in certain scenarios.
- More commonly, we assume that

$$\begin{aligned} &D_i \perp \{Y_i(0), Y_i(1)\} | \mathbf{X}_i \text{ (unconfoundedness)}, \\ &0 < P(D_i = 1 | \mathbf{X}_i) < 1 \text{ (positivity)}. \end{aligned}$$

- The two parts altogether is called strong ignorability, conditional exogeneity, or exchangeability.
- Note that positivity is automatically satisfied in experiments but not in observational studies.

From block randomization to observational studies

- The assumptions are exactly what we make in block randomization.
- When analyzing an observational study under these two assumptions, we actually assume that the data are generated by a hypothetical block randomization.
- There may not be actual blocks when $\mathbf{X}_i$ include continuous variables.
- We may know the variables used for blocking but not the assignment mechanism.
- Remember that we have two approaches to estimate the ATE in blocking experiments.
- We either estimate the CATEs across blocks and aggregate them, or weight each unit with the probability of being treated.
- Each approach has a regression representation.

From block randomization to observational studies

- Hence, in an observational study under the two assumptions, we can control the covariates by
  1. classify observations into groups defined by the covariates,
  2. try to estimate the probability of being treated,
  3. direct model the relationship between Y_i and $(D_i, \mathbf{X}_i)$.
- The first approach leads to matching,
- The second one leads to weighting.
- The third one leads to regression.
- They are equivalent in block randomization but not in observational studies due to difference in structural restrictions.

The role of the propensity score

- We have seen that the first two approaches are both valid in block randomization.
- If we know the probability of being treated, we don't need the blocks formed by the covariates.
- In other words, if we can control the difference in the probability of being treated, we have controlled for the difference in covariates.
- This conclusion is first reached by Rosenbaum and Rubin (1983) in observational studies.
- They call the probability of being treated, $g(\mathbf{X}_i)$, the propensity score.
- They show that under strong ignorability,

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

The role of the propensity score (*)

- First note that propensity score is a balancing score in the sense that

$$D_i \perp \mathbf{X}_i | g(\mathbf{X}_i).$$

- The reason is that

$$\begin{aligned} & P[D_i = 1 | \mathbf{X}_i, g(\mathbf{X}_i)] = P[D_i = 1 | \mathbf{X}_i] = g(\mathbf{X}_i), \\ & P[D_i = 1 | g(\mathbf{X}_i)] = E[D_i | g(\mathbf{X}_i)] \\ = & E[E[D_i | g(\mathbf{X}_i), \mathbf{X}_i] | g(\mathbf{X}_i)] \\ = & E[P[D_i = 1 | \mathbf{X}_i, g(\mathbf{X}_i)] | g(\mathbf{X}_i)] \\ = & E[g(\mathbf{X}_i) | g(\mathbf{X}_i)] = g(\mathbf{X}_i). \end{aligned}$$

The role of the propensity score (*)

- Then we can see that

$$
\begin{aligned}
& P[D_i = 1|Y_i(0), Y_i(1), g(\mathbf{X}_i)] \\
=& E[D_i|Y_i(0), Y_i(1), g(\mathbf{X}_i)] \\
=& E[E[D_i|Y_i(0), Y_i(1), g(\mathbf{X}_i), \mathbf{X}_i]|Y_i(0), Y_i(1), g(\mathbf{X}_i)] \\
=& E[E[D_i|g(\mathbf{X}_i), \mathbf{X}_i]|Y_i(0), Y_i(1), g(\mathbf{X}_i)] \\
=& E[E[D_i|g(\mathbf{X}_i)]|Y_i(0), Y_i(1), g(\mathbf{X}_i)] \\
=& E[D_i|g(\mathbf{X}_i)] \\
=& P[D_i = 1|g(\mathbf{X}_i)]
\end{aligned}
$$

- The propensity score, as a uni-dimensional variable, contains all the information in the high-dimensional covariates $\mathbf{X}_i$.

Estimate the propensity score

- This is a prediction problem.
- We want to find a $\hat{g}(\mathbf{X}_i)$ that approximates $g(\mathbf{X}_i)$ well.
- It is common for researchers to impose structural restrictions on $g(\cdot)$, such as
$$g(\mathbf{X}_i) = \frac{e^{\mathbf{X}_i'\beta}}{1 + e^{\mathbf{X}_i'\beta}}.$$
- This is known as the logistic model.
- It is ensured that $\hat{g}(\mathbf{X}_i) \in [0, 1]$.
- The model is a transformation of the linear model $\mathbf{X}_i'\beta$.
- The transformation is known as the link function.
- We estimate the parameter β via maximum likelihood estimation.

Estimate the propensity score

- Suppose we know $g(\mathbf{X}_i)$, then the probability for us to observe $\mathcal{D} = (D_1, D_2, \ldots, D_N)$ is

$$L = \prod_{i=1}^{N} g(\mathbf{X}_i)^{D_i}(1 - g(\mathbf{X}_i))^{1-D_i}.$$

- We find $\hat{\beta}$ such that

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} L \\
&= \arg\max_{\beta} \log L \\
&= \arg\max_{\beta} \sum_{i=1}^{N} [D_i \log(g(\mathbf{X}_i)) + (1 - D_i) \log(1 - g(\mathbf{X}_i))]
\end{aligned}$$

Estimate the propensity score

- Again, $\hat{\beta}$ can be found via the first order condition; note that

$$
\begin{aligned}
\frac{d\log(g(\mathbf{X}_i))}{d\beta} =& \frac{1}{g(\mathbf{X}_i)}\frac{dg(\mathbf{X}_i)}{d\beta} \\
=& \frac{1}{g(\mathbf{X}_i)}\frac{\mathbf{X}_i e^{\mathbf{X}_i'\beta}(1+e^{\mathbf{X}_i'\beta}) - \mathbf{X}_i e^{\mathbf{X}_i'\beta} e^{\mathbf{X}_i'\beta}}{(1+e^{\mathbf{X}_i'\beta})^2} \\
=& \frac{1+e^{\mathbf{X}_i'\beta}}{e^{\mathbf{X}_i'\beta}} * \frac{\mathbf{X}_i e^{\mathbf{X}_i'\beta}}{(1+e^{\mathbf{X}_i'\beta})^2} \\
=& \frac{\mathbf{X}_i}{1+e^{\mathbf{X}_i'\beta}} = \mathbf{X}_i(1-g(\mathbf{X}_i)).
\end{aligned}
$$

- Similarly, $\frac{d\log(1-g(\mathbf{X}_i))}{d\beta} = -\mathbf{X}_i g(\mathbf{X}_i)$

Estimate the propensity score

- Eventually, we have

$$
\begin{aligned}
\frac{d \log L}{d\beta} &= \sum_{i=1}^{N} \left[D_i \frac{d \log(g(\mathbf{X}_i))}{d\beta} + (1 - D_i) \frac{d \log(1 - g(\mathbf{X}_i))}{d\beta} \right] \\
&= \sum_{i=1}^{N} [D_i \mathbf{X}_i (1 - g(\mathbf{X}_i)) - (1 - D_i) \mathbf{X}_i g(\mathbf{X}_i)] \\
&= \sum_{i=1}^{N} (D_i - g(\mathbf{X}_i)) \mathbf{X}_i
\end{aligned}
$$

- The equation $\sum_{i=1}^{N} (D_i - g(\mathbf{X}_i))\mathbf{X}_i = \mathbf{0}$ does not have a close-form solution and must be solved numerically.
- Statistical software will do this for us!
- We can adopt more complex models for $g(\mathbf{X}_i)$ (e.g., ML algorithms like the probability forest).

Estimate the response surface

- We often call the two conditional expectations $E[Y_i|D_i = 1, \mathbf{X}_i]$ and $E[Y_i|D_i = 0, \mathbf{X}_i]$ the response surfaces.
- We denote them as $m_1(\mathbf{X}_i)$ and $m_0(\mathbf{X}_i)$.
- If we can estimate them consistently, then an estimator for the SATE is

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}[\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)].$$

- We may impose the structural restriction that both $m_1(\mathbf{X}_i)$ and $m_0(\mathbf{X}_i)$ are linear functions, hence

$$\begin{aligned}
m_1(\mathbf{X}_i) &= \tau_1 + \mathbf{X}_i'\beta_1, \\
m_0(\mathbf{X}_i) &= \tau_0 + \mathbf{X}_i'\beta_0, \\
\hat{\tau} &= \frac{1}{N}\sum_{i=1}^{N}[\hat{\tau}_1 - \hat{\tau}_0 + \mathbf{X}_i'\hat{\beta}_1 - \mathbf{X}_i'\hat{\beta}_0].
\end{aligned}$$

Summary

- We call either the propensity score or the response surface "nuisance parameters."
- The target parameter is the SATE while the nuisance parameters are just intermediates we have to estimate.
- Note that the first approaches allow the effects to vary arbitrarily across units.
- The third approach assumes that all the heterogeneity can be explained by $\mathbf{X}_i$.
- This may lead to problems in practice.
- We can combine the these approaches to achieve more robust estimates
- But the pre-condition is that strong ignorability is satisfied.
- It is crucial to validate this assumption through various means.
- Methods only differ in the way to handle nuisance parameters.

Evaluate methods for observational studies

- We will use the classic example from LaLonde (1986) in the next few lectures.
- This study compares the treatment group in an experiment with both the real control group and a control group drawn from the population (CPS and PSID).
- It thus provides a benchmark (the experimental estimate) for evaluating different methods in observational studies.
- Treatment: skill training in the National Supported Work Demonstration (NSW) program.
- Outcome: annual income in 1978.
- Covariates: age, education, race, married, plus income and employment status in 1974 and 1975.

Evaluate methods for observational studies

```
## The OLS estimate is 1794.343

## The SE of OLS estimate is 670.9967

## The Lin regression estimate is 1583.468

## The SE of Lin regression estimate is 678.0574
```

Evaluate methods for observational studies

```
## The OLS estimate is -15204.78
## The SE of OLS estimate is 657.0765
## The Lin regression estimate is -8746.283
## The SE of Lin regression estimate is 4398.952
```

Evaluate methods for observational studies

```
##                 mean.Tr  mean.Co   sdiff T pval
## age              25.816   25.054  10.655  0.266
## education        10.346   10.088  12.806  0.150
## black             0.843    0.827   4.477  0.647
## hispanic          0.059    0.108 -20.341  0.064
## married           0.189    0.154   9.000  0.334
## nodegree          0.708    0.835 -27.751  0.002
## re74           2095.574 2107.027  -0.234  0.982
## re75           1532.056 1266.909   8.236  0.385
## u74               0.708    0.750  -9.190  0.330
## u75               0.600    0.685 -17.225  0.068
```

Evaluate methods for observational studies

```
##              mean.Tr   mean.Co    sdiff T pval
## age           25.816    34.851 -126.266  0.000
## education     10.346    12.117  -88.077  0.000
## black          0.843     0.251  162.564  0.000
## hispanic       0.059     0.033   11.357  0.132
## married        0.189     0.866 -172.406  0.000
## nodegree       0.708     0.305   88.378  0.000
## re74        2095.574 19428.746 -354.707  0.000
## re75        1532.056 19063.338 -544.576  0.000
## u74            0.708     0.086  136.391  0.000
## u75            0.600     0.100  101.786  0.000
```

Evaluate methods for observational studies

- What causes the striking differences?
- Heckman, Ichimura, and Todd (1997) suggest that there are four possibilities.
  1. the treated and the non-experimental untreated differ in unobservable attributes (selection bias).
  2. the treated and the non-experimental untreated differ in observable attributes.
  3. different questionnaires are used in the experiment and the observation study.
  4. these individuals come from different economic environments.
- Their analysis finds that 2, 3, and 4 are more important than 1.

References I

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–54.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 604–20.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.