Heterogeneous Treatment Effects I

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We discussed the causal interpretation of the OLS estimator in the previous class.
- In randomized experiments, the OLS estimator equals the Hajek estimator.
- The HC2 variance estimator equals the Neyman variance estimator.
- We may use regression adjustment to control for covariates and enhance the efficiency of the OLS estimator.
- This is justified by the FWL theorem when the model specification is correct.
- Otherwise, we can rely on Lin's regression to ensure the increase in efficiency.

From ATE to CATE

- Sometimes we want to know the average treatment effect on a sub-population:
$$\tau(\mathbf{x}) = E[\tau_i|\mathbf{X} = \mathbf{x}].$$
- This is known as the conditional average treatment effect (CATE).
- It allows us to see how the effects vary within the population and helps researchers to design more personalized policy or medicine.
- Note that $\mathbf{X}$ should not be affected by the treatment.
- It is sometimes called the moderator.

From CATE to optimal assignment

- CATE allows us to figure out the optimal assignment of the treatment.
- It provides a natural measure of the benefit for each subgroup.
- An assignment mechanism is a mapping from the covariates to the probability of being treated.
- The optimal assignment mechanism hinges on our knowledge of CATE.
- If the average effect is positive for women and negative for men, we should only treat women in the sample:

$$P(D_i = 1) = \begin{cases} 1 & male_i = 0 \\ 0 & male_i = 1 \end{cases}.$$

Optimal assignment

- In general, we want to find a mapping (also known as a policy) $\pi(\mathbf{X}) \in \Pi$ that maximizes a welfare function $W(\pi)$.
- $\pi(\mathbf{X})$ can be deterministic or stochastic.
- We usually need to impose restrictions on Π, such that it is not too complicated.
- For example, we can rely on the linear eligibility score:

$$P(D_i = 1) = \begin{cases} 1 & \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} \geq 0, \\ 0 & \text{Otherwise.} \end{cases}$$

- The optimal policy in Π may not be the first-best policy:

$$P(D_i = 1) = \begin{cases} 1 & \tau(\mathbf{X}_i) \geq 0, \\ 0 & \text{Otherwise.} \end{cases}$$

Optimal assignment

- $W(\pi)$ is decided by the objective of the researcher.
- Do we want to maximize the total utility? Do we want to prevent harm? Do we want to promote fairness?
- Different objects lead to different $\pi^*(\mathbf{X})$.
- If we know $\tau(\mathbf{x})$, finding $\pi^*(\mathbf{X})$ is a pure optimization problem.
- E.g, we can find $\beta = (\beta_0, \beta_1, \ldots, \beta_P)$ that maximizes

$$\sum_{i=1}^{N} \tau(\mathbf{X}_i) \mathbf{1}\left\{ \beta_0 + \sum_{p=1}^{P} \beta_p x_{ip} \geq 0 \right\}.$$

- In practice, we need to estimate $\tau(\mathbf{x})$ first and find $\hat{\pi}^*(\mathbf{X})$ that minimizes the "regret:"

$$E[W(\pi^*(\mathbf{X}_i)) - W(\hat{\pi}^*(\mathbf{X}_i))].$$

Optimal assignment

- Scholars in this field are working on deriving the optimal assignment mechanism in various scenarios.
- How do we incorporate different constraints into this problem?
- What if the treatment status of one unit affects the outcome of other units?
- In dynamic experiments, how can we learn the optimal combination of treatments and implement it ASAP?
- How do we combine information from multiple studies to make policy learning more accurate?

From CATE to external validity

- CATE is also closely connected to the external validity of a study.
- Remember that if we have a representative sample, the estimate of SATE is consistent for PATE as well.
- But this is rarely the case.
- We want to know some general laws of human behavior.
- But the sample often comes from one country or even one county.
- How do we generalize our estimate obtained from one sample to the population?

External validity

- We need to understand how SATE differs from PATE.
- One possibility: it is completely driven by the difference in demographic composition.
- Suppose the only variable that affects the effect's size is age and our experiment is conducted in a county with more senior people.
- To generalize the conclusion to the whole country, we just need to reweigh our sample with the proportion of senior residents in America.
- A more severe issue is known as the site-selection bias.
- There are unobservable factors that are correlated with both the effects and where the experiment is implemented.
- It is an open question in the literature.

Estimate CATE

- The remaining question: how do we estimate the CATE?
- If **X** only includes binary variables, we can estimate the ATE conditional on each value of **X**.
- It is equivalent to estimating a regression model with an interaction term:

$$Y_i = \mu + \tau D_i + \beta X_i + \delta D_i * X_i + \varepsilon_i.$$

- Such a model is "saturated" as it covers all the combinations of D_i and X_i.
- The estimated effect of D_i equals $\hat{\tau}$ if $X_i = 0$ and $\hat{\tau} + \hat{\delta}$ if $X_i = 1$.

Estimate CATE

- Note that X_i is not randomly assigned, hence the difference between $\tau(1)$ and $\tau(0)$ does not have a causal interpretation.
- E.g., we cannot say "turning old increases the effect by 20%."
- It is different from

$$Y_i = \mu + \tau D_{1i} + \beta D_{2i} + \delta D_{1i} * D_{2i} + \varepsilon_i,$$

  where both D_1 and D_2 are randomly assigned.
- If interested in the interaction effect, we have to control for confounders that affect X_i.

Estimate CATE

- If **X** includes continuous variables, the convention is to fit the same regression model.
- We have learned that Lin's regression is the better approach:

$$Y_i = \mu + \tau D_i + (X_i - \bar{X})\beta + \delta D_i * (X_i - \bar{X}) + \varepsilon_i.$$

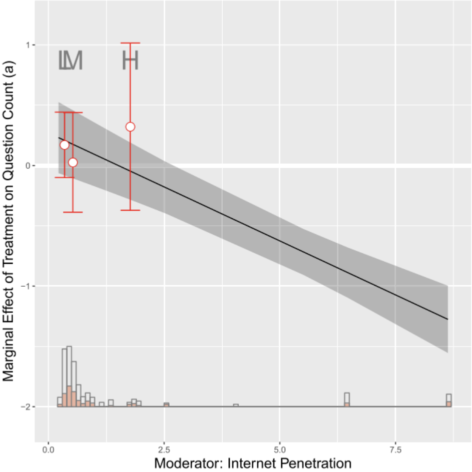
- The estimated moderator effect equals $\hat{\tau} + \hat{\delta}(X_i - \bar{X})$, a linear function of X.
- There is no guarantee that this linear relationship holds.

Caveats of interaction models

- Consider the following application in Malesky, Schuler, and Tran (2012).
- It is an experiment implemented in Vietnam.
- Treatment: an online profile for randomly selected legislators that documents their performance.
- Outcome: questions a legislator asked in Congress.
- Their ATE estimate is not significant.
- But the interaction model shows that the effect is significant in regions where the Internet penetration rate is high.

Caveats of interaction models

- Hainmueller, Mummolo, and Xu (2019) show that the estimate is entirely driven by certain regions.



Caveats of interaction models

- This example illuminates the problems of relying on linear models.
- The predictions can be very inaccurate if the true pattern is not quite linear.
- The results can be influenced by a few observations in the sample.
- It is because regression is a global model.

Estimate the CATE flexibly

- Remember that we want to estimate

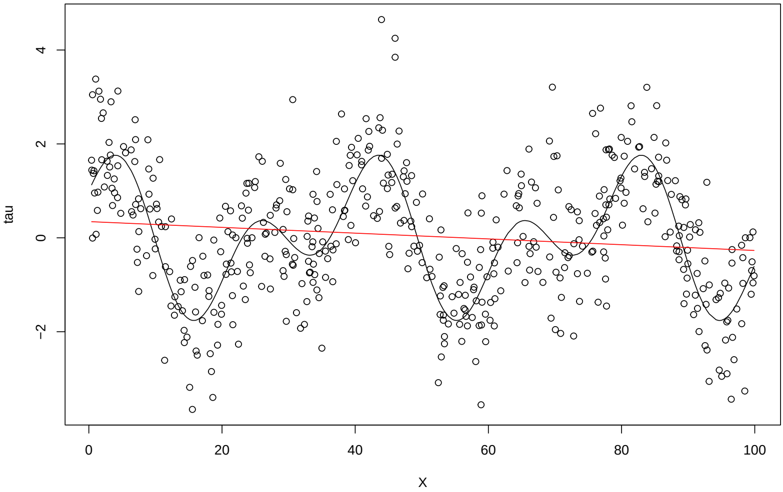
$$\tau(x) = E[\tau_i | X_i = x].$$

  without assuming a linear relationship.
- Let's first assume we know the value of each τ_i.
- It becomes a problem of estimating the conditional expectation of a variable.
- This is a prediction problem rather than a causal inference problem.

Estimate conditional expectation

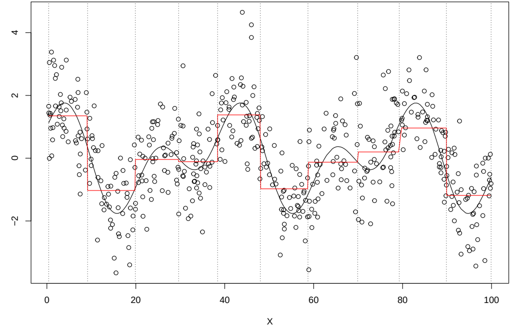
- Later we discuss how to deal with the problem of estimating the CATE using similar techniques.
- We have learn the regression approach, which assumes that $\tau(x) = \beta x$.
- Instead of linearity, let's only assume the smoothness of $\tau(x)$.
- This is much weaker and satisfied in many scenarios.
- A common form of such an assumption is the sth order derivative of $\tau(x)$ exists.

Estimate conditional expectation



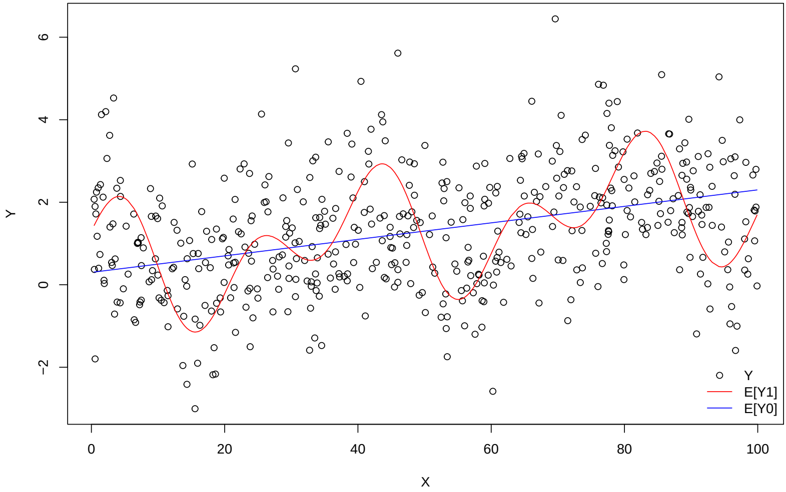
The binscatter estimator

- Smoothness means that if x' is close to x, then $\tau(x')$ is close to $\tau(x)$.
- Therefore, we can estimate $\tau(x)$ using information from $\tau(x')$.
- A natural estimator is to divide the support of X into K bins and estimate $\tau(x)$ using the average of τ_i within each bin.

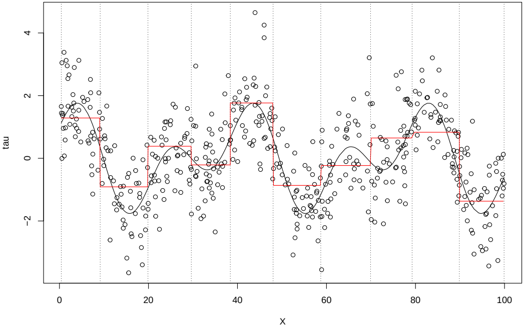


The binscatter estimator for the CATE

- With unknown τ_i, we apply the HT or HA estimator in each of the bins.

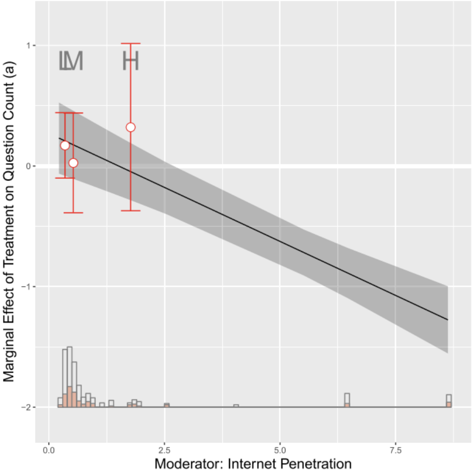


The binscatter estimator for the CATE



The binscatter estimator for the CATE

- Hainmueller, Mummolo, and Xu (2019) suggest that we use three bins.



- There are a lot of different choices (Cattaneo et al. 2019).
- Note that the estimator is clearly biased.

References I

Cattaneo, Matias D, Richard K Crump, Max H Farrell, and Yingjie Feng. 2019. "On Binscatter." *arXiv Preprint arXiv:1902.09608*.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27 (2): 163–92.

Malesky, Edmund, Paul Schuler, and Anh Tran. 2012. "The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly." *American Political Science Review* 106 (4): 762–86.