Instrumental Variable for Causal Effects

Paper review on Angrist & Imbens (1995), and Syrgkanis et al. (2019)

Presenters:
Sakina Dhorajiwala (M. International Development Policy)
Shota Miki (M.S. Economics & Computation)

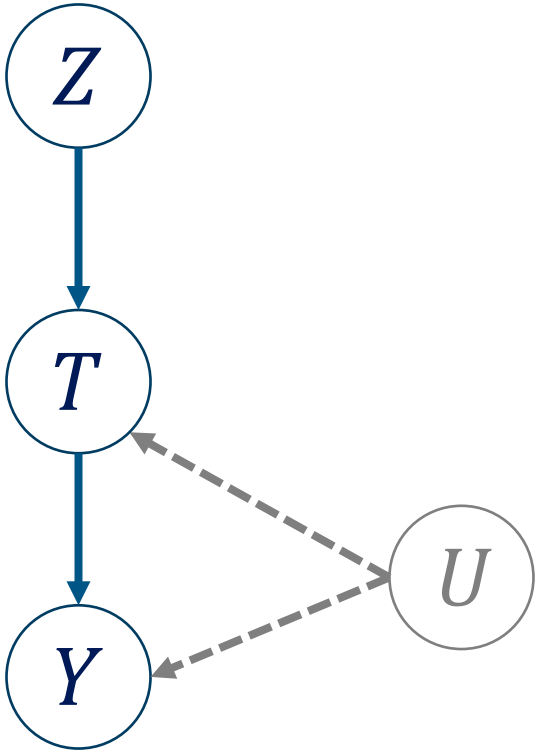
Duke

Outline

- Angrist and Imbens (1995)
    - What is IV & When do we use it?
    - Example of American Charter Schools
    - Using IV to estimate LATE
- Syrgkanis et al. (2019)
    - Overview of DRIV
    - How it works
    - Demonstration

Duke

Overview



When do we Use IV?

- Concerns of endogeneity i.e., when T is correlated with the error term (U)
- So, we find a variable (Z) that is correlated with the predictor variable of interest (T), but is not correlated with the error term (U)

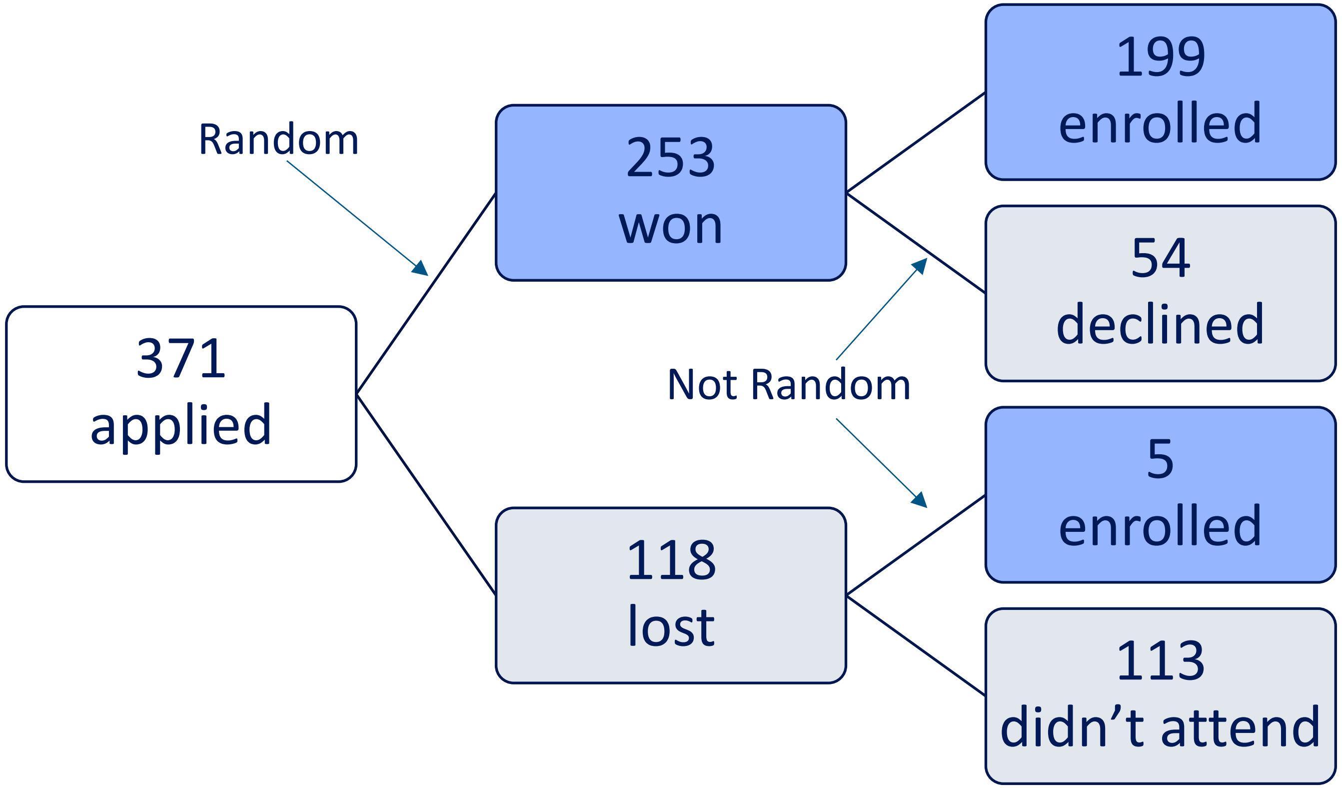
Simple IV estimation is like…

- Z: instrument
- T: treatment
- Y: outcome
- U: confounding variables (unobserved)

Duke

Example: Do Charter Schools students have better quality education?

Does *attending* a charter school lead to better educational outcomes?



Duke

Source: https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one

Calculating the treatment effect

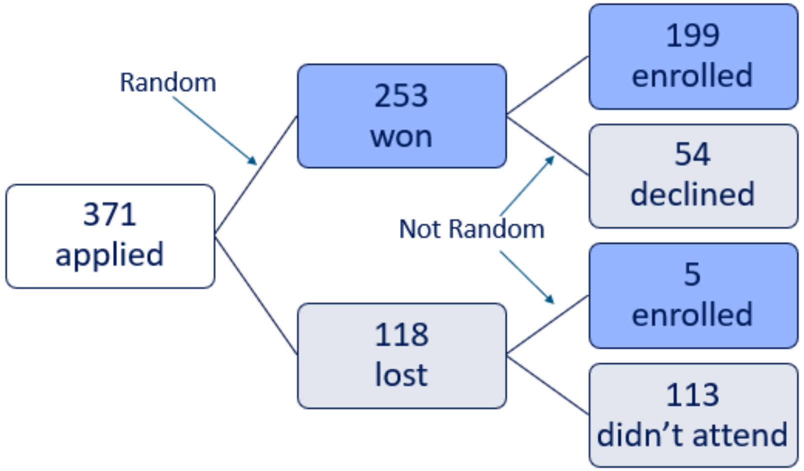
Instrument

Winning Lottery → Charter Attendance → Math Scores

First stage: Effect of winning on attendance

Second stage: Effect of attendance Math Scores

Reduced form: Effect of winning on scores



Duke

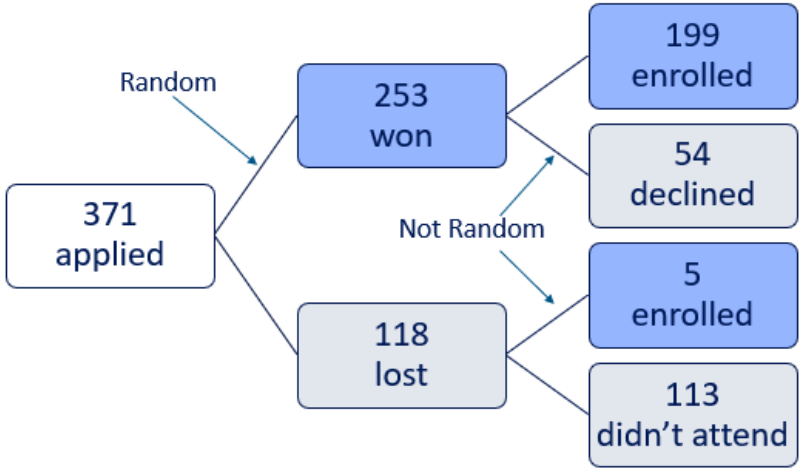
Source: https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one

Calculating the treatment effect

Effect of winning on attendance = % of winners who attend - % of losers who attend

Effect of winning on scores = scores of winners – scores of loser

Effect of winning on attendance	×	Effect of attendance on scores	=	Effect of winning on math score
First stage		Second stage		Reduced form



Duke

Source: https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one

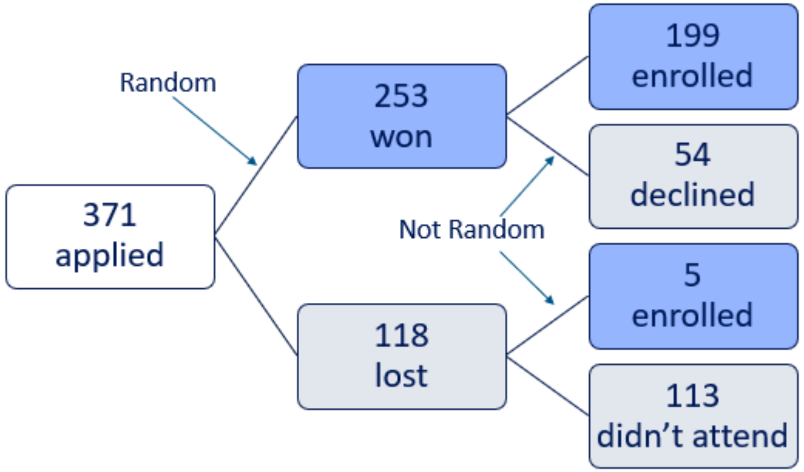
Calculating the treatment effect

Effect of winning on attendance = % of winners who attend - % of losers who attend

Effect of winning on scores = scores of winners – scores of loser

Effect of winning on attendance × Effect of attendance on scores = Effect of winning on math score

First stage | Second stage | Reduced form



$$\text{Effect of Charter Attendance on Math scores} = \frac{\text{Effect of winning on scores}}{\text{Effect of winning on attendance}} = \frac{0.36\sigma\ (given)}{(199/253) - (5/118)\ = 0.74} = \boxed{0.48\sigma}$$

Source: https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one

Calculating the treatment effect

Effect of winning on attendance	×	Effect of attendance on scores	=	Effect of winning on math score
First stage		Second stage		Reduced form

$$\text{Effect of Charter Attendance on Math scores} = \frac{\text{Effect of winning on scores}}{\text{Effect of winning on attendance}}$$

These estimates are for kids opting into the lottery, whose enrollment status is changed by winning. That's not necessarily a random sample of all children

Duke

Source: https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one

Angrist & Imbens (1991, NBER)

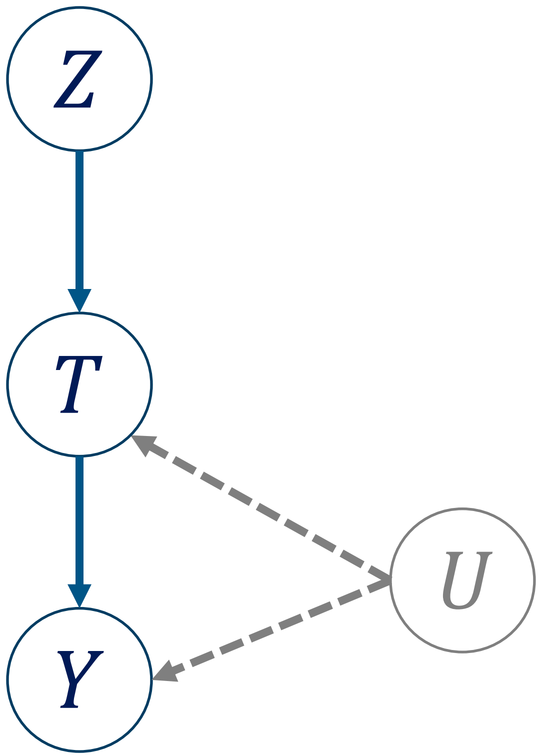
Later published in Econometrica (1995)

Duke

Angrist & Imbens (1995)

- Use of instrumental variables (IV) to estimate LATEs, i.e., average treatment effect for individuals who would only receive treatment if they complied with the treatment
- Helps in identifying the ATE when there is no group available for whom the probability of treatment is zero
- This is done for individuals whose treatment status is influenced by changing an exogeneous regressor that satisfies the exclusion restriction
- The incentives for participation are randomized, not the participation status itself

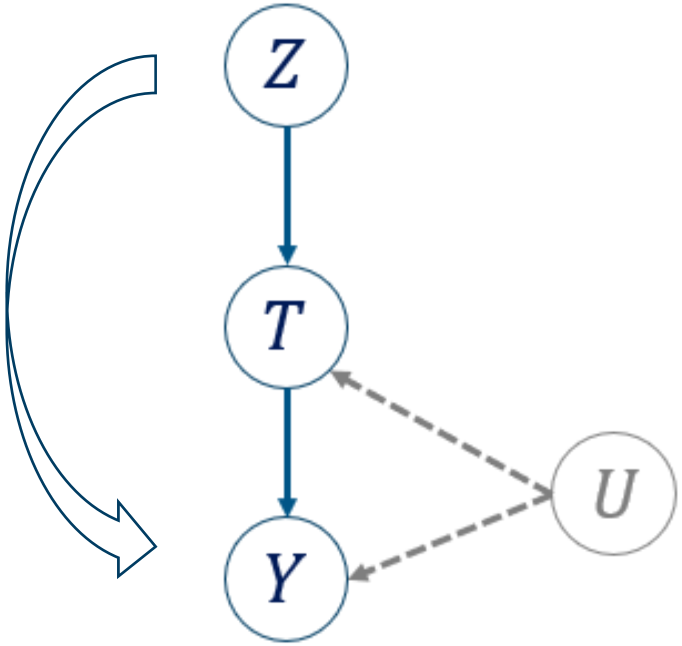
Condition 1: Existence of Instruments



1. Relevance: Z has a causal effect on T
2. Exclusion Restriction: The causal effect of Z on Y is only through T (no direct path from Z to T)
3. Instrumental Unconfoundedness: No unblockable paths from Z to Y
   - If there is a backdoor path (observed W), you can condition on it. Making Z a conditional instrument

Duke

Non parametric Identification of Local ATE



$Y(T=1)\ Y(T=0)$ or $Y(0)\ Y(1)$

Potential Outcomes when treatment takes values 0 or 1

$T(Z=1)\ T(Z=0)$ *or* $T(1)\ T(0)$

Potential Treatments (under the instrument) when you intervene on instrument Z {0,1}

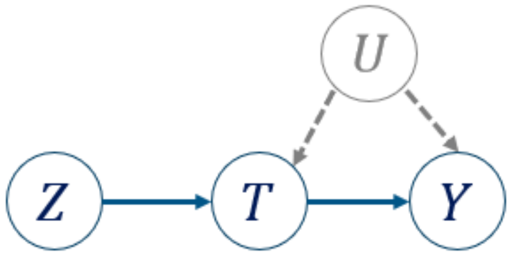
$Y(Z=1)\ Y(Z=0)$

Potential Outcomes when we are intervening on the instrument (instead of the treatment)

Duke

Content source: Brady Neal, Intro to Causal Inference

Principal Strata



$\mathrm{T}(\mathrm{Z}=1)=1, \qquad T(\mathrm{Z}=0)=0$

Compliers: Who always take the treatment they are assigned (Z)

$\mathrm{T}(\mathrm{Z}=1)=0, \qquad T(\mathrm{Z}=0)=1$

Defiers: Who always take the treatment they are NOT assigned

$\mathrm{T}(\mathrm{Z}=1)=1, \qquad T(\mathrm{Z}=0)=1$

Always Takers: Who ALWAYS take the treatment irrespective of assignment

$\mathrm{T}(\mathrm{Z}=1)=0, \qquad T(\mathrm{Z}=0)=0$

Never Takers: Who NEVER take the treatment irrespective of assignment

Duke

Content source: Brady Neal, Intro to Causal Inference

The Monotonicity Assumption (No Defiers)

Condition 2, according to Angrist and Imbens

$\forall_i, T_i(\mathrm{Z}=1) \geq T_i(\mathrm{Z}=0)$

For every individual i, the value of treatment they would take, given that they are given encouragement (*Z*=1), is greater than or equal to the value they would take if (Z=0)

$\mathrm{T}(\mathrm{Z}=1)=1,$	$T(\mathrm{Z}=0)=0$	**Compliers**: Who always take the treatment they are assigned (Z)
$\cancel{\mathrm{T}(\mathrm{Z}=1)=0,}$	$\cancel{T(\mathrm{Z}=0)=1}$	~~**Defiers**: Who always take the treatment they are NOT assigned~~
$\mathrm{T}(\mathrm{Z}=1)=1,$	$T(\mathrm{Z}=0)=1$	**Always Takers**: Who ALWAYS take the treatment irrespective of assignment
$\mathrm{T}(\mathrm{Z}=1)=0,$	$T(\mathrm{Z}=0)=0$	**Never Takers**: Who NEVER take the treatment irrespective of assignment

Duke

Content source: Brady Neal, Intro to Causal Inference

Derivation

- $\mathbb{E}[Y(Z=1) - Y(Z=0)] =$

$$\mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=1, T(0)=0)P(T(1)=1, T(0)=0)$$ Compliers

$$+ \cancel{\mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=0, T(0)=1)\ P(T(1)=0, T(0)=1)}$$ Defiers

$$+ \cancel{\mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=1, T(0)=1)P(T(1)=1, T(0)=1)}$$ Always Takers

$$+ \cancel{\mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=0, T(0)=0)]P(T(1)=0, T(0)=0)}$$ Never Takers

$$= \mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=1, T(0)=0) * P(T(1)=1, T(0)=0)$$

$$\mathbb{E}[Y(Z=1) - Y(Z=0)|T(1)=1, T(0)=0) = \frac{\mathbb{E}[Y(Z=1) - Y(Z=0)]}{P(T(1)=1, T(0)=0)}$$

Treatment effect only for compliers

Duke

Content source: Brady Neal, Intro to Causal Inference

Derivation

$\mathbb{E}[Y(Z=1) - Y(Z=0) | T(1)=1, T(0)=0)$ *can be written as*
$\mathbb{E}[Y(T=1) - Y(T=0) | T(1)=1, T(0)=0)$

because we are considering only compliers. F
or them when Z=1 => T=1 & when Z=0 => T=0

Therefore,

- $\mathbb{E}[Y(T=1) - Y(T=0) | T(1)=1, T(0)=0) = \dfrac{\mathbb{E}[Y(Z=1) - Y(Z=0)]}{P(T(1)=1, T(0)=0)}$

$\downarrow$

Local Average Treatment (LATE) or
Complier Average Causal Effect (CACE)

Duke

Content source: Brady Neal, Intro to Causal Inference

Derivation

$\mathbb{E}[Y(Z=1) - Y(Z=0) | T(1) = 1, T(0) = 0)$ *can be written as*
$\mathbb{E}[Y(T=1) - Y(T=0) | T(1) = 1, T(0) = 0)$

because we are considering only compliers. F
or them when Z=1 => T=1 & when Z=0 => T=0

Changed associational difference because of assumption of IV unconfoundedness

Therefore,

- $\mathbb{E}[Y(T=1) - Y(T=0) | T(1) = 1, T(0) = 0)$ = $\dfrac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{P[T(1)=1, T(0)=0)]}$

Local Average Treatment (LATE) or
Complier Average Causal Effect (CACE)

Probability of being a complier

Content source: Brady Neal, Intro to Causal Inference

Derivation

$$\mathbb{E}[Y(Z=1)-Y(Z=0)|T(1)=1\,,T(0)=0)\ \textit{can be written as}$$
$$\mathbb{E}[Y(T=1)-Y(T=0)|T(1)=1\,,T(0)=0)$$

because we are considering only compliers. F
or them when Z=1 => T=1 & when Z=0 => T =0

Therefore,

- $\mathbb{E}[Y(T=1)-Y(T=0)|T(1)=1\,,T(0)=0)$ = $\dfrac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[Y|T=1]-\mathbb{E}[Y|T=0]}$

Local Average Treatment (LATE) or
Complier Average Causal Effect (CACE)

You can quantify this as the probability of (1- non compliers) & then use the monotonicity assumption to quantify from observational data

Duke

Content source: Brady Neal, Intro to Causal Inference

Derivation

$$\underbrace{\mathbb{E}[Y(T=1) - Y(T=0) \mid T(1)=1, T(0)=0)}_{\text{Local Average Treatment (LATE)}} = \underbrace{\frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]}}_{\text{Wald's Estimand}}$$

Limitations

- Condition of monotonicity is not always satisfied
- We get estimates only for compliers, we might be interested in the broader group
- We might not be sure who the compliers are – then we cannot be sure of who the LATE affects

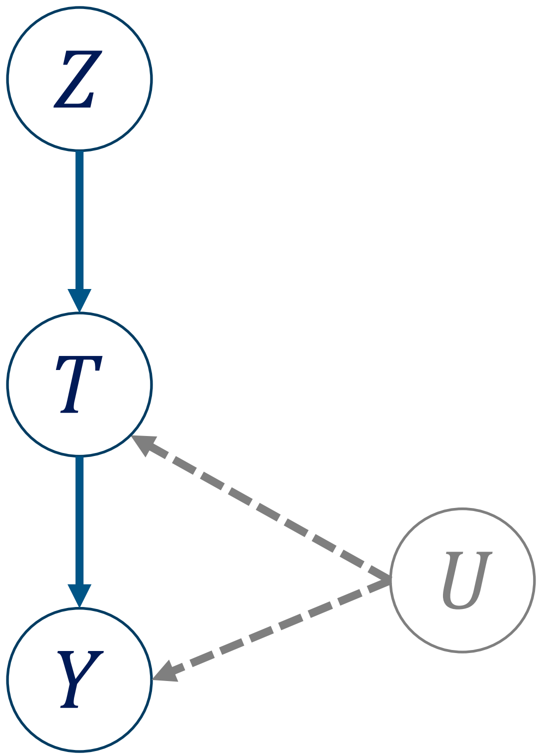
Duke

Syrgkanis et al. (2019)

"Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments," *Advances in Neural Information Processing Systems*, 32.

Duke

Overview

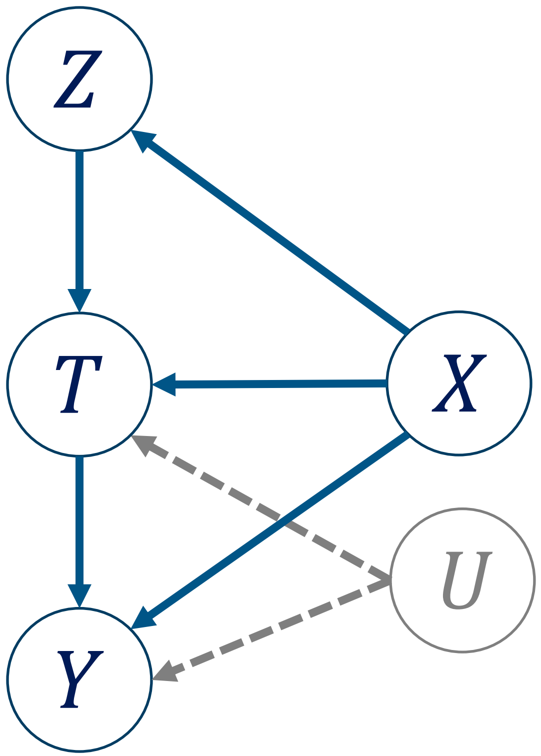


Simple IV estimation is like…

- Z: instrument
- T: treatment
- Y: outcome
- U: confounding variables (unobserved)

Duke

Overview

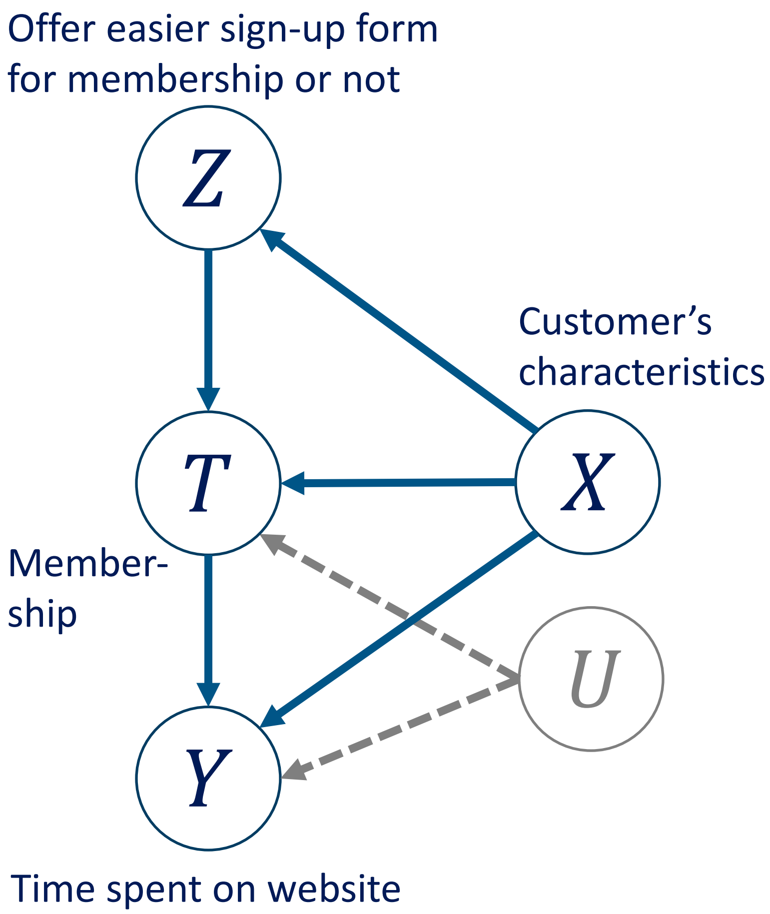


In this paper…

- Z: instrument
- T: treatment
- Y: outcome
- U: confounding variables (unobserved)
- X: confounding variables (observed) can affect Z, T, Y, and treatment effect

Duke

Overview



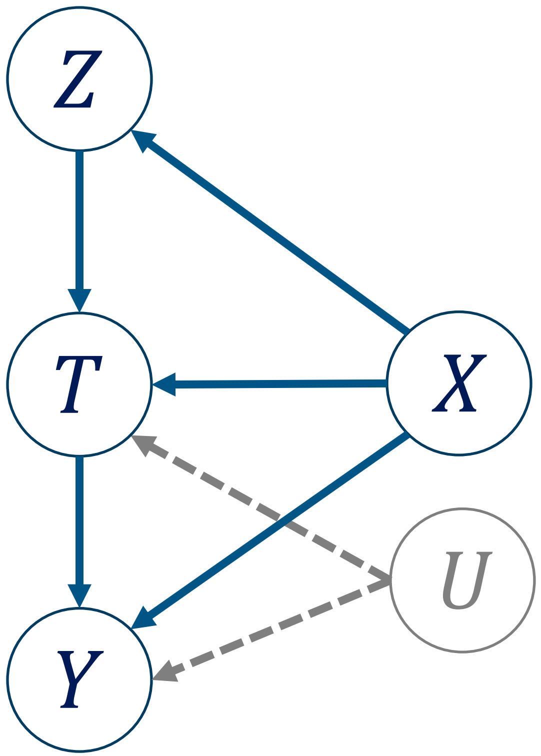
- Want to estimate CATE = $E[Y_1 - Y_0 | X]$.
  e.g. If we want our customers to spend more time on our website, what kind of customers should we approach?
- How can we deal with X?
- How can we deal with U?
- Want to use flexible models.
- But we're afraid of estimation errors.
- Want to interpret the estimation.

Duke

Tripadvisor

Image source: https://www.tripadvisor.com/

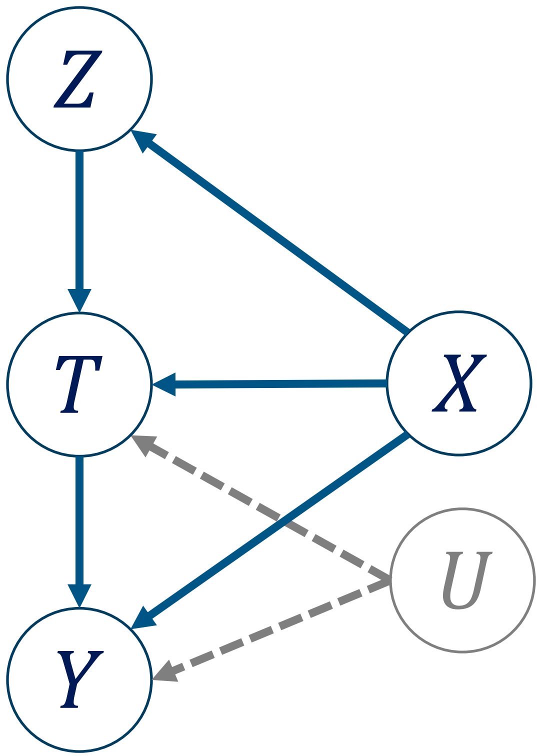
Overview



- Want to estimate CATE = $E[Y_1 - Y_0|X]$.
  → DRIV (proposed in this paper): general algorithm to estimate CATE using IV.
- How can we deal with X?
  → Condition on X (block the path Z- X- Y).
- How can we deal with U?
  → IV estimation.
- Want to use flexible models.
  → DRIV can accommodate ML.
- But we're afraid of estimation errors.
  → DRIV is doubly robust.
- Want to interpret the estimation.
  → DRIV can incorporate interpretable models.

Duke

How it works

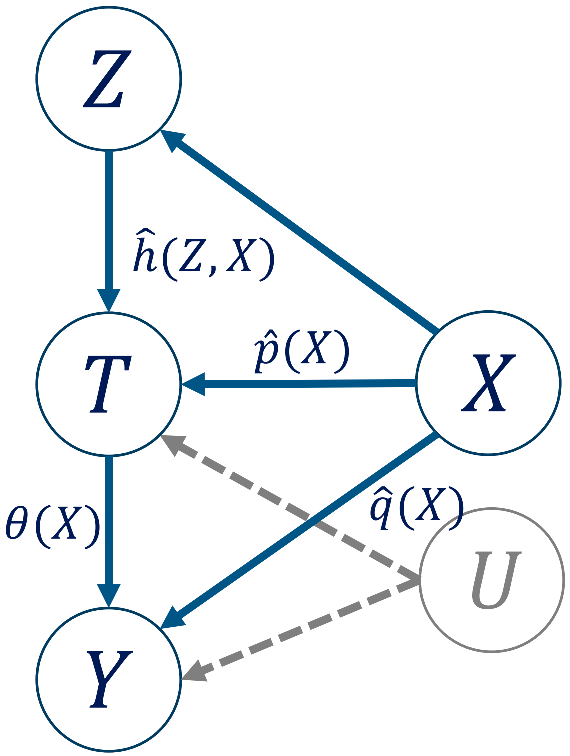


DRIV: two-step optimization

1. Make a preliminary estimate for CATE
2. Make it more robust to estimation errors

Duke

First step

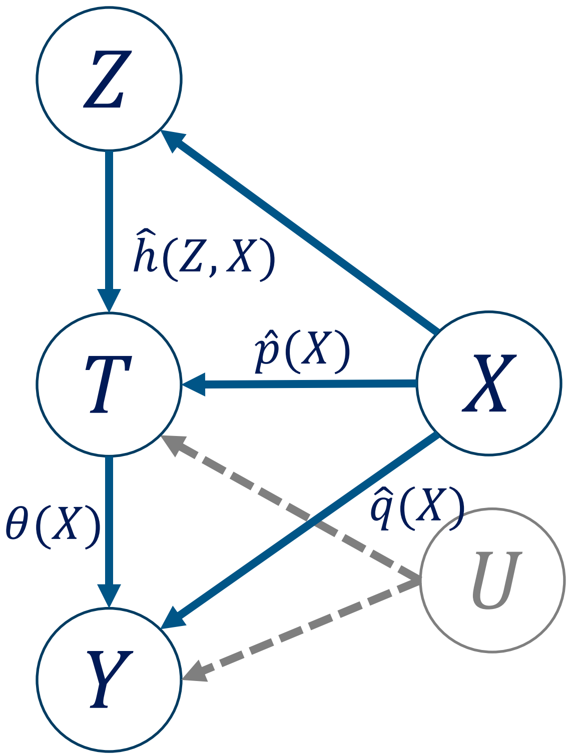


- Make a preliminary estimate for CATE θ_{pre} by minimizing the loss:

$$L^1(\theta) = E\left[\left(Y - \hat{q}(X) - \theta(X)\{\hat{h}(Z,X) - \hat{p}(X)\}\right)^2\right]$$

Duke

First step



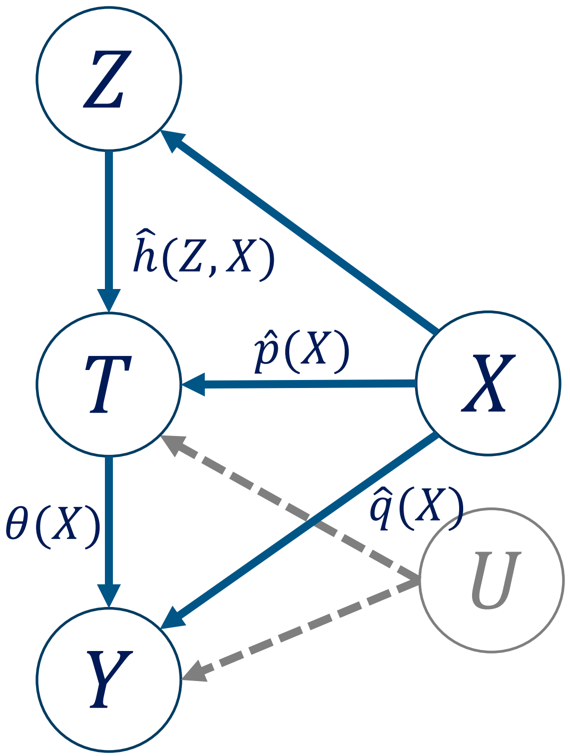
- Make a preliminary estimate for CATE θ_{pre} by minimizing the loss:

$$L^1(\theta) = E\left[\left(Y - \underbrace{\hat{q}(X)}_{\hat{E}[Y|X]} - \theta(X)\{\underbrace{\hat{h}(Z,X)}_{\hat{E}[T|Z,X]} - \underbrace{\hat{p}(X)}_{\hat{E}[T|X]}\}\right)^2\right]$$

- You can use ML to estimate these conditional means.

Duke

First step (cont'd)



Where did $L^1(\theta)$ come from?

$\rightarrow$ Moment condition:

$$E[e|Z,X] = 0$$

Suppose the true model is:

$$Y = \theta_0(X)T + f_0(X) + e$$

Let

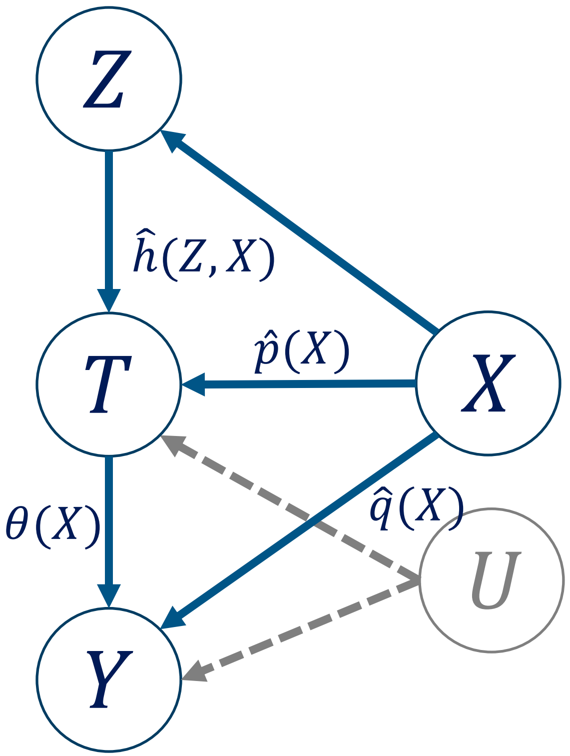
$$h_0(Z,X) = E[T|Z,X]$$

…

$\theta(X)$ satisfying the moment condition is equivalent to the minimizer of $L^1(\theta)$

Duke

First step (cont'd)



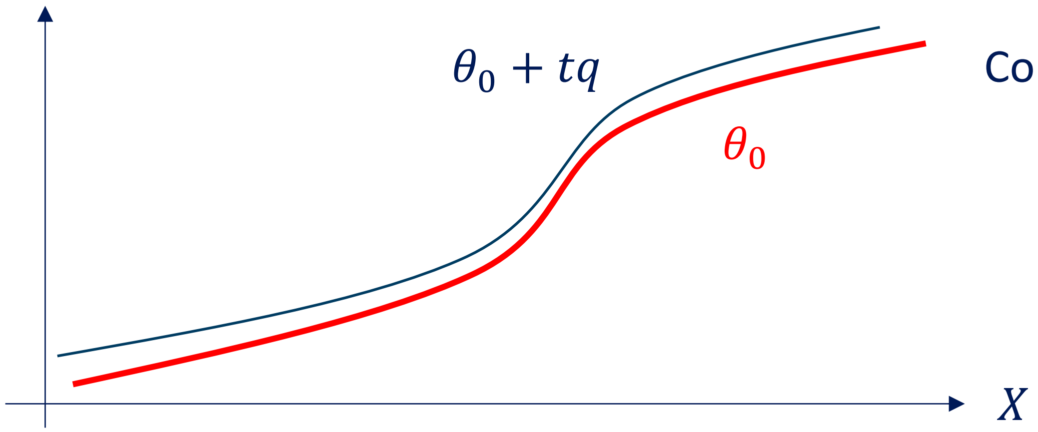
$$L^1(\theta) = E\left[\left(Y - \hat{q}(X) - \theta(X)\{\hat{h}(Z,X) - \hat{p}(X)\}\right)^2\right]$$

Pros:

- Robust to estimation errors in $\hat{q}(X)$ and $\hat{p}(X)$
- Easy to minimize because of convexity

Cons:

- NOT robust to estimation errors in $\hat{h}(Z,X)$ → 2nd step

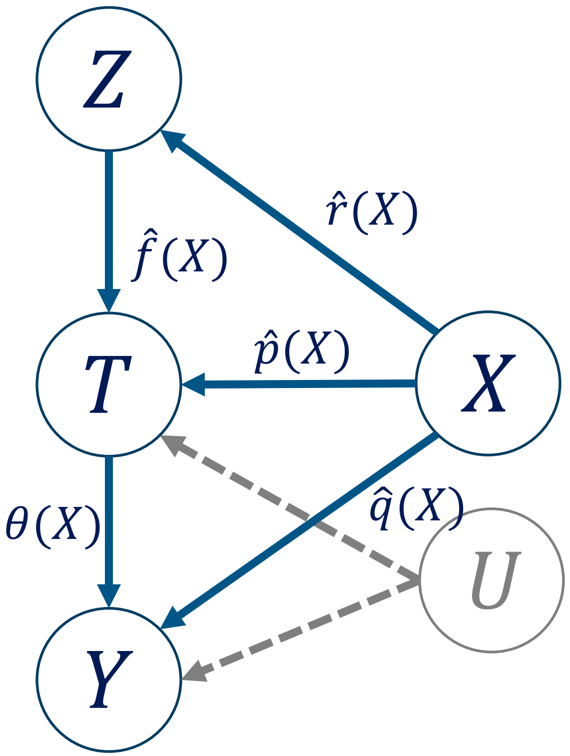


Concept of Neyman orthogonality:

$$\left.\frac{\partial}{\partial t} L^1(\theta_0 + tq)\right|_{t=0} = 0$$

Duke

Second step

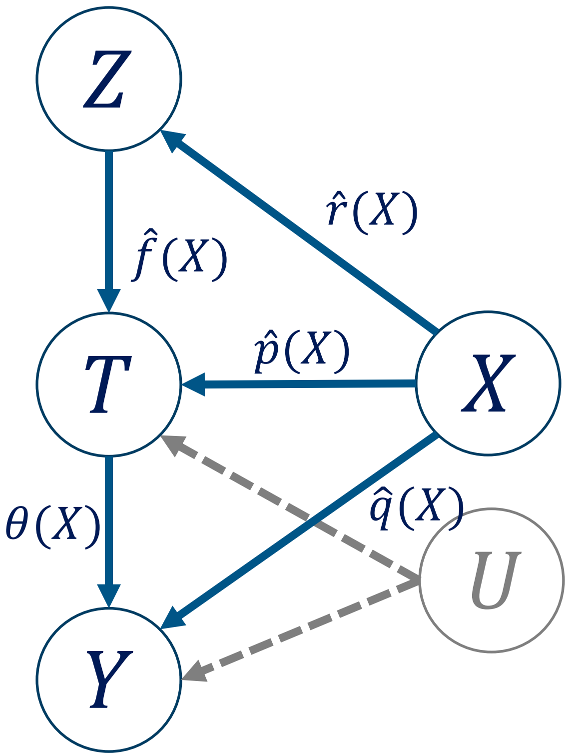


- Make an estimate for CATE θ by minimizing the loss:

$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Duke

Second step

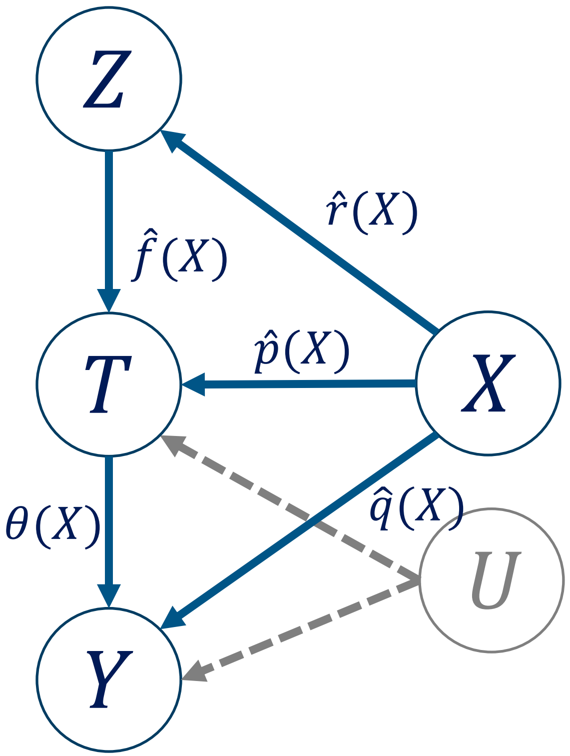


- Make an estimate for CATE θ by minimizing the loss:

$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Preliminary estimate from step 1

Second step



- Make an estimate for CATE θ by minimizing the loss:

$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Preliminary estimate from step 1

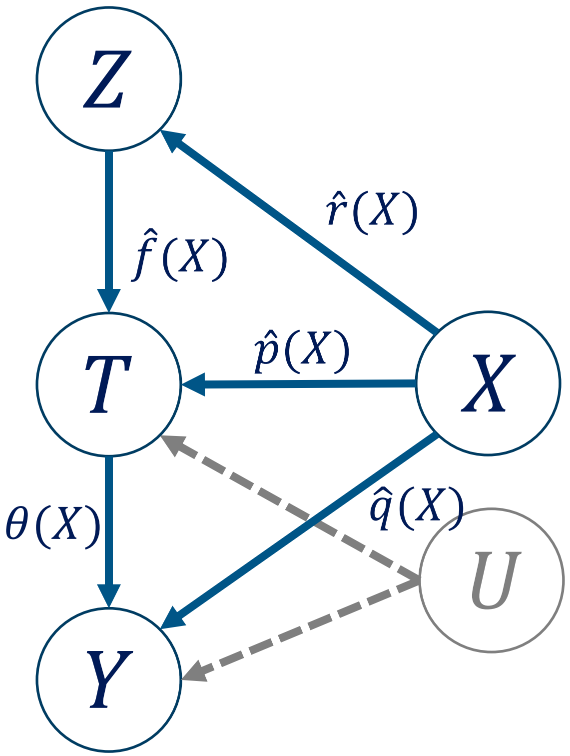
$Y - \hat{E}[Y|X]$

$T - \hat{E}[T|X]$

$Z - \hat{E}[Z|X]$

Duke

Second step



- Make an estimate for CATE θ by minimizing the loss:

$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Preliminary estimate from step 1

$Y - \hat{E}[Y|X]$

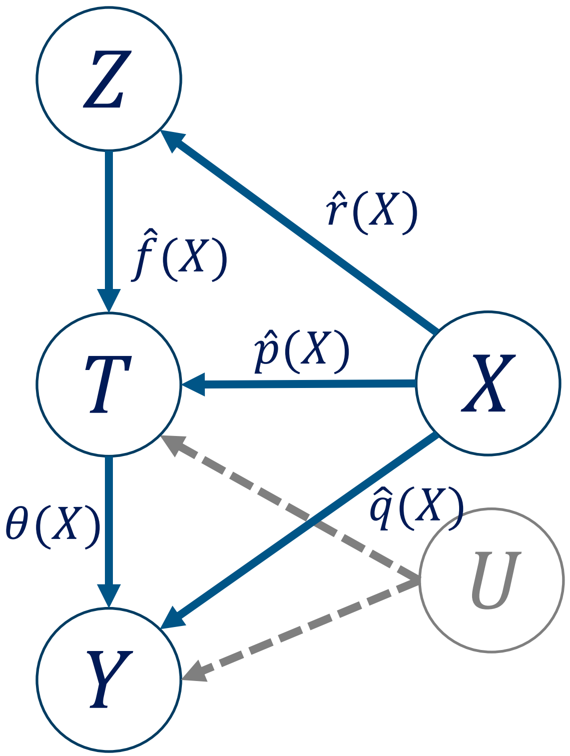
$T - \hat{E}[T|X]$

$Z - \hat{E}[Z|X]$

$\hat{E}[TZ|X] - \hat{E}[T|X]\hat{E}[Z|X]$

Duke

Second step



- Make an estimate for CATE θ by minimizing the loss:

$$L^2(\theta) = E\left[\left(\boxed{\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)}} - \theta(X)\right)^2\right]$$

Doubly robust estimator

Duke

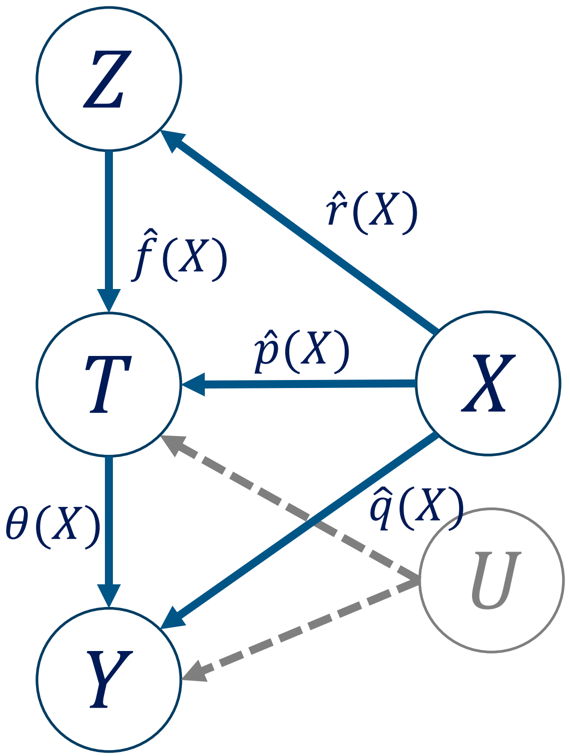
Doubly robust approach

- AIPW = Augmented Inverse-Propensity Weighting
- Combine two estimators:
    1. Inverse-Propensity Weighting: $E\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]$
    2. Regression-based: $E[Y|T=1,X] - E[Y|T=0,X]$
- Consistent if either $e(X)$ or $E[Y|T=t,X]$ is correct $\Longrightarrow$ Doubly robust

- $L^2(\theta) = E\left[\left(\boxed{\underbrace{\theta_{pre}}_{\text{Regression}} + \underbrace{\frac{(\tilde{Y}-\theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)}}_{\text{IPW}}} - \theta(X)\right)^2\right]$: robust to error in $\hat{\beta}(X)$

Duke

Second step (cont'd)



$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Pros:

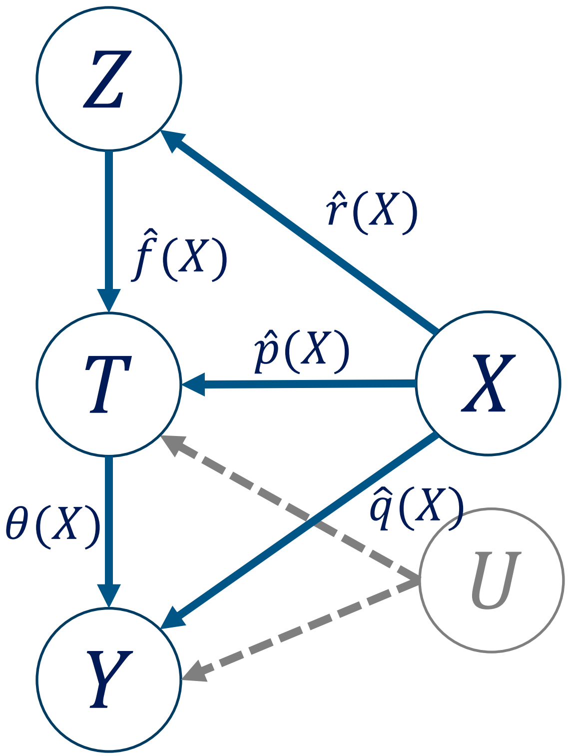
- Robust to estimation errors in $\hat{p}(X)$, $\hat{q}(X)$, $\hat{r}(X)$, $\hat{\beta}(X)$, and θ_{pre}
- Easy to minimize because of convexity
- Enables interpretable $\theta(X)$

Cons:

- Second order impact from θ_{pre}?

Duke

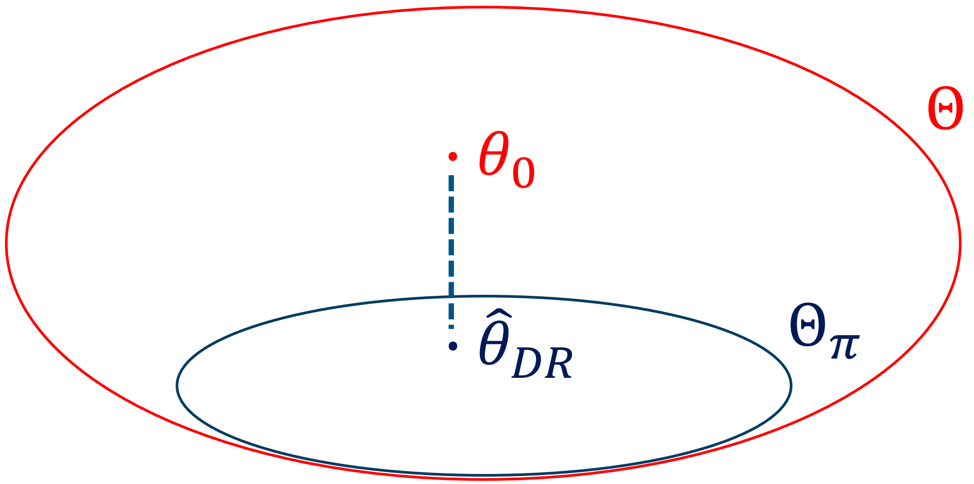
Second step (cont'd)



$$L^2(\theta) = E\left[\left(\theta_{pre} + \frac{(\tilde{Y} - \theta_{pre}\tilde{T})\tilde{Z}}{\hat{\beta}(X)} - \theta(X)\right)^2\right]$$

Why does this enable interpretable $\theta(X)$?

$\rightarrow$ We can choose hypothesis space Θ_π



Duke

Takeaways

- Proposed approach = DRIV (Doubly Robust IV?)
- Eliminate bias through IV estimation
- Utilize power of machine learning
- Doubly robust approach to fight against estimation errors
- Can produce interpretable results
- Two-step optimization
  - 1st step: preliminary estimate for CATE
  - 2nd step: make it more robust and interpretable

Duke

Demonstration

- Synthetic data:
  - DRIV correctly estimated ATE and CATE
  - Estimate by DMLATEIV (Chernozhukov et al. [2018]) is more biased

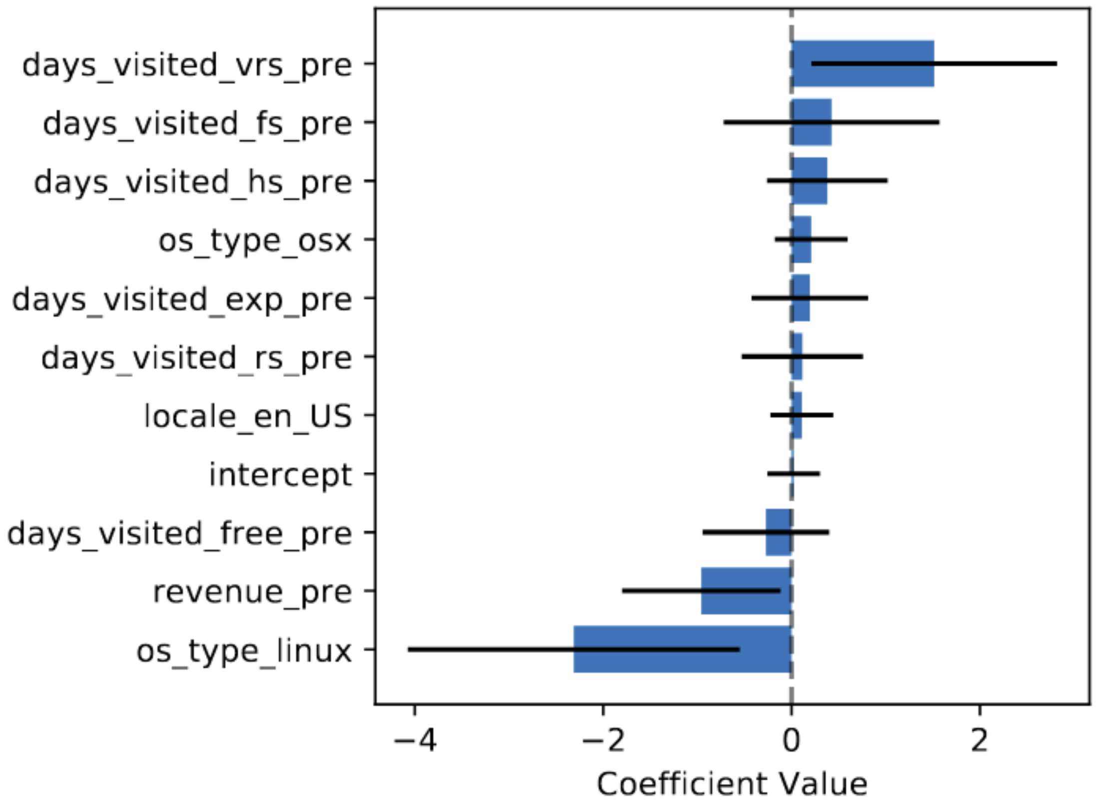
		Observational Data		Semi-Synthetic Data		
Nuisance	Method	ATE Est	95% CI	ATE Est	95% CI	Cover ‡
LM	DMLATEIV	0.137	[0.027, 0.248]	0.654	[0.621, 0.687]	10%
LM	DRIV	0.065	[-0.02, 0.151]	0.587	[0.521, 0.652]†	92%

† Contains the true ATE (0.609) ‡ Coverage for 95% CI over 100 Monte Carlo simulations

Table 2: NLSYM ATE Estimates for Observational and Semi-synthetic Data

Duke

Demonstration (cont'd)



From Figure 1 of Syrgkanis et al. (2019)

- TripAdvisor data
  - Z: A/B test assignment for membership sign-up process
  - T: becoming a member
  - Y: # of days a user visits TripAdvisor
  - X: 28-day pre-experiment summary about browsing and purchasing activity
- Θ_π: linear functions
- Implication:
  - More approach to users with high "days_visited_vrs_pre"
  - Improve approach to users with high "revenue_pre"

Image source: https://www.tripadvisor.com/

Duke

References

- Angrist, Joshua, and Guido Imbens. 1991. “Identification and Estimation of Local Average Treatment Effects.” NBER Technical Working Paper Series 118.
- Angrist, Joshua. 2022. “Introduction to Instrumental Variables, Part One.” Marginal Revolution University. Retrieved (https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one).
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.
- Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, *18*(1), 36-56.
- Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., & Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments. Advances in Neural Information Processing Systems, 32.
- Wager, Stefan, (2022). Average Treatment Effects. Double Robustness. YouTube. (https://www.youtube.com/watch?v=IfZHUFFlsGc)
- Neal, B (2020). Introduction to Causal Inference. https://www.bradyneal.com/causal-inference-course

Duke