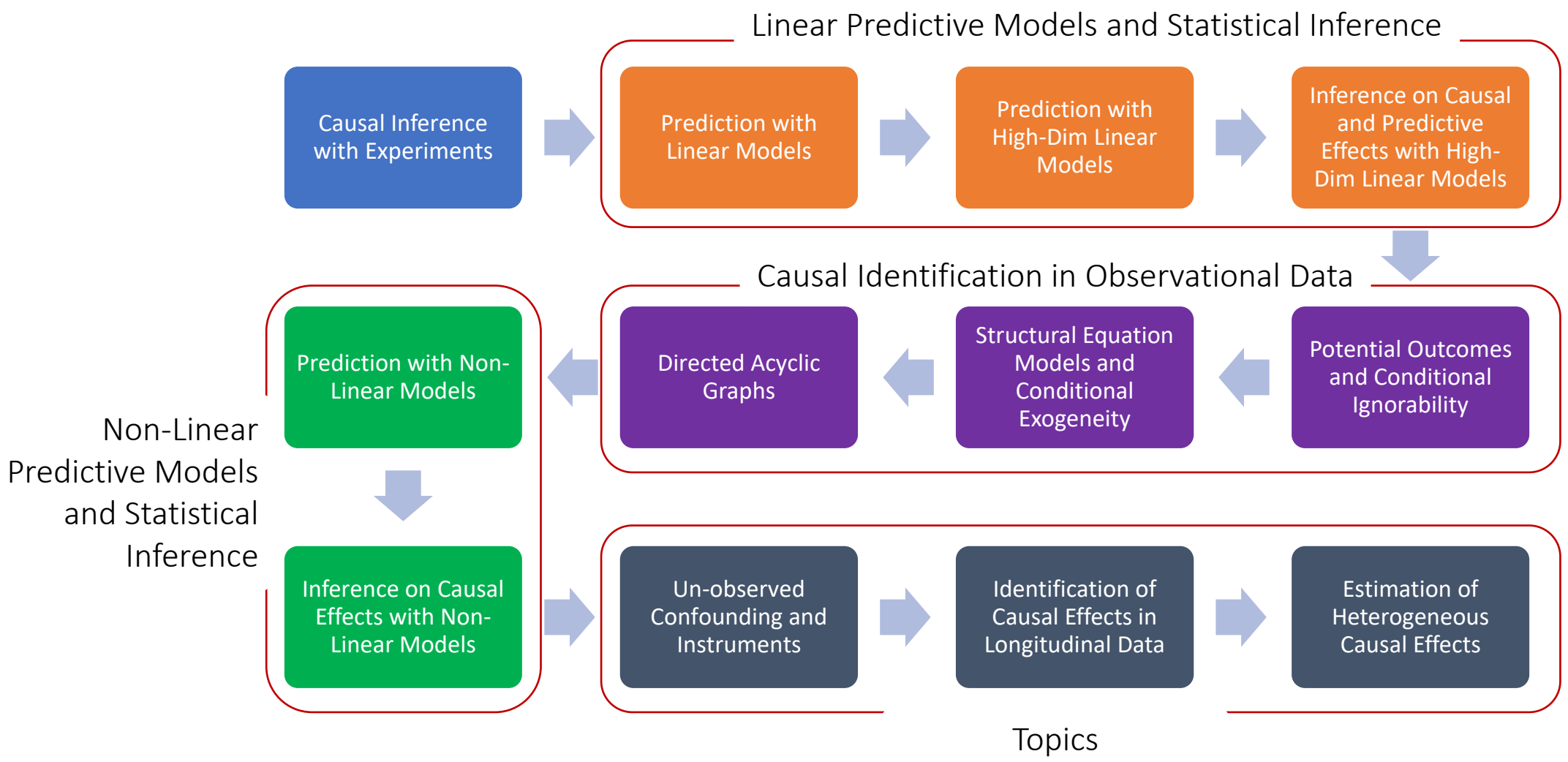
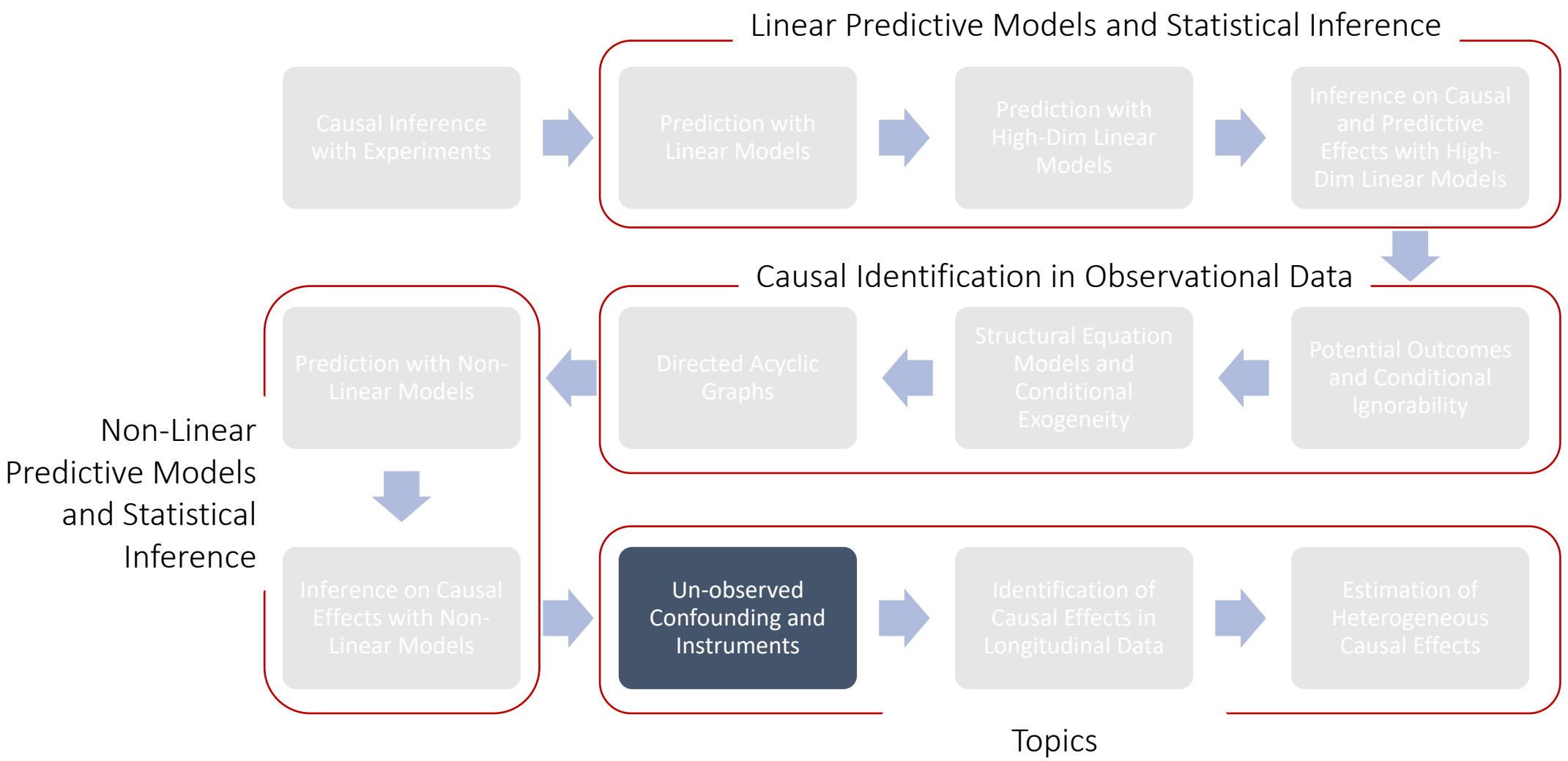
MS&E 228: Unobserved Confounding

Vasilis Syrgkanis

MS&E, Stanford


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Prediction with Non-Linear Models
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics

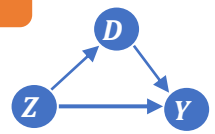
Goals for Today

- What can we do when we have un-observed confounding
- Omitted variable bias bounds
- Introduction to “Instruments”

Causal Inference Pipeline

Theory

DAGs/ASEMs



$$a_d = E[E[Y|D=d,Z]]$$

Construct estimate $\hat{g}(d,z)$ of predictive model

$$E[Y \mid D=d, Z=z]$$

Estimate mean counterfactual

$$\hat{a}_d = E_n[\hat{g}(d,Z)]$$

$$a_d \in \hat{a}_d \pm 2\hat{\sigma}$$

Practice

Data Collection → Domain Assumption Elicitation → Identification → Estimation (training) → Validation (testing) → Inference (Confidence Intervals) → Sensitivity Analysis → Decision Policy → Adaptive Experiments → Data Collection

What if my conditional ignorability assumption was violated? How much would the estimate change?

$$a_d \in \hat{a}_d \pm (2\hat{\sigma} + \epsilon)$$

Bias Bounds

Reduction in unexplained variance of Y when adding A in the model that predicts Y from treatment and controls

- The analyst provides bounds on the partial R^2

$$R^2_{Y\sim A|D,X} \leq C^2_Y, \qquad R^2_{D\sim A|X} \leq C^2_D$$

Reduction in unexplained variance of D when adding A in the model that predicts D from controls

- Based on these bounds we can conclude that

$$\theta_0 \in \theta_s \pm \sqrt{C^2_Y \frac{C^2_D}{1-C_D} \frac{E\left[\left(\tilde{Y}-\theta_s\tilde{D}\right)^2\right]}{E\left[\tilde{D}^2\right]}}$$

Measurable from the data

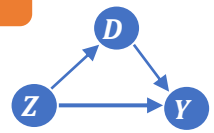
For more details:
[Making Sense of Sensitivity: Extending Omitted Variable Bias]

For more general analysis see:
[Long Story Short: Omitted Variable Bias in Causal Machine Learning]

Causal Inference Pipeline

Theory

DAGs/ASEMs



$$a_d = E[E[Y|D=d,Z]]$$

Construct estimate $\hat{g}(d,z)$ of predictive model

$$E[Y \mid D=d, Z=z]$$

Estimate mean counterfactual

$$\hat{a}_d = E_n[\hat{g}(d,Z)]$$

$$a_d \in \hat{a}_d \pm 2\hat{\sigma}$$

Practice

Data Collection → Domain Assumption Elicitation → Identification → Estimation (training) → Validation (testing) → Inference (Confidence Intervals) → Sensitivity Analysis → Decision Policy → Adaptive Experiments → Data Collection

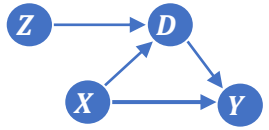
Are there other ways of identifying the treatment effect I care about (other than identification by conditioning) when I have un-observed confounding?

$$a_d \in \hat{a}_d \pm (2\hat{\sigma} + \epsilon)$$

Causal Inference Pipeline

Theory

DAGs/ASEMs



$$\alpha = \frac{E[\tilde{Y}\tilde{Z}]}{E[\tilde{D}\tilde{Z}]}$$

Estimate approximate residuals $\check{Y}, \check{Z}, \check{D}$ using cross-fitting

Solve moment restriction

$$E_n[(\check{Y} - \hat{a}\check{D})\check{Z}] = 0$$

$$a \in \hat{a} \pm 2\hat{\sigma}$$

Practice

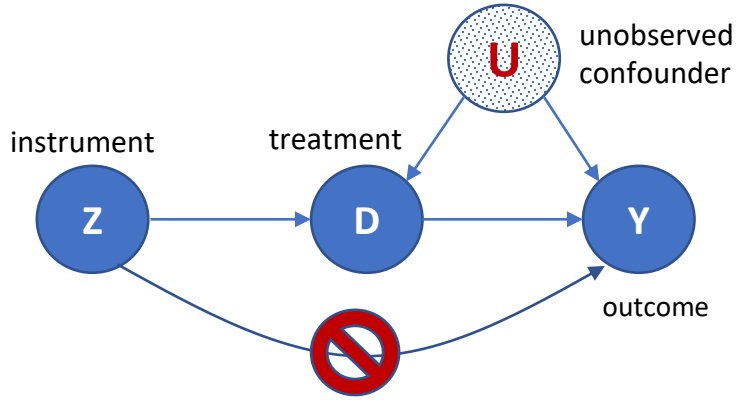
Data Collection → Domain Assumption Elicitation → Identification → Estimation (training) → Validation (testing) → Inference (Confidence Intervals) → Sensitivity Analysis → Decision Policy → Adaptive Experiments → Data Collection

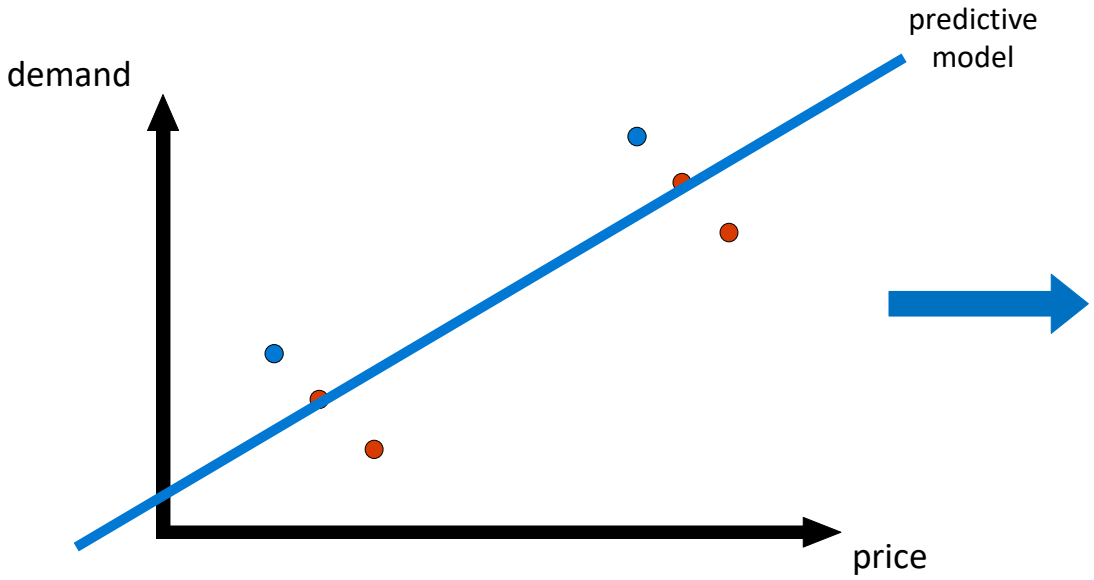
Instrumental Variable: any random variable Z that affects the treatment (log-price) D but does not affect the outcome (log-demand) Y other than through the treatment

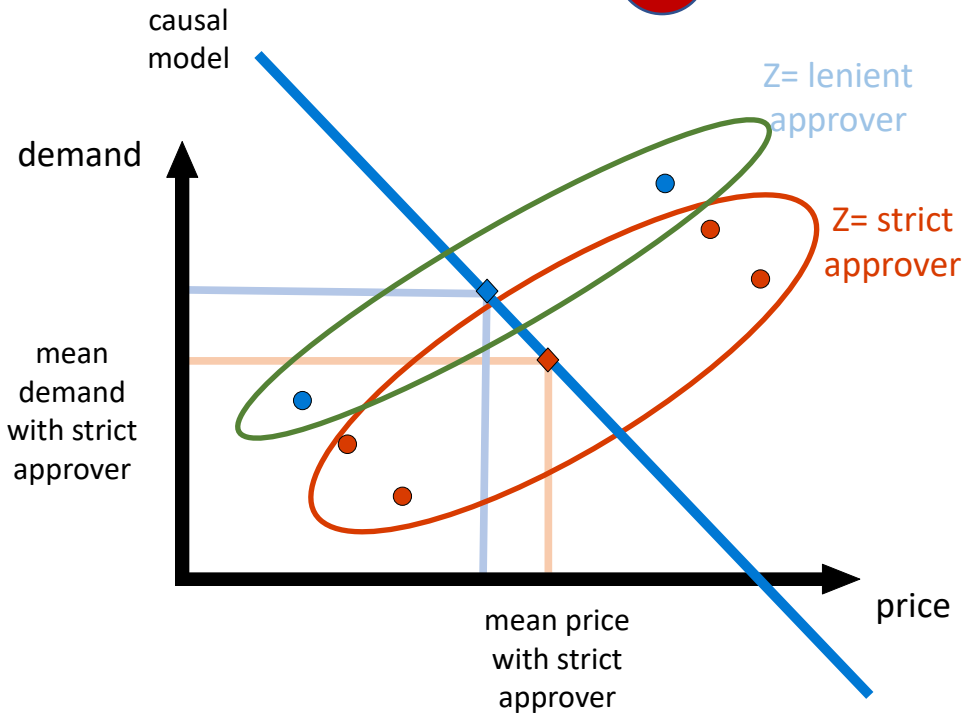
$$a \in \hat{a} \pm (2\hat{\sigma} + \epsilon)$$

Instrumental Variables and 2SLS

Instrumental Variable: any random variable Z that affects the treatment (log-price) D but does not affect the outcome (log-demand) Y other than through the treatment [Wright'28, Bowden-Turkington'90, Angrist-Krueger'91, Imbens-Angrist'94]







Partially Linear Instrumental Variable Model

- After partialling out the observed controls X

$$\tilde{Y} \coloneqq \theta_0 \tilde{D} + \tilde{A} + \epsilon_Y$$
$$\tilde{D} \coloneqq \beta \tilde{Z} + \gamma \tilde{A} + \epsilon_D$$
$$\tilde{Z} \coloneqq \epsilon_Z$$
$$\tilde{A} \coloneqq \epsilon_A$$

- Setting falls into the general moment estimation framework

$$M(\theta, h, p, m) = E\left[\Big(Y - h(X) - \theta\left(D - p(X)\right)\Big)\ \left(Z - m(X)\right)\right] = 0$$

- Where $h(X) = E[Y|X]$, $p(X) = E[D|X]$, $m(Z) = E[Z|X]$

Orthogonal Method: Double ML for IV

Double ML. Split samples in half

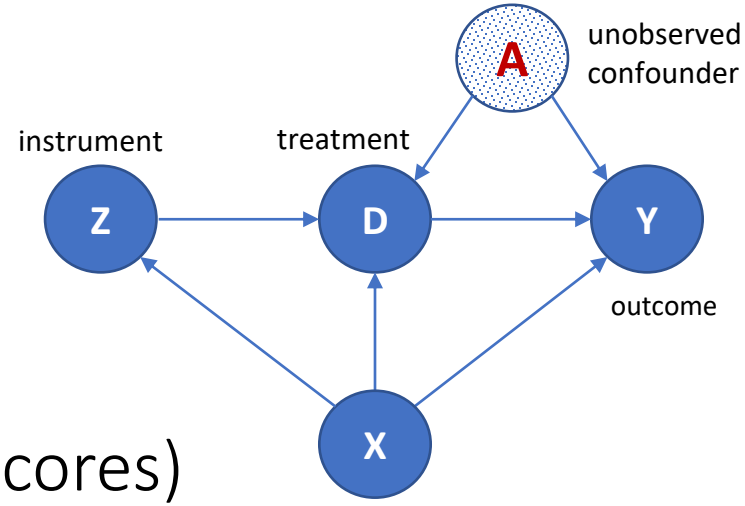
- Regress $Y \sim X$ with ML on first half, to get estimate $\hat{h}(S)$ of $E[Y|X]$
- Regress $D \sim X$ with ML on first half, to get estimate $\hat{p}(S)$ of $E[D|X]$
- Regress $Z \sim X$ with ML on first half, to get estimate $\hat{m}(S)$ of $E[Z|X]$
- Construct residuals on other half, $\hat{Z} = Z - \hat{m}(X), \hat{D} \coloneqq D - \hat{p}(X)$ and $\hat{Y} \coloneqq Y - \hat{h}(X)$
- Solve moment condition:

$$E_n\left[\left(\hat{Y} - \theta\hat{D}\right)\hat{D}\right] = 0$$

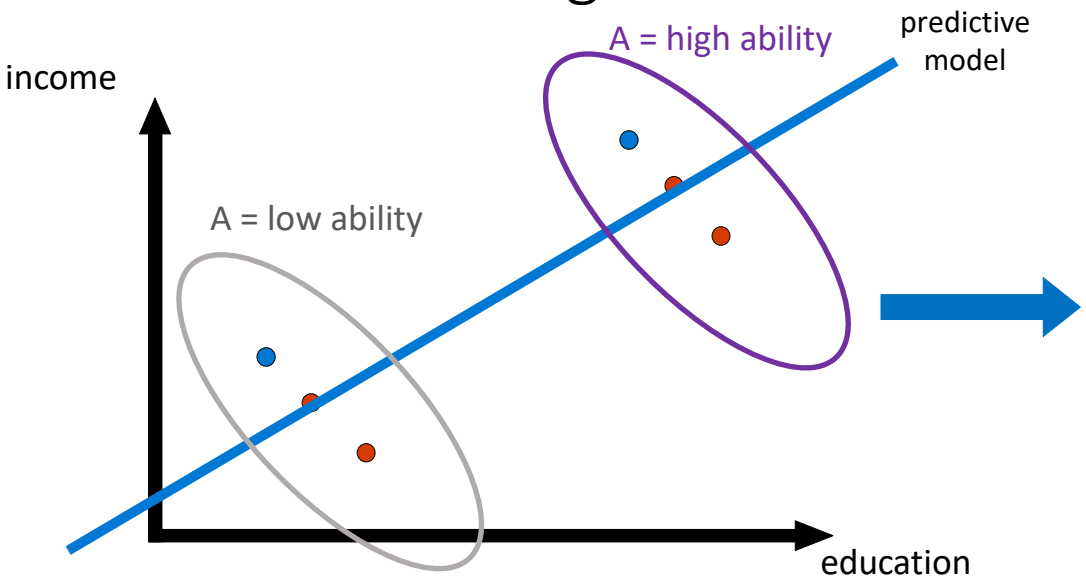
```
from econml.iv.dml import OrthoIV
orthoiv = OrthoIV()
orthoiv.fit(y, D, Z, W=X).effect_inference()
```

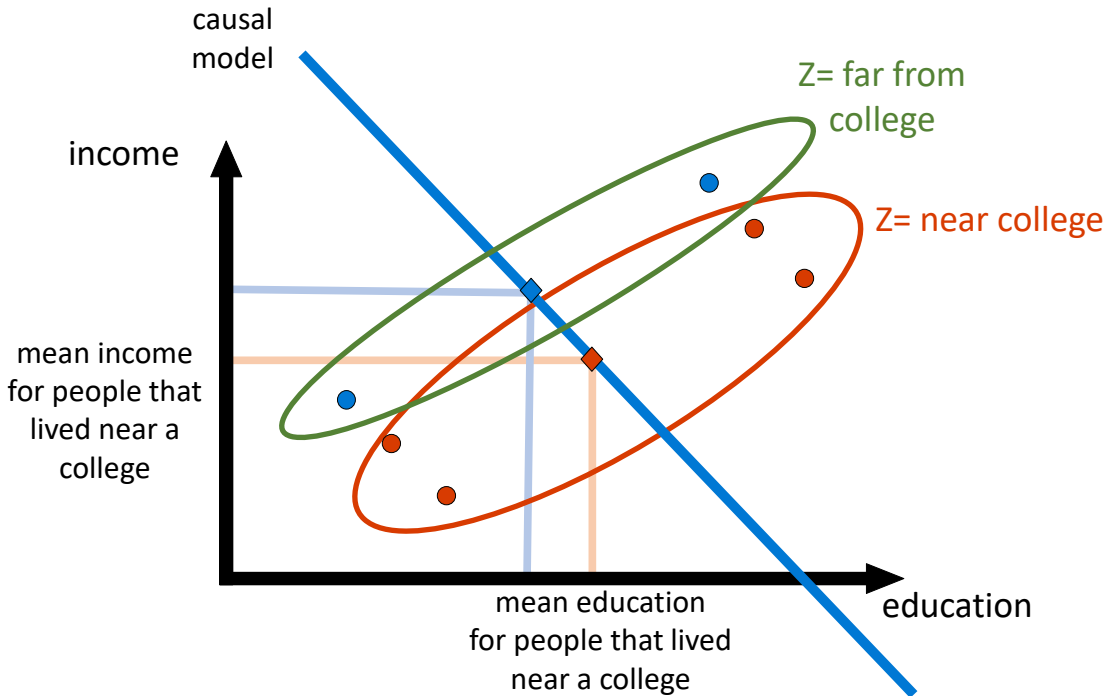
Examples

Returns to Education

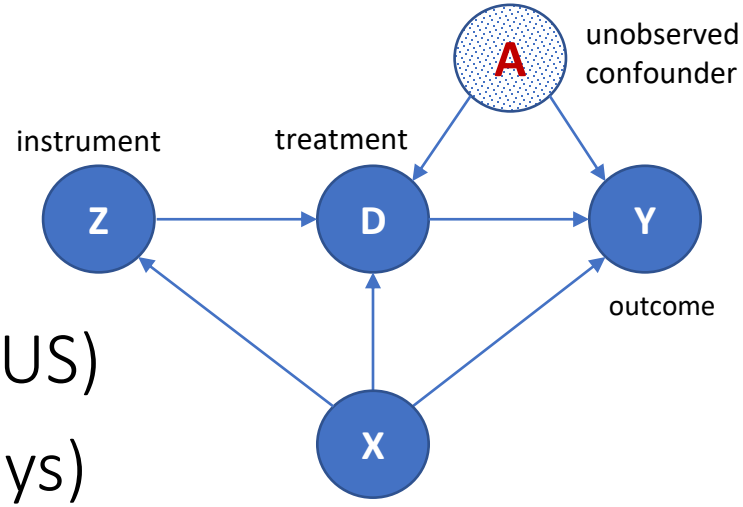


- D: years of college, Y: income
- X: observable characteristics of a student (e.g. test scores)
- A: unobserved “ability”
- Z: distance to college

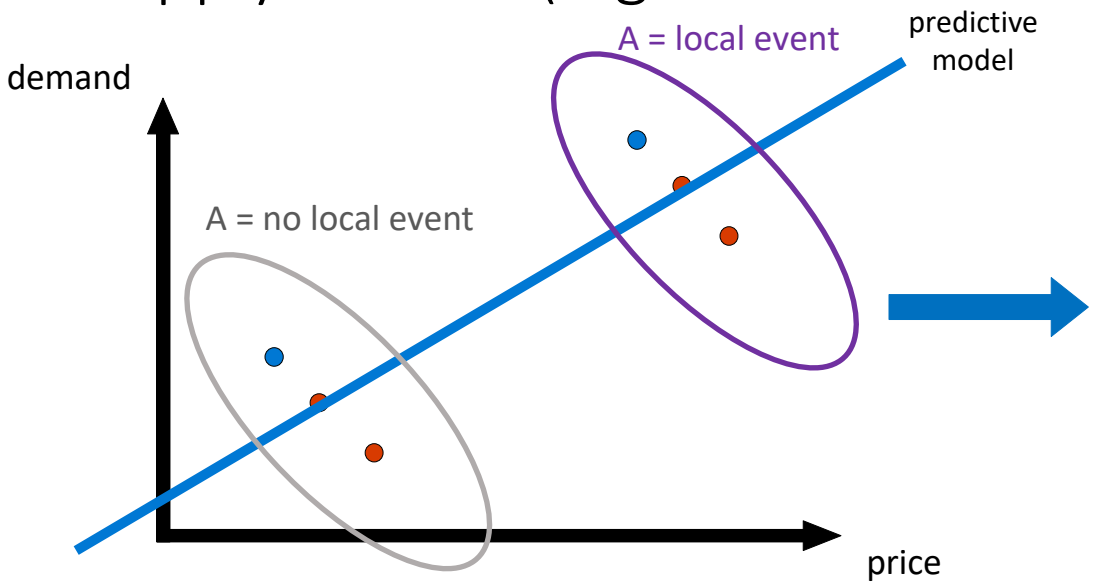


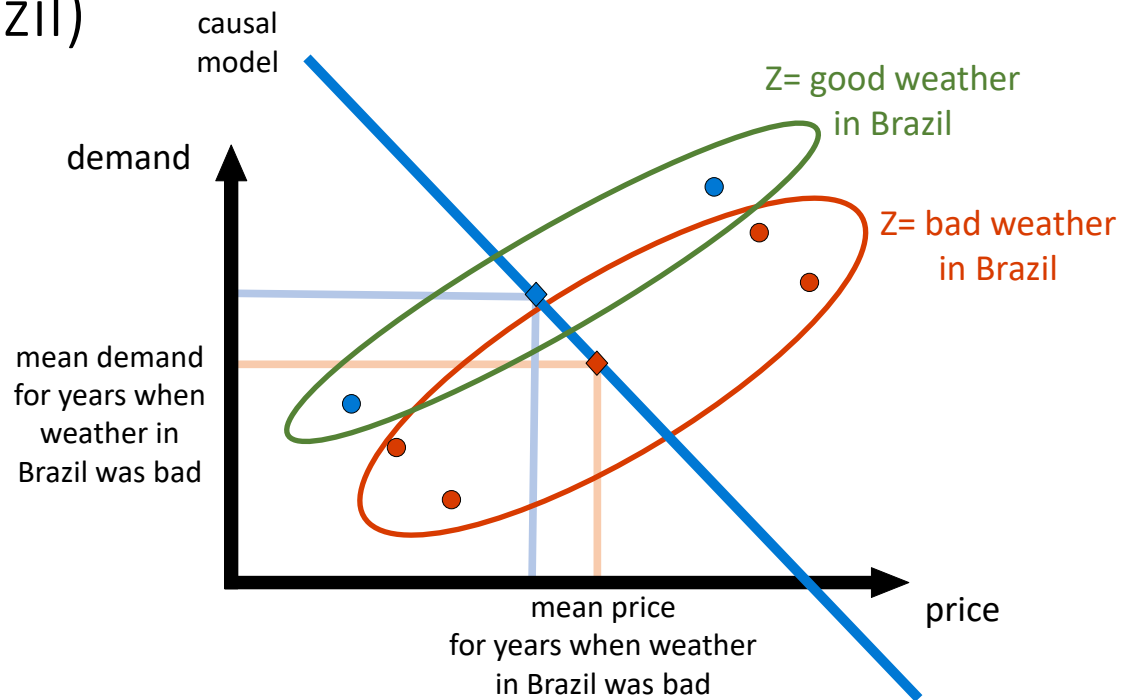


Demand Estimation

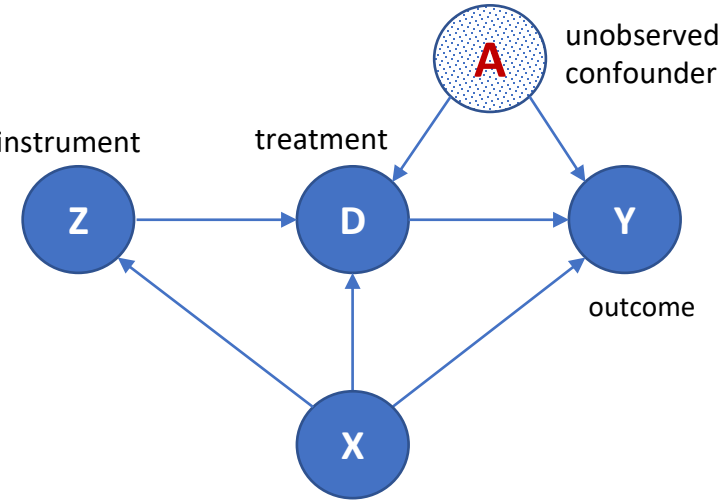


- D: price (e.g. of coffee), Y: demand (e.g. of coffee in US)
- X: observable characteristics of a market (e.g. holidays)
- A: unobserved “demand shocks” (e.g. local event)
- Z: supply shifters (e.g. weather in brazil)

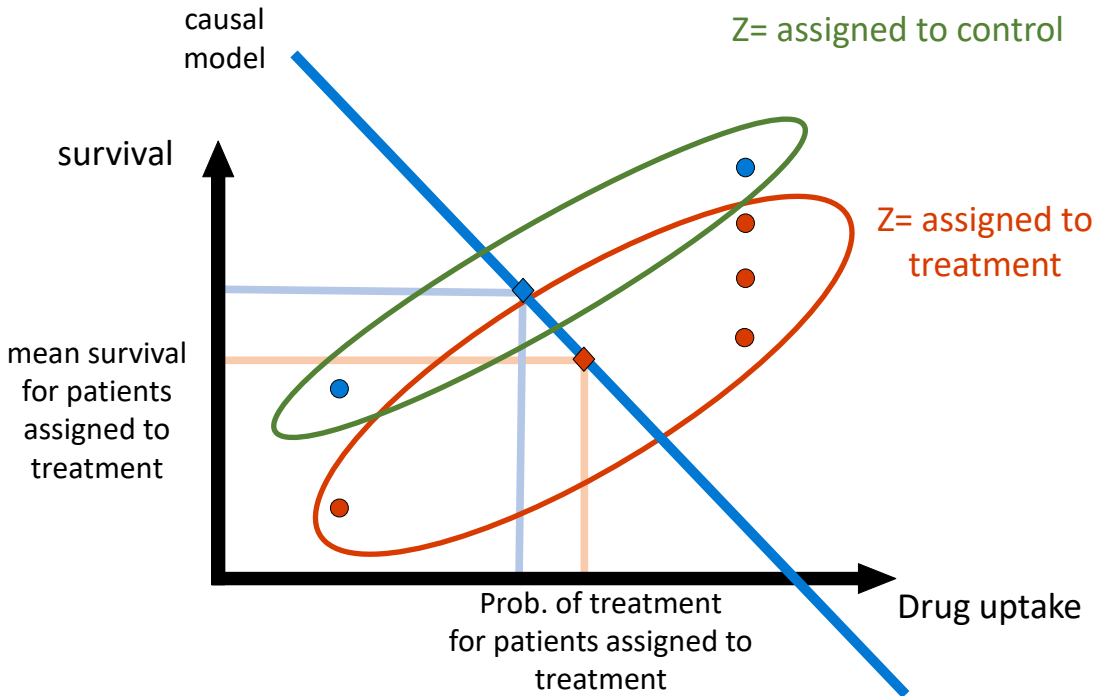




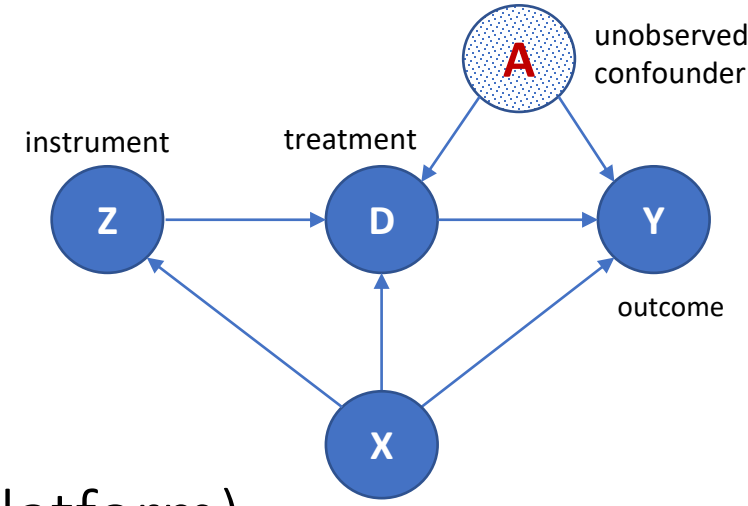
Clinical Trials with Non-Compliance



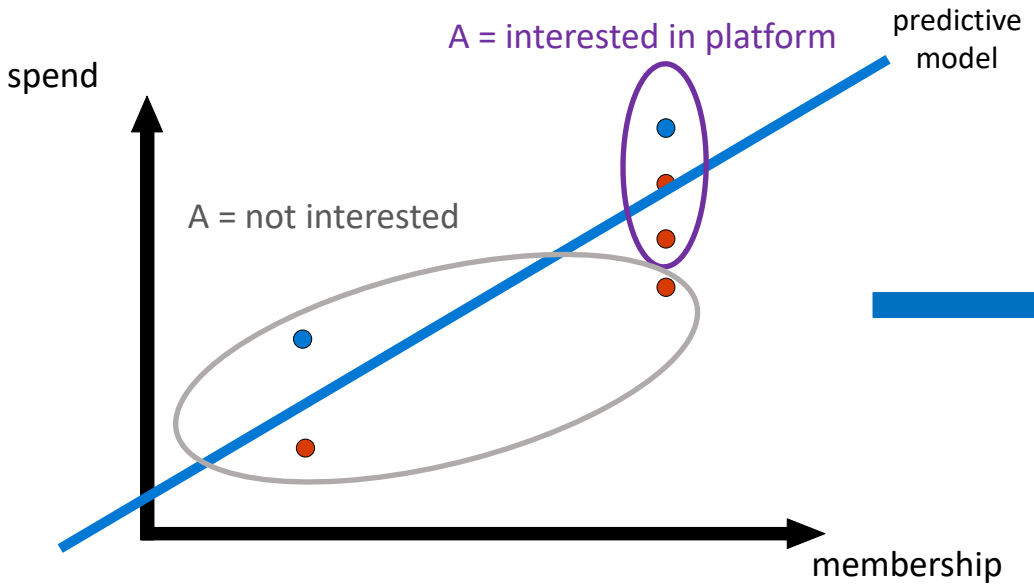
- D: drug treatment, Y: survival
- X: observable characteristics of a patient
- A: unobserved “compliance factors” (e.g. health habits)
- Z: randomized cohort assignment

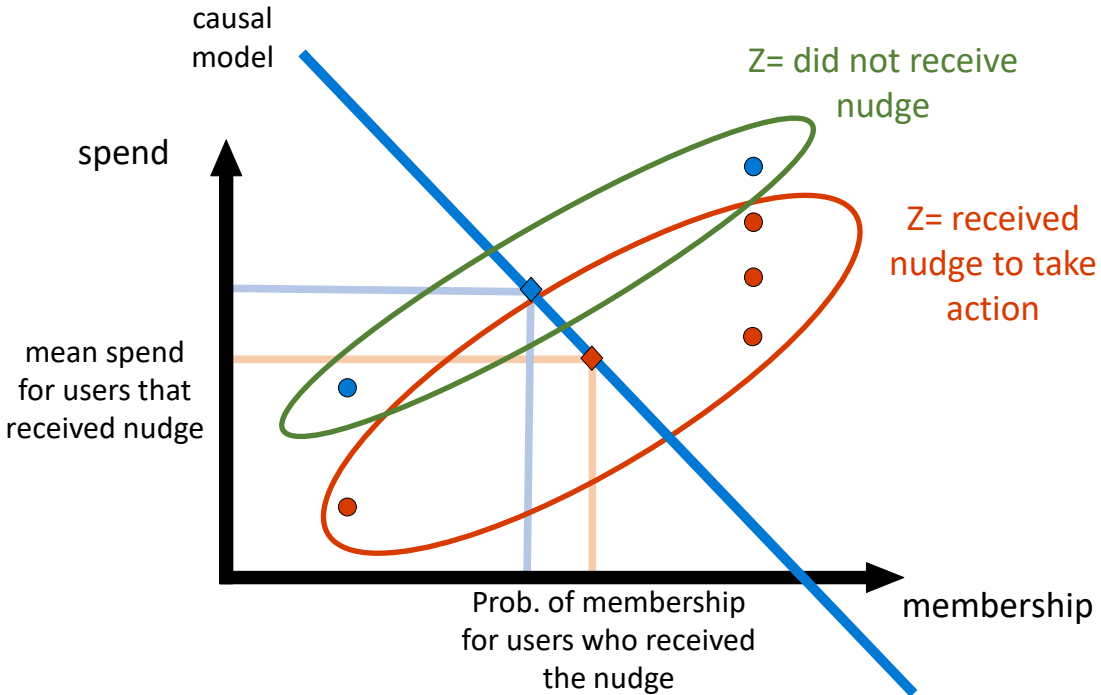


Digital Recommendation A/B tests



- D: action taken by user (e.g. membership), Y: spend
- X: observable characteristics of a user
- A: unobserved confounding factors (e.g. interest in platform)
- Z: randomized nudge to take action (e.g. one-click sign-up pop-up)





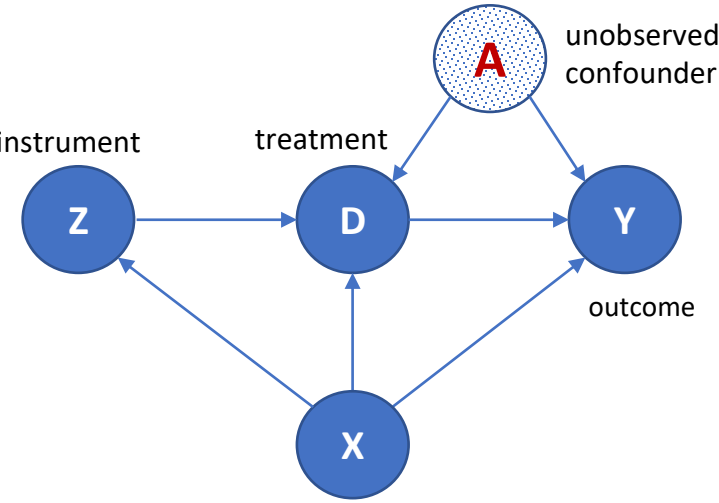
Limits of Identification via Instruments

- ATE identification via Instruments not based solely on DAG restrictions
- Requires further restrictions on structural equation models

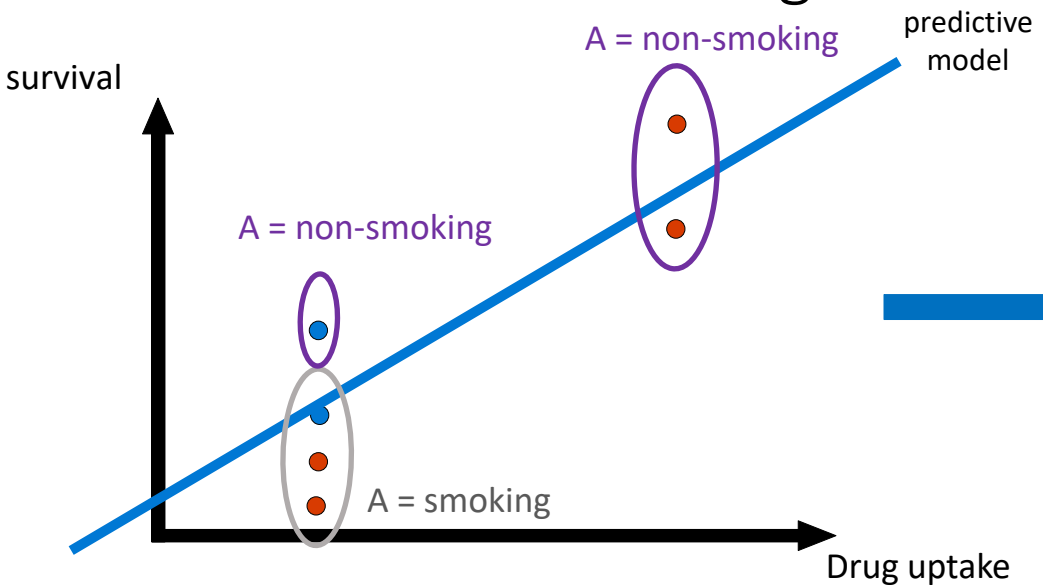
Example

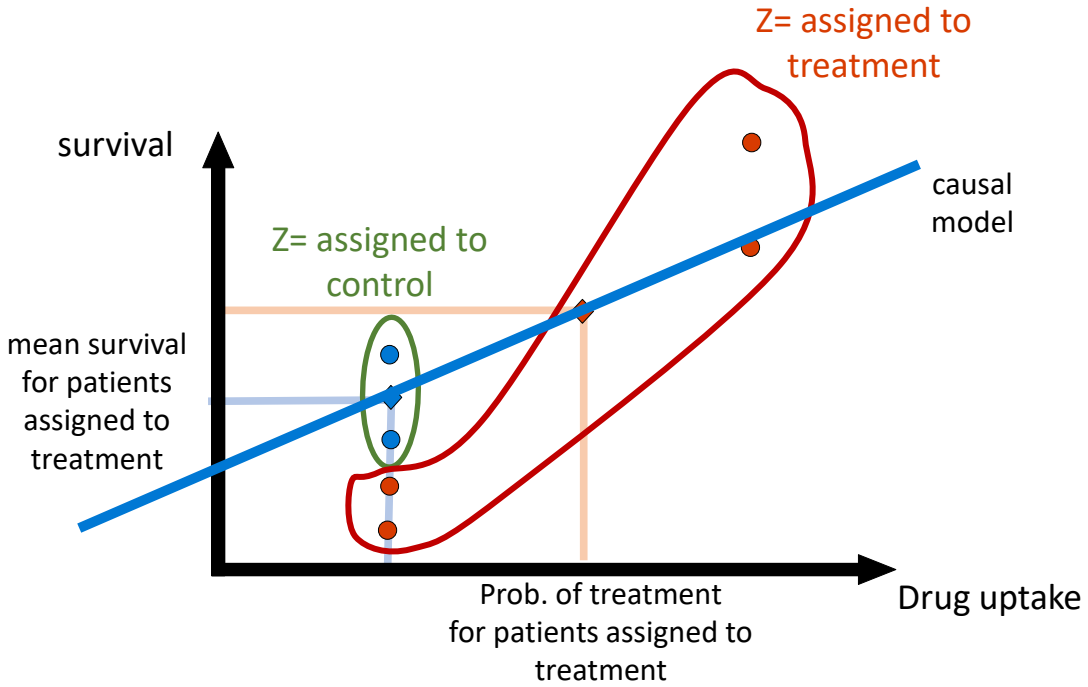
- Binary treatment D (drug) and binary instrument Z (drug recommendation)
- Consider an unobserved confounder A = “smoking”
- Suppose that smokers (A=1) never take the drug (never comply) and non-smokers (A=0) always follow the recommendation (comply)
- Suppose that drug has positive effects for non-smokers but has severe side-effects for smokers

Clinical Trials with Non-Compliance



- D: drug treatment, Y: survival
- X: observable characteristics of a patient
- A: unobserved “compliance factors” (e.g. health habits)
- Z: randomized cohort assignment





Limits of Identification via Instruments

- ATE identification via Instruments not based solely on DAG restrictions
- Requires further restrictions on structural equation models

Example

- IV regression will never be able to uncover the side effects of drug treatment on smokers
- Nothing in the data is informative of that
- Effect will be biased as compared to average effect in whole population

What do we need for ATE

- Either the compliance behavior (effect of instrument on treatment) does not vary with A (or X)
- Or the treatment effect (effect of treatment on outcome) does not vary with A (or X)

$$
\begin{aligned}
Y &:= g_Y(\epsilon_Y)\, D + f_Y(X, A, \epsilon_Y) \\
D &:= f_D(Z, X, A, \epsilon_D) \\
Z &= f_Z(X, \epsilon_Z) \\
A &:= f_A(X, \epsilon_A)
\end{aligned}
$$

$$
\begin{aligned}
Y &:= g_Y(X, A, \epsilon_Y)\, D + f(X, A, \epsilon_Y) \\
D &:= g_D(\epsilon_D) Z + f_D(X, A, \epsilon_D) \\
Z &= f_Z(X) + \epsilon_Z \\
A &:= f_A(X, \epsilon_A)
\end{aligned}
$$

Joint Variation on Observables

- If joint variation is captured through observables then ATE is feasible

$$Y \coloneqq g_Y(X, \epsilon_Y)\, D + f_Y(X, A, \epsilon_Y)$$
$$D \coloneqq f_D(Z, X, A, \epsilon_D)$$
$$Z = f_Z(X, \epsilon_Z)$$
$$A \coloneqq f_A(X, \epsilon_A)$$

$$Y \coloneqq g_Y(X, A, \epsilon_Y)\, D + f(X, A, \epsilon_Y)$$
$$D \coloneqq g_D(X, \epsilon_D)Z + f_D(X, A, \epsilon_D)$$
$$Z = f_Z(X, \epsilon_Z)$$
$$A \coloneqq f_A(X, \epsilon_A)$$

- We just need to do our identification analysis conditional on X and then average

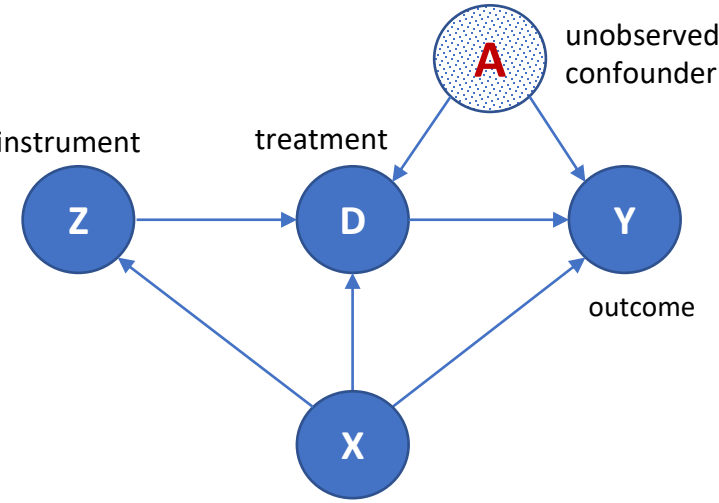
$$\beta(X) = \frac{E[\tilde{Y}\tilde{Z} \mid X]}{E[\tilde{D}\tilde{Z} \mid X]}, \qquad a = E[\beta(X)]$$

- Roughly: reweighting data based on compliance level $E[\tilde{D}\tilde{Z} \mid X]$

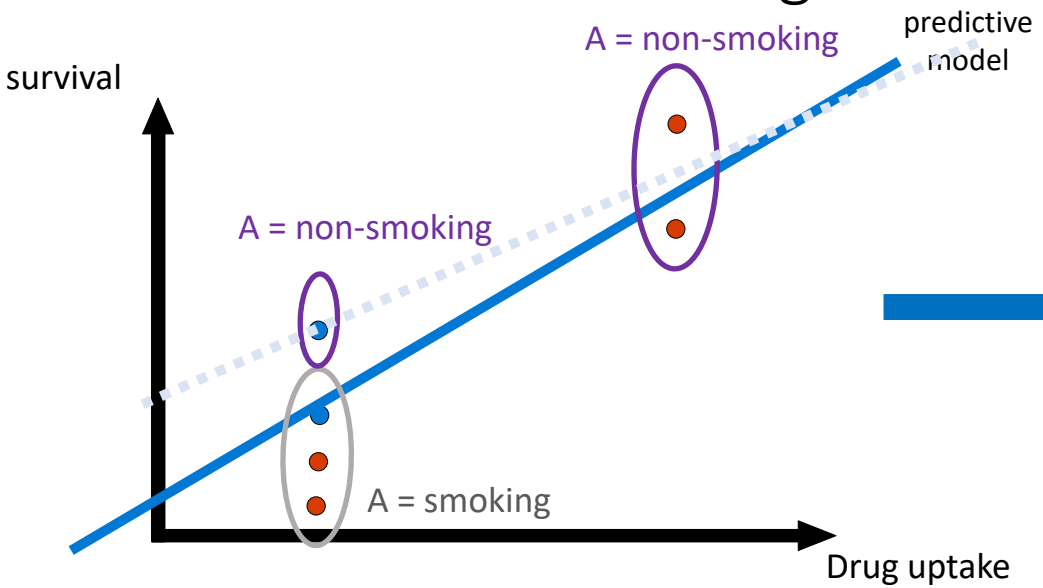
What if joint variation happens through unobservables?

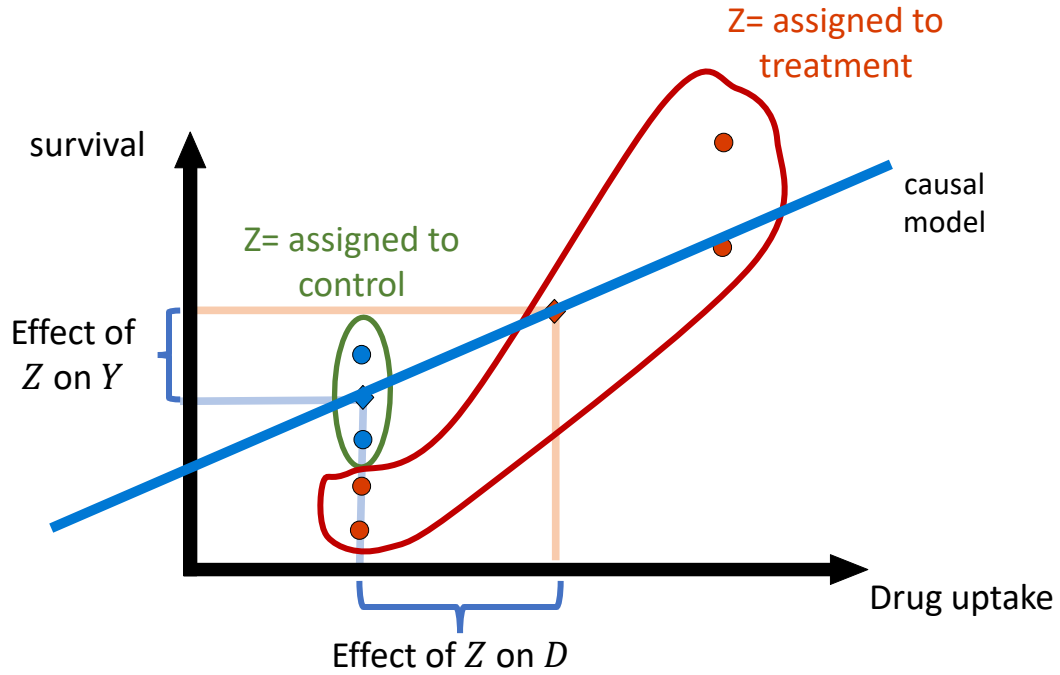
$$
\begin{aligned}
Y &\coloneqq f_Y(D, X, A, \epsilon_Y) \\
D &\coloneqq f_D(Z, X, A, \epsilon_D) \\
Z &= f_Z(X, \epsilon_Z) \\
A &\coloneqq f_A(X, \epsilon_A)
\end{aligned}
$$

Clinical Trials with Non-Compliance



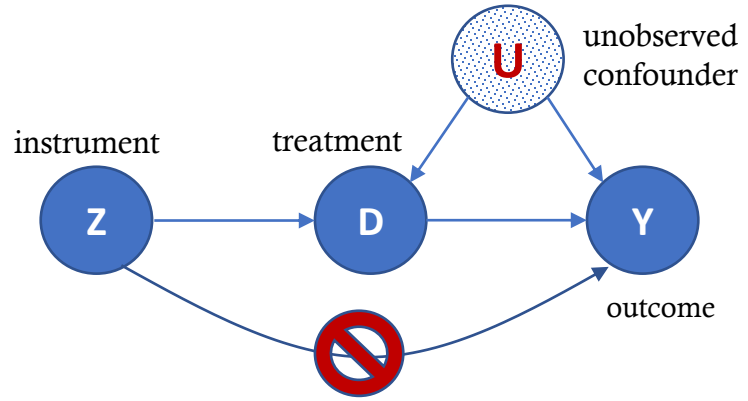
- D: drug treatment, Y: survival
- X: observable characteristics of a patient
- A: unobserved “compliance factors” (e.g. health habits)
- Z: randomized cohort assignment





Does the IV estimate coincide with the average effect for some sub-population?

The Binary Case



Imbens-Angrist (1994): core contribution of Nobel 2022 award

- Instrument/Treatment are binary (instrument=recommended treatment)
- Assume monotonicity: $D^{(1)} \geq D^{(0)}$
- Recommended treatment cannot reverse taken treatment
- Object of interest: Local Average Treatment Effect (ATE among compliers)

$$\theta_0 = E\left[\mathrm{Y}^{(1)} - \mathrm{Y}^{(0)} \middle| \mathrm{D}^{(1)} > \mathrm{D}^{(0)}\right]$$

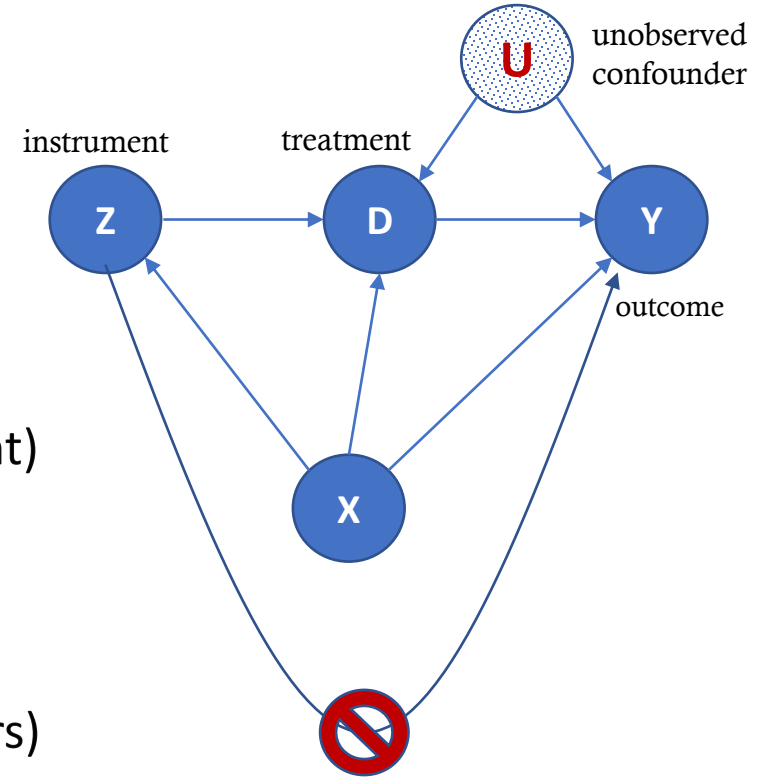
- Proof [Angrist-Imbens'94]:

$$\theta_0 = \frac{E\left[\left(Y^{(1)} - Y^{(0)}\right)1\left\{D^{(1)} > D^{(0)}\right\}\right]}{E\left[1\left\{D^{(1)} > D^{(0)}\right\}\right]} = \frac{E\left[Y^{(D(1))} - Y^{(D(0))}\right]}{E\left[D^{(1)} - D^{(0)}\right]} = \frac{ATE(Z \to Y)}{ATE(Z \to D)}$$

$$\gamma = \frac{\mathbb{E}\left[\tilde{Z}\tilde{y}\right]}{\mathbb{E}\left[\tilde{Z}^2\right]}$$

$$\delta = \frac{\mathbb{E}\left[\tilde{\mathrm{Z}}\tilde{\mathrm{D}}\right]}{\mathbb{E}\left[\tilde{\mathrm{Z}}^2\right]}$$

The Binary Case



Imbens-Angrist (1994): core contribution of Nobel 2022 award

- Instrument/Treatment are binary (instrument=recommended treatment)
- Assume monotonicity: $D^{(1)} \geq D^{(0)}$
- Recommended treatment cannot reverse taken treatment
- Object of interest: Local Average Treatment Effect (ATE among compliers)

$$\theta_0 = E\left[\mathrm{Y}^{(1)} - \mathrm{Y}^{(0)} \middle| \mathrm{D}^{(1)} > \mathrm{D}^{(0)}\right]$$

- Proof [Angrist-Imbens'94]:

$$\theta_0 = \frac{E\left[\left(Y^{(1)} - Y^{(0)}\right)1\left\{D^{(1)} > D^{(0)}\right\}\right]}{E\left[1\left\{D^{(1)} > D^{(0)}\right\}\right]} = \frac{E\left[Y^{(D(1))} - Y^{(D(0))}\right]}{E\left[D^{(1)} - D^{(0)}\right]} = \frac{ATE(Z \to Y)}{ATE(Z \to D)}$$

$$E\left[E[Y|Z=1,X] - E[Y|Z=0,X]\right]$$

$$E\left[E[D|Z=1,X] - E[D|Z=0,X]\right]$$

LATE in the Binary Case

- Under monotonicity

$$\theta_0 = \frac{E\big[E[\,Y \mid Z=1,X\,] - E[Y|Z=0,X]\big]}{E\big[E[D|Z=1,X] - E[D|Z=0,X]\big]}$$

- Moment formulation

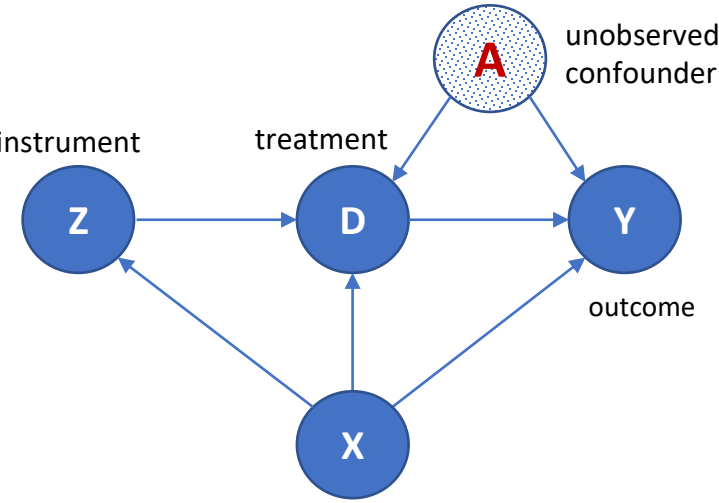
$$E[E[Y|Z=1,X] - E[Y|Z=0,X] - \theta_0(E[D|Z=1,X] - E[D|Z=0,X])] = 0$$

$$+ \boldsymbol{H(Z,X)(Y - E[Y|Z,X])} \qquad + \boldsymbol{H(Z,X)(D - E[D|Z,X])}$$

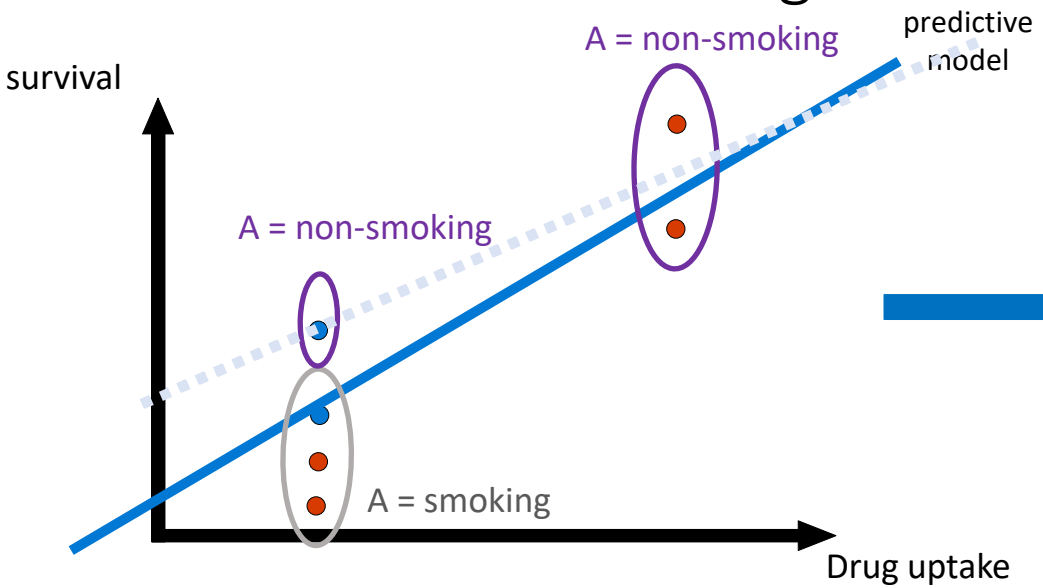
$$H(Z,X) = \frac{Z}{P(Z=1|X)} - \frac{1-Z}{1-P(Z=1|X)}$$

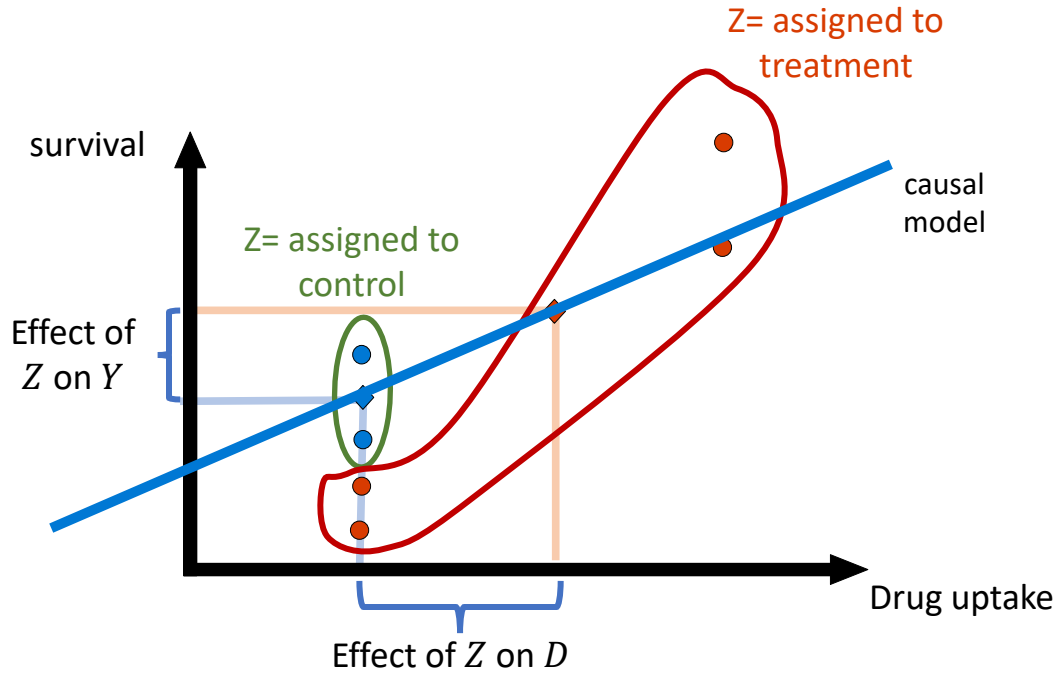
- Orthogonal moment formulation: apply ATE debiasing twice

Clinical Trials with Non-Compliance



- D: drug treatment, Y: survival
- X: observable characteristics of a patient
- A: unobserved “compliance factors” (e.g. health habits)
- Z: randomized cohort assignment





Confidence Intervals

Main Theorem (expanded) Define RMSE: $\|h\|_{L^2} = \sqrt{E[h(X)^2]}$

- If moment is Neyman orthogonal and RMSE of $\hat{g}$ goes down at rate $n^{1/4}$, plus regularity conditions
$$n^{1/4}\|\hat{g} - g_0\|_{L^2} \approx 0$$

- Then the estimate $\hat{\theta}$ is *asymptotically linear*
$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \approx \sqrt{n}\, E_n[\phi_0(Z)], \qquad \phi_0(Z) = -J_0^{-1}\, m(Z; \theta_0, g_0), \qquad J_0 \coloneqq \partial_\theta E[m(Z; \theta_0, g_0)]$$

- Consequently, it is *asymptotically normal*
$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \sim_a N(0, V), \qquad V \coloneqq E[\phi_0(Z)\phi_0(Z)']$$

- *Confidence intervals* for any projection based on estimate of variance are asymptotically valid

$$\ell'\theta \in \left[\ell'\hat{\theta} \pm c\sqrt{\frac{\ell'\hat{\mathbb{V}}\ell}{n}}\right], \qquad \hat{\mathbb{V}} = \mathrm{Var_n}\left(\hat{\phi}(Z)\right), \qquad \hat{\phi}(Z) \coloneqq -\hat{J}^{-1}m\left(Z; \hat{\theta}, \hat{g}\right), \qquad \hat{J} = \partial_\theta E_n\left[m\left(Z; \hat{\theta}, \hat{g}\right)\right]$$

Partially Linear Instrumental Variable Model

- Suppose that we can identify parameter of interest via moment

$$E\left[\left(\tilde{Y} - \theta_0 \tilde{D}\right)\tilde{Z}\right] = 0$$

- Setting falls into the general moment estimation framework

$$M(\theta, h, p, m) = E\left[\left(Y - h(X) - \theta\left(D - p(X)\right)\right)\ \left(Z - m(X)\right)\right] = 0$$

Where $h(X) = E[Y|X]$, $p(X) = E[D|X]$, $m(Z) = E[Z|X]$

Inference with DML in PLIV Setting

- If RMSE of propensity model and outcome model goes down at rate $n^{1/4}$, plus regularity conditions
- Then the estimate $\hat{\theta}$ is *asymptotically linear*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \approx \sqrt{n}\, E_n[\phi_0(Z)], \qquad \phi_0(Z) = -J_0^{-1}\left(\tilde{Y} - \theta_0 \tilde{D}\right)\tilde{Z}, \qquad J_0 := E\left[\tilde{D}\tilde{Z}\right]$$

Jacobian of moments with respect to parameter; relates to strength of instrument

- Consequently, it is *asymptotically normal*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \sim_a N(0, V), \qquad V := \frac{E\left[\left(\tilde{Y} - \theta_0 \tilde{D}\right)^2 \tilde{Z}^2\right]}{E\left[\tilde{D}\tilde{Z}\right]^2}$$

- *Confidence intervals* for any projection based on estimate of variance are asymptotically valid

$$\ell'\theta \in \left[\ell'\hat{\theta} \pm c\sqrt{\frac{\ell'\hat{V}\ell}{n}}\right], \qquad \hat{\mathrm{V}} = \mathrm{Var_n}\left(\hat{\phi}(Z)\right), \qquad \hat{\phi}(Z) := -\hat{J}^{-1}\left(\hat{Y} - \hat{\theta}\hat{D}\right)\hat{Z}, \qquad \hat{J} = E_n\left[\hat{D}\hat{Z}\right]$$

LATE in the Binary Case

- Under monotonicity

$$\theta_0 = \frac{E\big[E[\,Y \mid Z=1,X\,] - E[Y|Z=0,X]\big]}{E\big[E[D|Z=1,X] - E[D|Z=0,X]\big]}$$

- Moment formulation

$$E[E[Y|Z=1,X] - E[Y|Z=0,X] - \theta_0(E[D|Z=1,X] - E[D|Z=0,X])] = 0$$

$$+ \boldsymbol{H(Z,X)(Y - E[Y|Z,X])} \qquad + \boldsymbol{H(Z,X)(D - E[D|Z,X])}$$

$$H(Z,X) = \frac{Z}{P(Z=1|X)} - \frac{1-Z}{1-P(Z=1|X)}$$

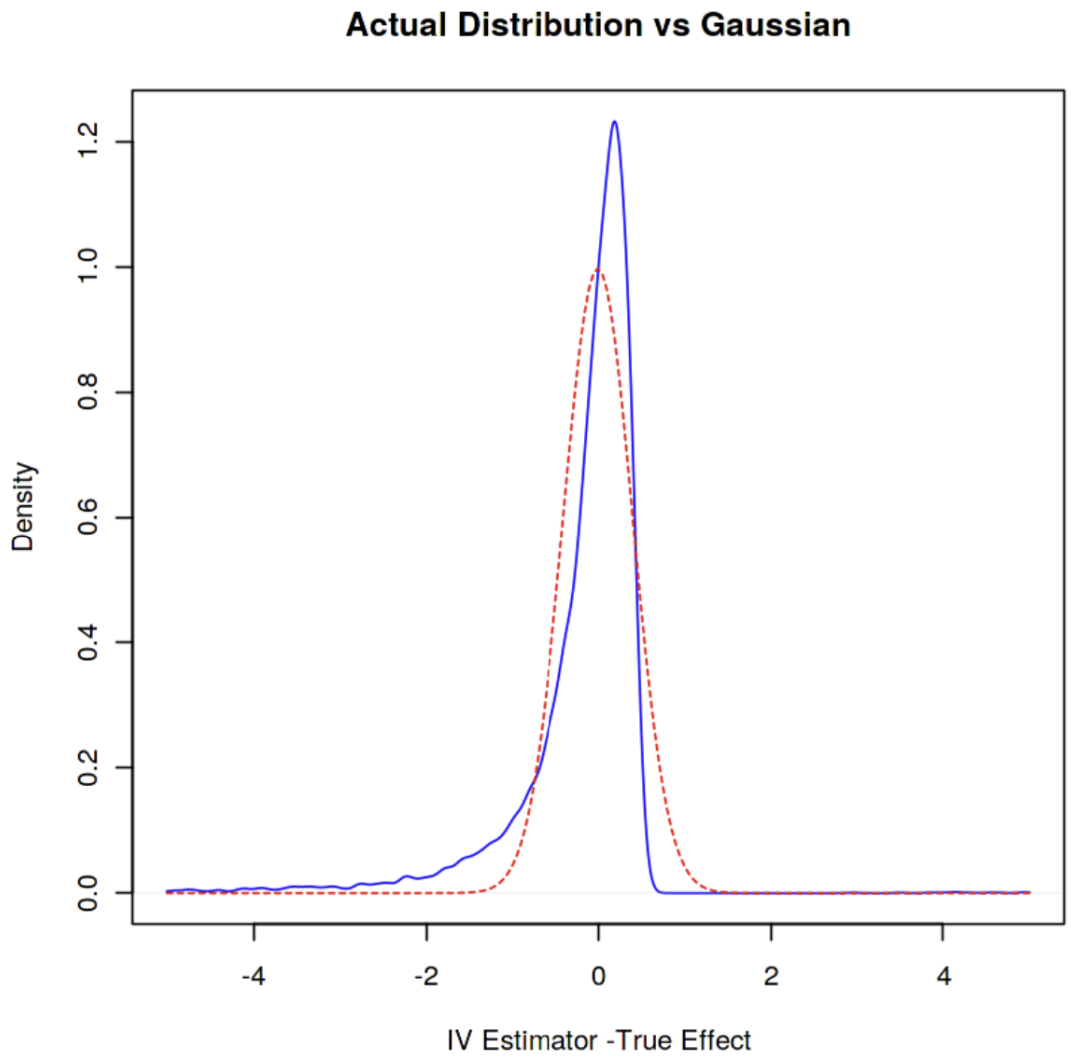
- Orthogonal moment formulation: apply ATE debiasing twice

Weak Identification

- If $\boldsymbol{E}\left[\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{Z}}\right]$ is small and comparable with the sample size, then approximation

$$E_n\left[\widetilde{D}\widetilde{Z}\right]^{-1} \approx E\left[\widetilde{D}\widetilde{Z}\right]$$

- Can be inaccurate in finite samples and normal based approximation will yield in-correct confidence intervals



A More Robust Inference Approach

- Even in the weak regime the moment constraint is still well-behaved

$$E\left[\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right]$$

- At the true parameter θ_0 we know that:

$$\mathrm{C}(\theta) \coloneqq \frac{\left(\sqrt{n}\, E_n\left[\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right]\right)^2}{Var_n\left(\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right)} \sim_a \left(N(0,1)\right)^2 = \chi^2(1)$$

- This statistic does not hinge on inversion of the Jacobian; approximation remains valid even with cross-fitted approximate residuals due to Neyman orthogonality
- We can perform a grid search over candidate parameters θ and for every such parameter test whether (for confidence interval with confidence α)

$$C(\theta) \leq (1-\alpha)\ \text{quantile of}\ \chi^2(1)$$

- Then by construction: $\Pr\left(\theta_0 \in C(\theta)\right) \approx 1 - \alpha$

General Moments and Weak Identification

- For a general Neyman orthogonal moment

$$E[m(Z;\theta_0, g_0)] = 0$$

- We can construct a statistic that is robust to weak identification (i.e. Jacobian $\partial_\theta E[m(Z;\theta_0, g_0)]$ very small)

$$C(\theta) = \frac{\left(\sqrt{n}E_n[m(Z;\theta,\hat{g})]\right)^2}{Var_n\left(m(Z;\theta,\hat{g})\right)} \sim_a \chi^2(1)$$

- Construct a α-confidence region by including all parameter values θ s.t.

$$C(\theta) \leq (1-\alpha) \text{ quantile of } \chi^2(1)$$

- Then by construction: $\Pr\left(\theta_0 \in C(\theta)\right) \approx 1 - \alpha$