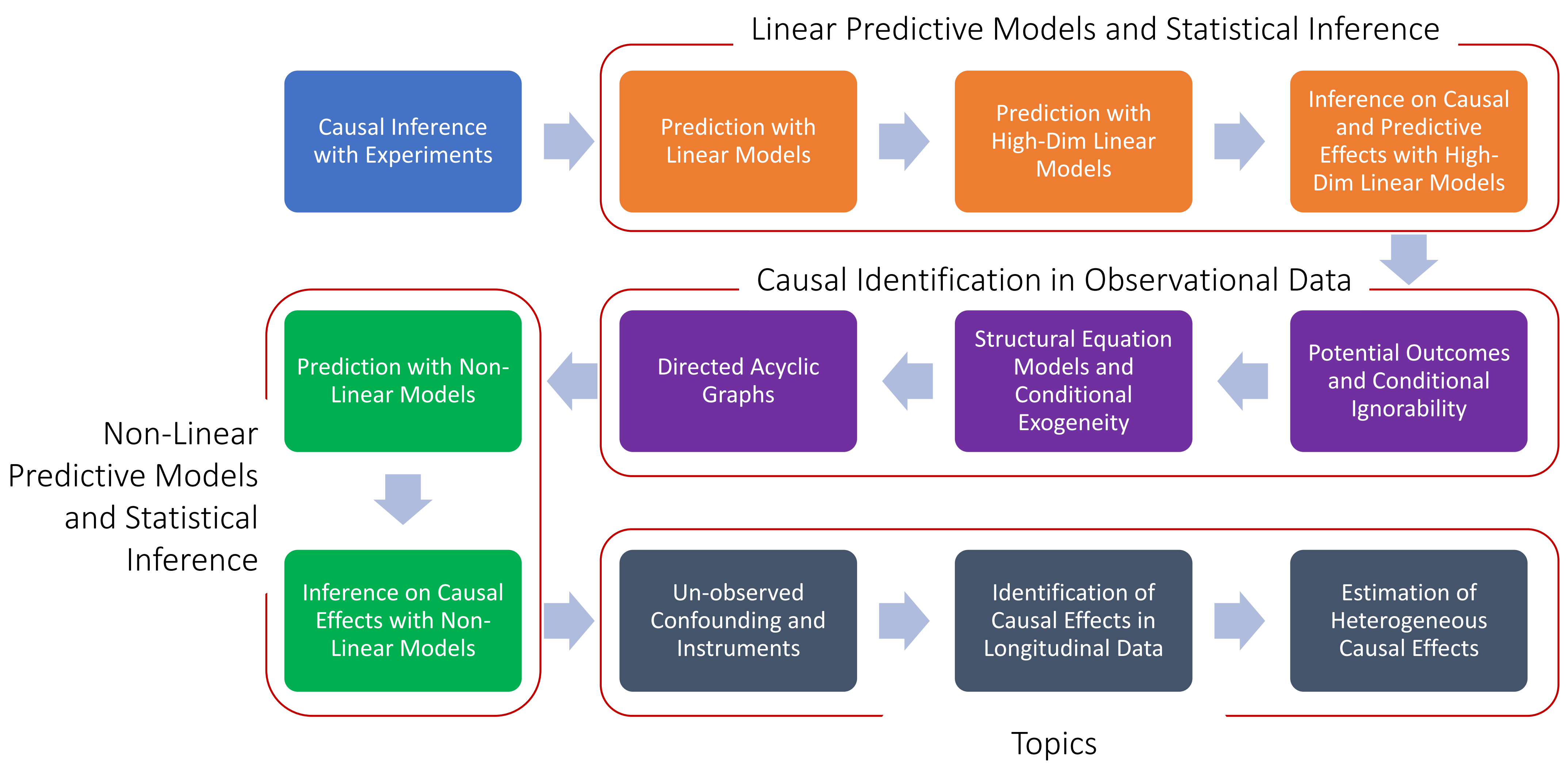
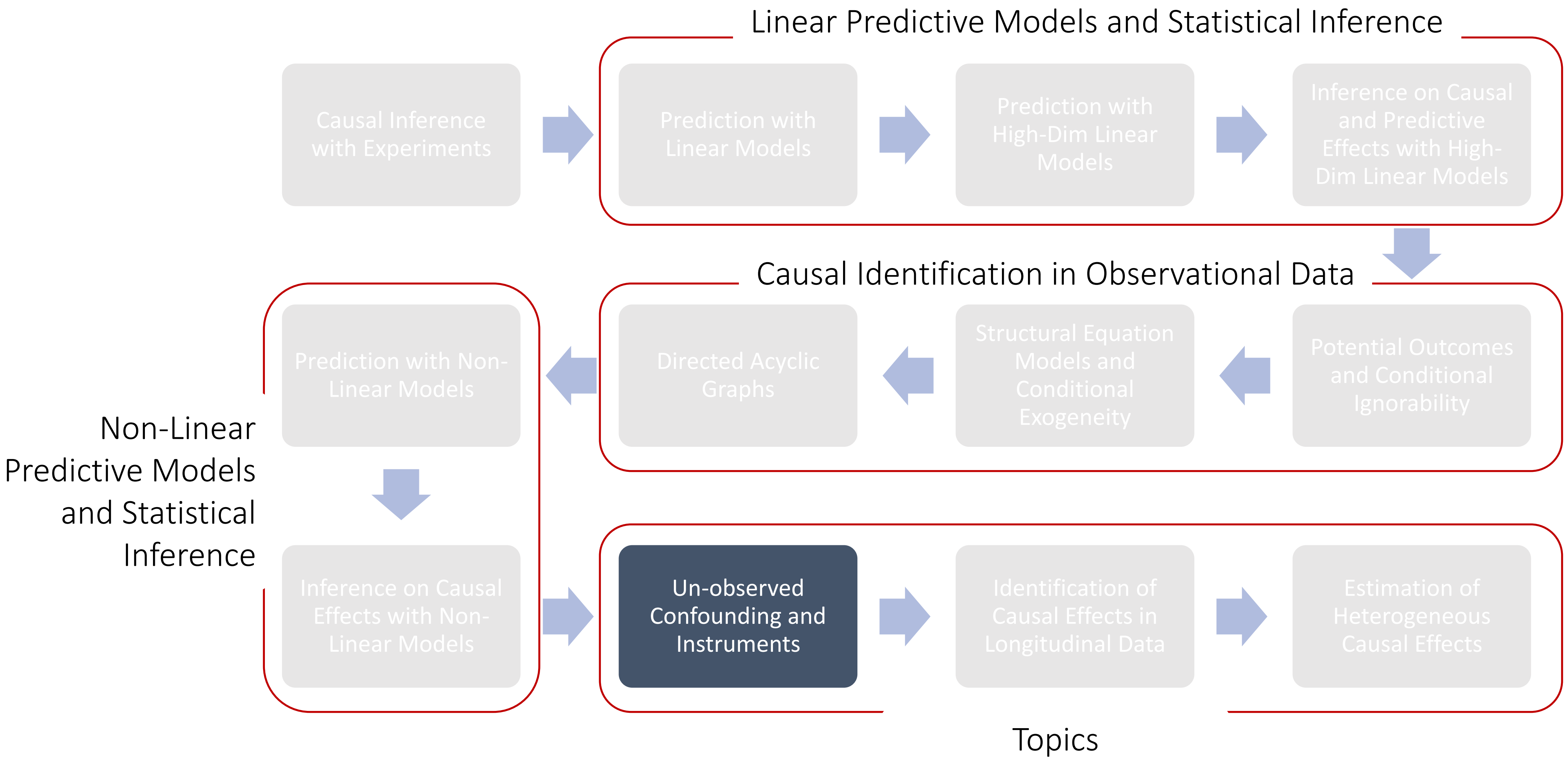
MS&E 228: Unobserved Confounding and Instruments

Vasilis Syrgkanis

MS&E, Stanford



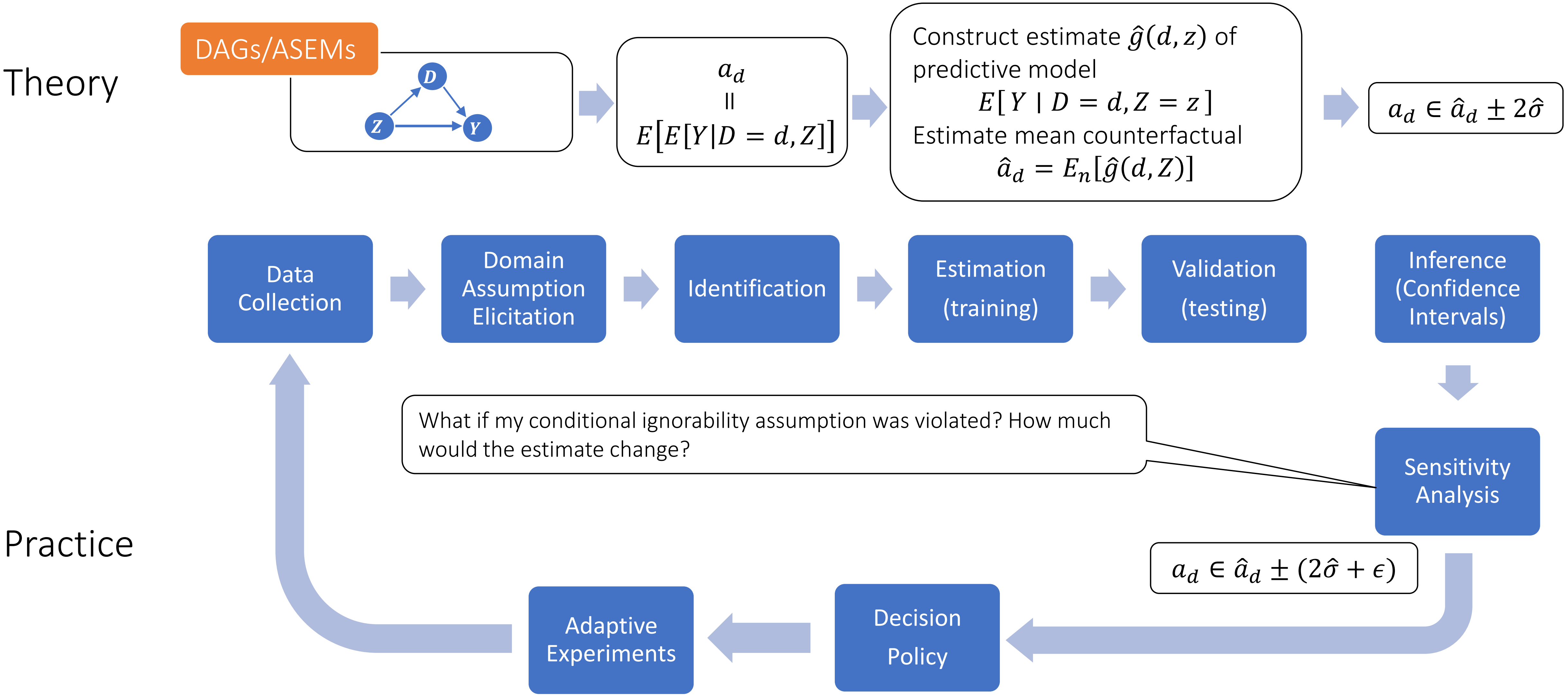
Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional 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
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

Goals for Today

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

Causal Inference Pipeline



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

$$\boxed{R^2_{Y\sim A|D,X}} \leq C_Y^2, \qquad \boxed{R^2_{D\sim A|X}} \leq C_D^2$$

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

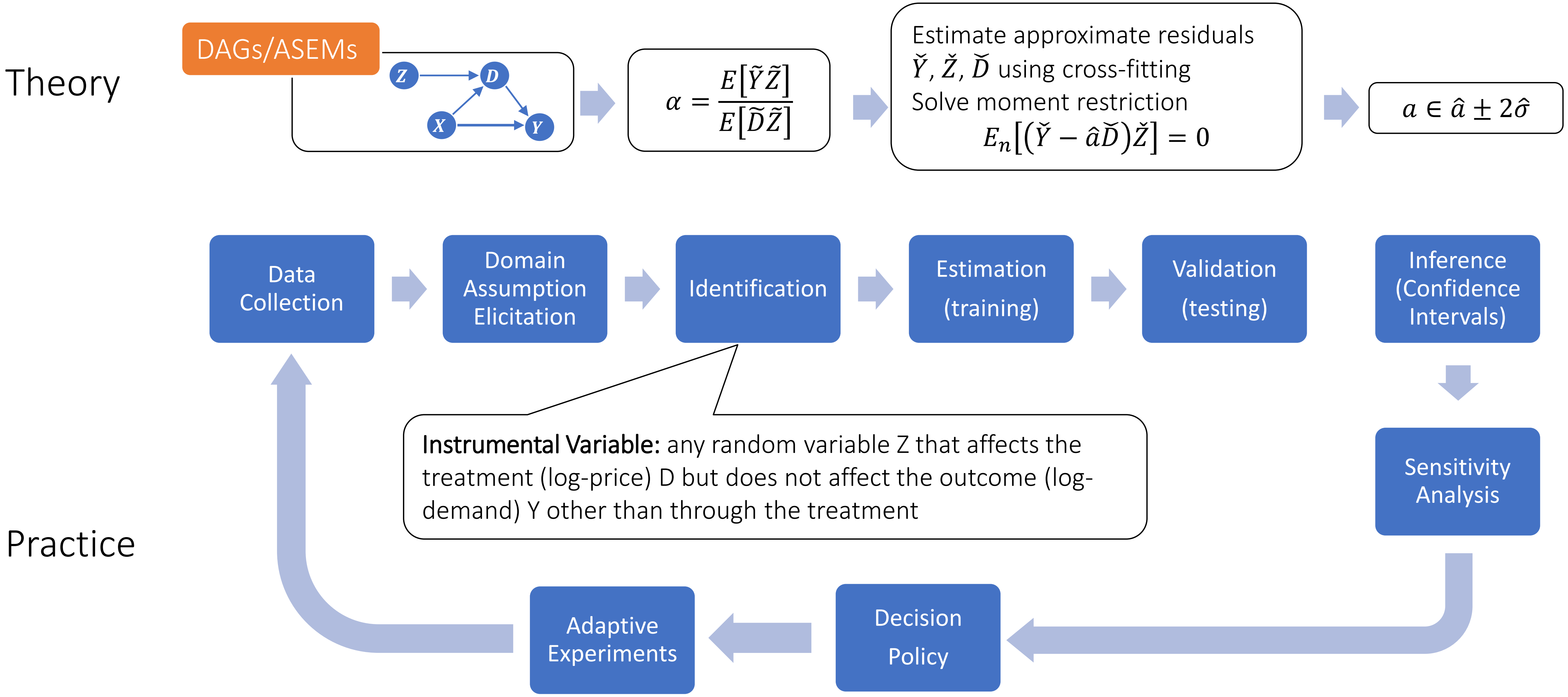
$$\theta_0 \in \theta_s \pm \sqrt{C_Y^2 \frac{C_D^2}{1 - C_D} \boxed{\frac{E\left[\left(\tilde{Y} - \theta_s \tilde{D}\right)^2\right]}{E\left[\tilde{D}^2\right]}}}$$

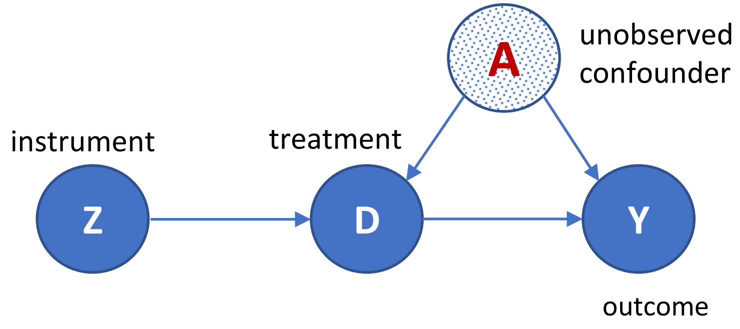
Measurable from the data

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

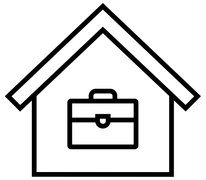




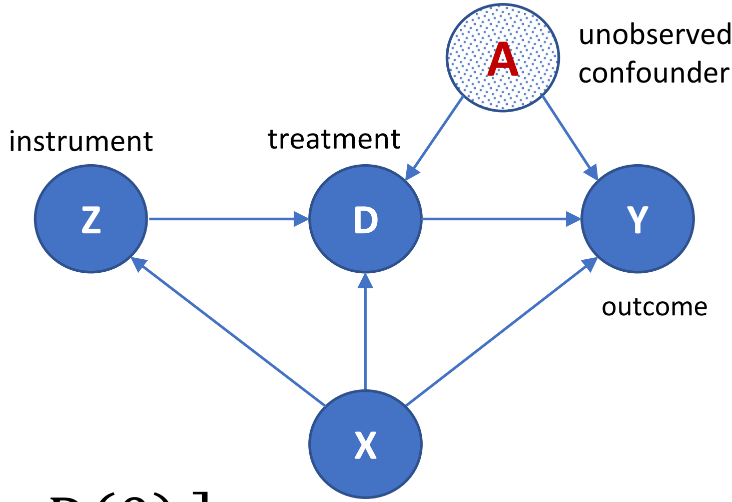
If we have access to a valid instrument Z, i.e. $Y(d), D(z) \perp\!\!\!\perp Z$ then typical approach is to estimate an effect via 2SLS ($\tilde{V} = V - E[V]$):

$$\theta = \frac{ATE(Z \rightarrow Y)}{ATE(Z \rightarrow D)} = \frac{E\left[\tilde{Y}\tilde{Z}\right]}{E\left[\tilde{D}\tilde{Z}\right]} = \frac{Cov(Y,Z)}{Cov(D,Z)}$$

If treatment and instrument are binary + monotonicity (instrument cannot “reverse” treatment choice), then this estimates the **Local Average Treatment Effect** (average effect among compliers):



$$\theta = \frac{ATE(Z \rightarrow Y)}{ATE(Z \rightarrow D)} = E[\,Y(1) - Y(0) \mid D(1) > D(0)\,]$$



If we only have a conditionally valid instrument, i.e.

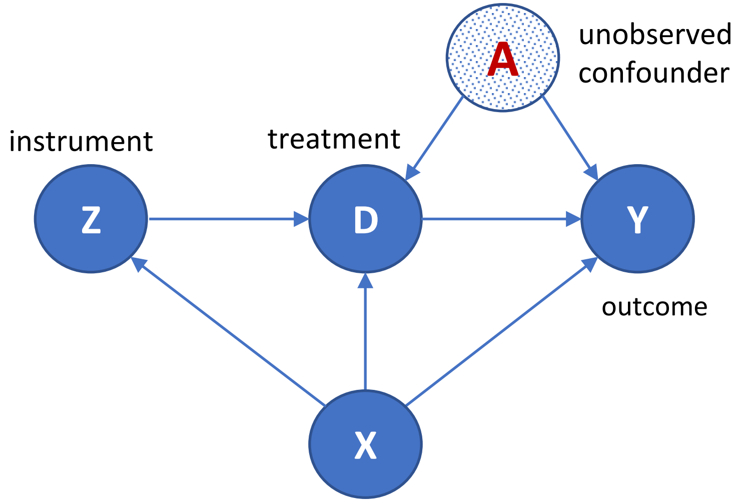
$$Y(d), D(z) \perp\!\!\!\perp Z \mid X$$

then still we have that:

$$\theta = \frac{ATE(Z \to Y)}{ATE(Z \to D)} = E[\, Y(1) - Y(0) \mid D(1) > D(0) \,]$$

But now this ratio is identified by a ratio of g-formulas

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



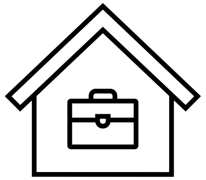
If we only have a conditionally valid instrument, i.e.

$$Y(d), D(z) \perp\!\!\!\perp Z \mid X$$

then still we have that:

$$\theta = \frac{ATE(Z \to Y)}{ATE(Z \to D)} = E[\,Y(1) - Y(0) \mid D(1) > D(0)\,]$$

But now this ratio is identified by a ratio of g-formulas

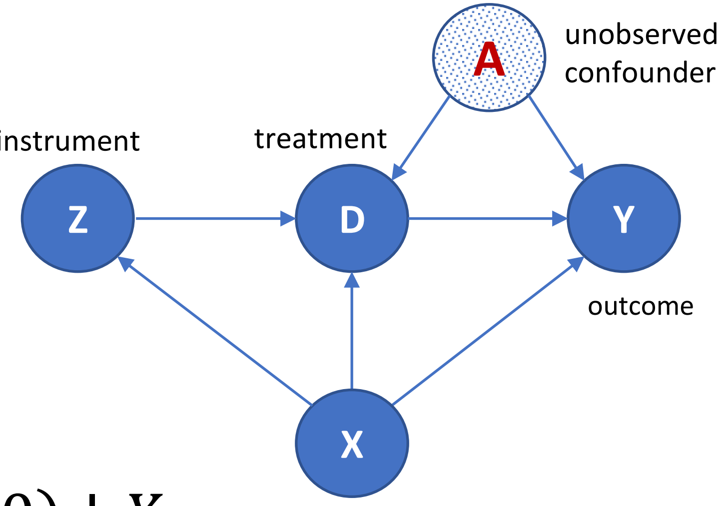


$$\theta = \frac{E\big[E[\,Y \mid Z = 1, X\,] - E[\,Y \mid Z = 0, X\,]\big]}{E\big[E[\,D \mid Z = 1, X\,] - E[\,D \mid Z = 0, X\,]\big]}$$

If we further assume that

$$Y(d = 1) - Y(d = 0) \perp\!\!\!\perp \mathrm{D(z = 1) - D(z = 0)}$$

Then this estimates the ATE



If we only have a conditionally valid instrument, i.e.

$$Y(d), D(z) \perp\!\!\!\perp Z \mid X$$

If we just assume that

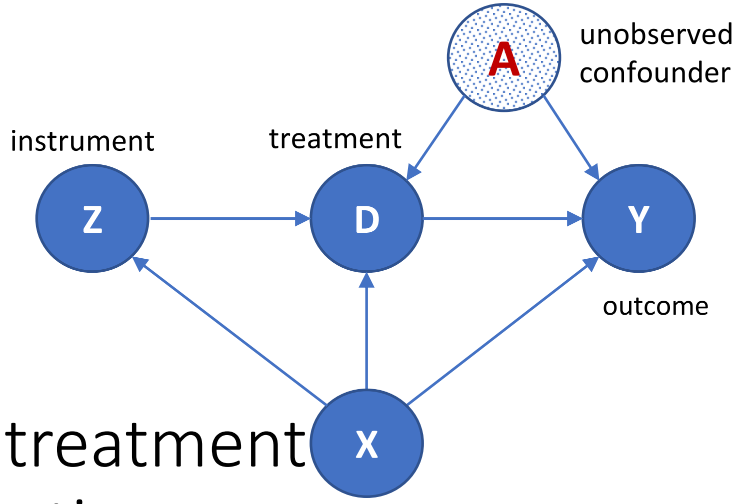
$$Y(d=1) - Y(d=0) \perp\!\!\!\perp \mathrm{D(z=1) - D(z=0)} \mid X$$

Then the ATE can be identified as:

$$\theta = E_X\left[\frac{ATE(Z \to Y \mid X)}{ATE(Z \to D \mid X)}\right] = E[Y(1) - Y(0)]$$

This is identified by the expected ratio of g-formulas

$$\theta = E_X\left[\frac{E[\,Y \mid Z = 1, X\,] - E[\,Y \mid Z = 0, X\,]}{E[\,Y \mid Z = 1, X\,] - E[\,Y \mid Z = 0, X\,]}\right]$$



If we have a possibly continuous instrument or treatment and assume the semi-parametric structural equation:

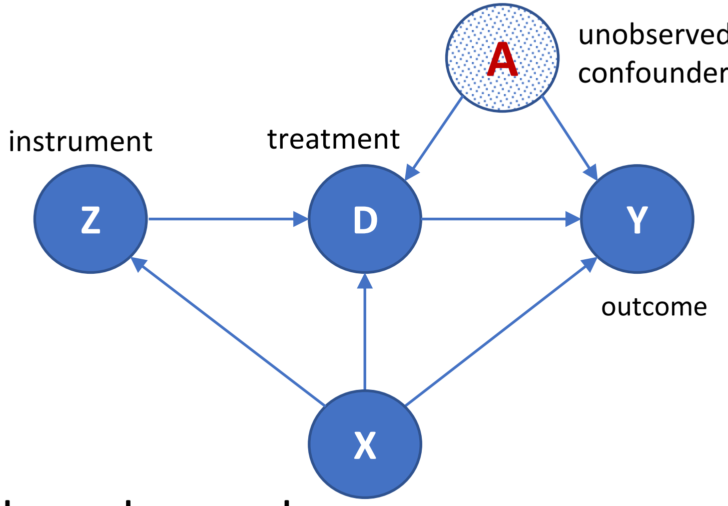
$$Y(d, x, a) \coloneqq \theta \cdot D + f_Y\left(x, a, \epsilon_y\right)$$

then we have the identifying moment condition:

$$E\left[\left(\tilde{Y} - \theta_0 \tilde{D}\right) \tilde{Z}\right] = E\left[f_Y(X, A, \epsilon_Y)\, \tilde{Z}\right] = 0$$

where $\tilde{V} = V - E[V|X]$

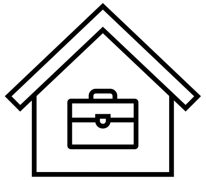
This moment is also Neyman orthogonal



If we have a possibly continuous instrument or treatment and assume the semi-parametric structural equation:

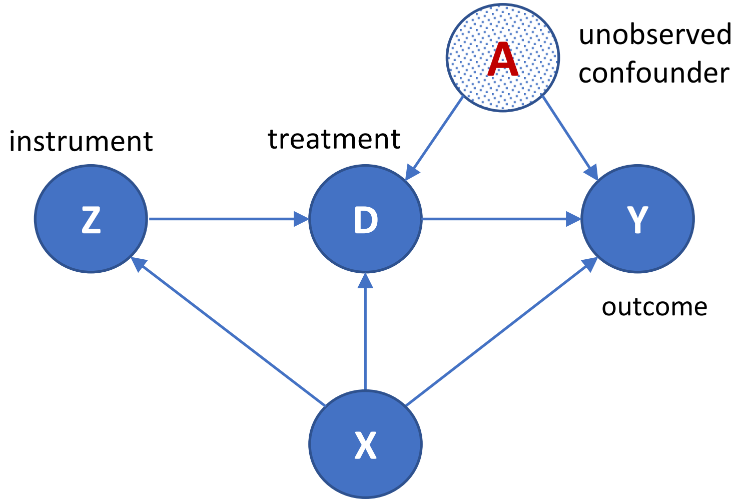
$$Y(d,x,a) \coloneqq g_Y(\epsilon_Y)\cdot D + f_Y\left(x,a,\epsilon_y\right)$$

then we have the identifying moment condition:



$$\begin{aligned} E\left[\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right] &= E\left[\left((g_Y(\epsilon_Y)-\theta)D + f_Y(X,A,\epsilon_Y)\right)\tilde{Z}\right] \\ &= E\left[(g_Y(\epsilon_Y)-\theta)\,D\tilde{Z}\right] \\ &= E[g_Y(\epsilon_Y)-\theta]\cdot E\left[D\tilde{Z}\right] \end{aligned}$$

$$E\left[\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right] = 0 \Rightarrow \theta = E[g(\epsilon_Y)] = \text{Average Marginal Effect}$$



If we have a possibly continuous instrument or treatment and assume the semi-parametric structural equation:

$$Y(d,x,a) \coloneqq g_Y(x,a,\epsilon_Y)\cdot D + f_Y\left(x,a,\epsilon_y\right)$$
$$D(z,x,a) \coloneqq g_D(\epsilon_D)\cdot Z + f_D(x,a,\epsilon_D)$$
$$Z(x) \coloneqq f_Z(x) + \epsilon_Z$$

then we can also argue that:

$$E\left[\left(\tilde{Y} - \theta\tilde{D}\right)\tilde{Z}\right] = 0 \Rightarrow \theta = E[g(X,A,\epsilon_Y)] = \text{Average Marginal Effect}$$

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

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

Inference with DML in PLIV Setting

- The estimate can be written as:

$$\hat{\theta} = \frac{E_n\big[\hat{Y}\hat{Z}\big]}{E_n\big[\hat{Z}\widehat{D}\big]}$$

- If RMSE of propensity models and outcome model goes down at rate $n^{1/4}$, plus regularity conditions

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

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

$$\ell'\theta \in \left[\ell'\hat{\theta} \pm c\sqrt{\frac{\ell'\hat{V}\ell}{n}}\right], \qquad \widehat{\mathrm{V}} = \frac{E_n\Big[\big(\hat{Y} - \hat{\theta}\,\widehat{D}\big)^2\hat{Z}^2\Big]}{E_n\big[\hat{Z}\widehat{D}\big]^2}$$

LATE in the Binary Case

- Under monotonicity

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

- Moment formulation

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

$$+ \qquad\qquad\qquad\qquad +$$

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

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

- Orthogonal moment formulation: apply ATE debiasing twice

Inference on LATE in the Binary Case

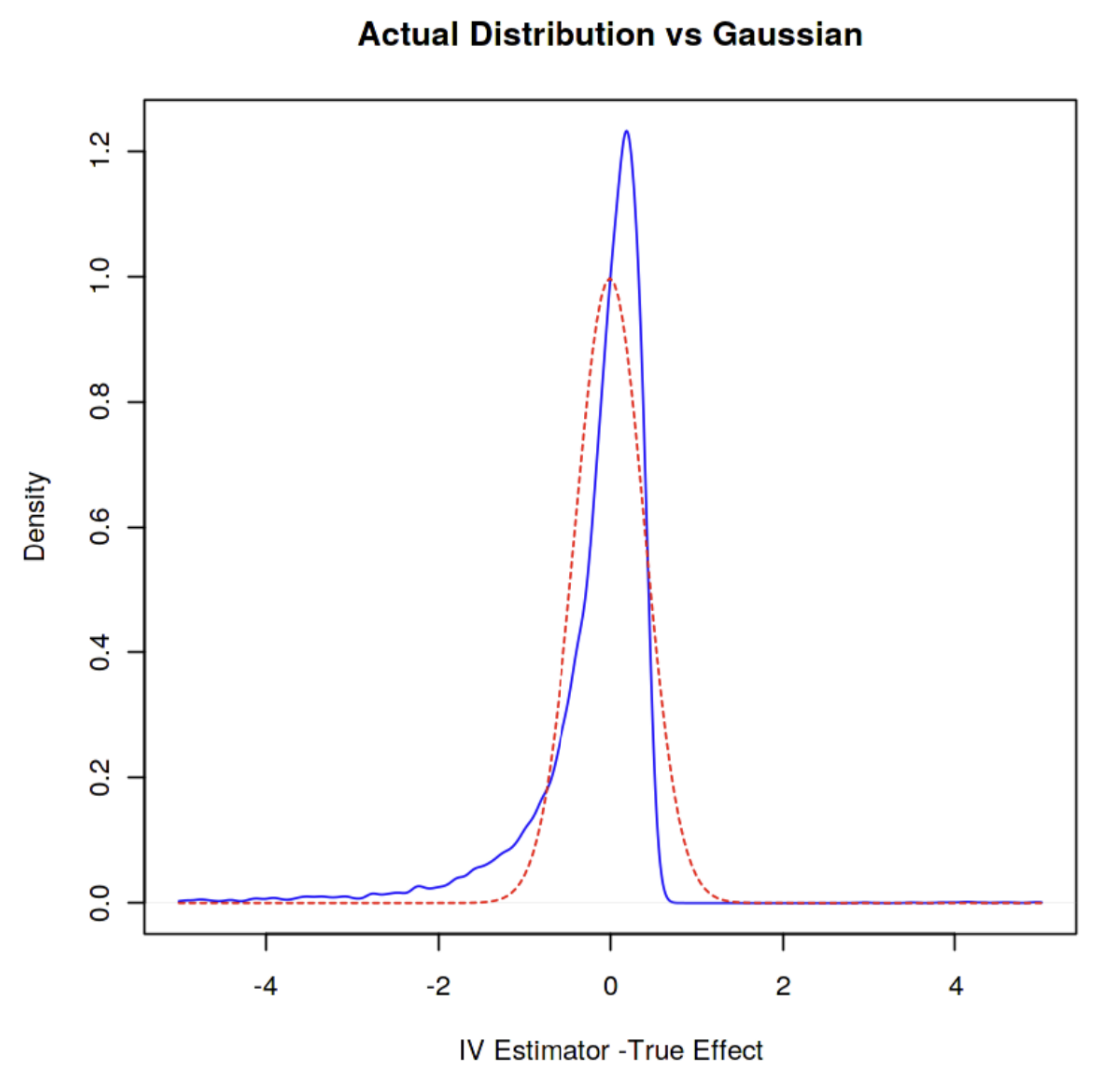
$$\hat{\theta} = \frac{E_n\left[\hat{h}_{Z\to Y}(1,X) - \hat{h}_{Z\to Y}(0,X) + a(Z,X)\left(Y - \hat{h}_{Z\to Y}(Z,X)\right)\right]}{E_n\left[\hat{h}_{Z\to D}(1,X) - \hat{h}_{Z\to D}(0,X) + a(Z,X)\left(D - \hat{h}_{Z\to D}(Z,X)\right)\right]}$$

- If RMSE of propensity models and outcome model goes down at rate $n^{1/4}$, plus regularity conditions

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \sim_a N(0,V)$$

Weak Identification

- If $\boldsymbol{E}[\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{Z}}]$ is small and comparable with the sample size, then approximation
$$E_n[\widetilde{D}\widetilde{Z}]^{-1} \approx E[\widetilde{D}\widetilde{Z}]^{-1}$$
- 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
$$E\left[\left(\tilde{Y} - \theta\widetilde{D}\right)\tilde{Z}\right]$$
- At the true parameter θ_0 we know that:
$$\mathrm{C}(\theta) := \frac{\left(\sqrt{n}\, E_n\left[\left(\tilde{Y} - \theta\widetilde{D}\right)\tilde{Z}\right]\right)^2}{Var_n\left(\left(\tilde{Y} - \theta\widetilde{D}\right)\tilde{Z}\right)} \sim_a \left(N(0,1)\right)^2 = \chi^2(1)$$
- This statistic does not hinge on inversion of $E\left[\widetilde{D}\tilde{Z}\right]$; 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: $\Pr\left(\theta_0\in C(\theta)\right)\approx 1-\alpha$