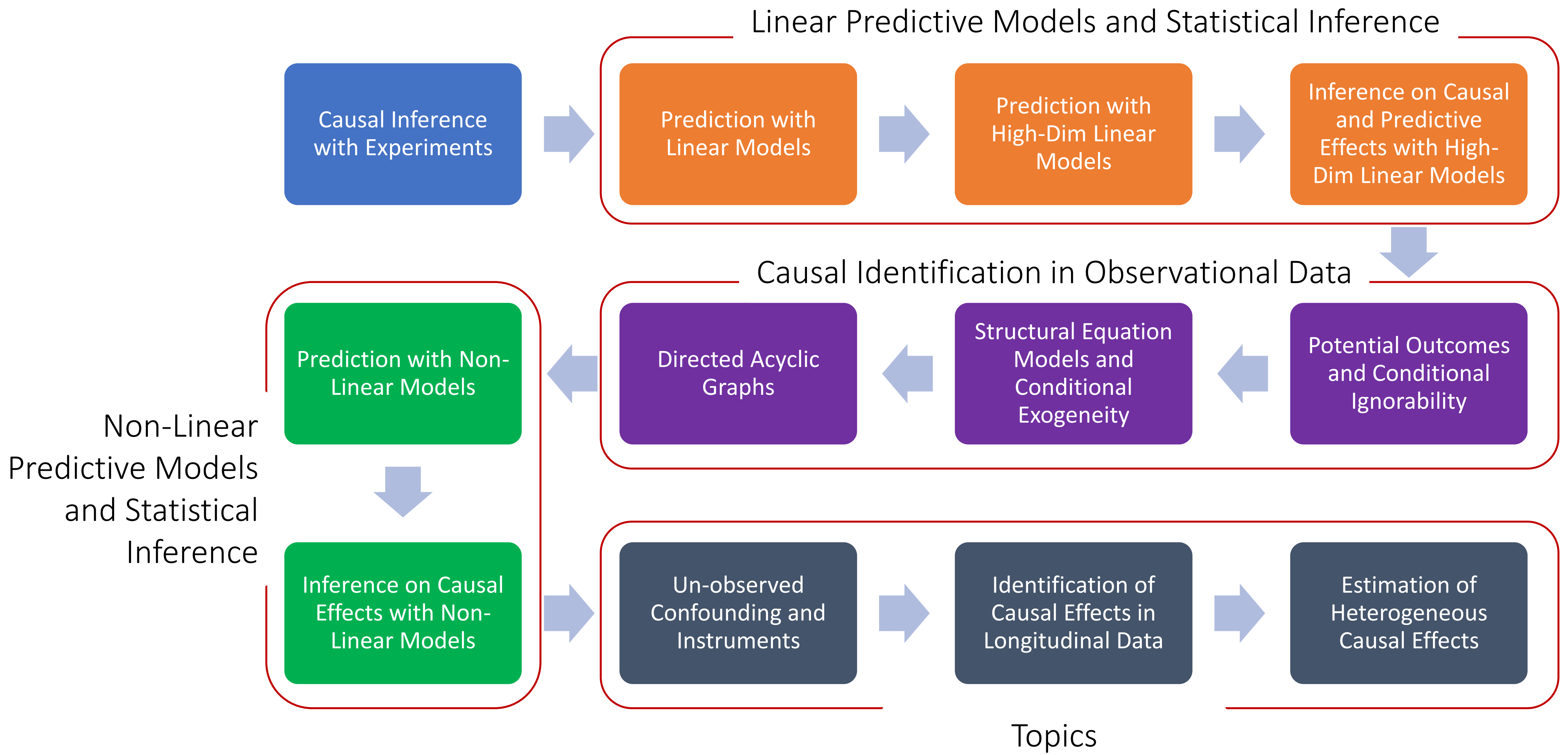
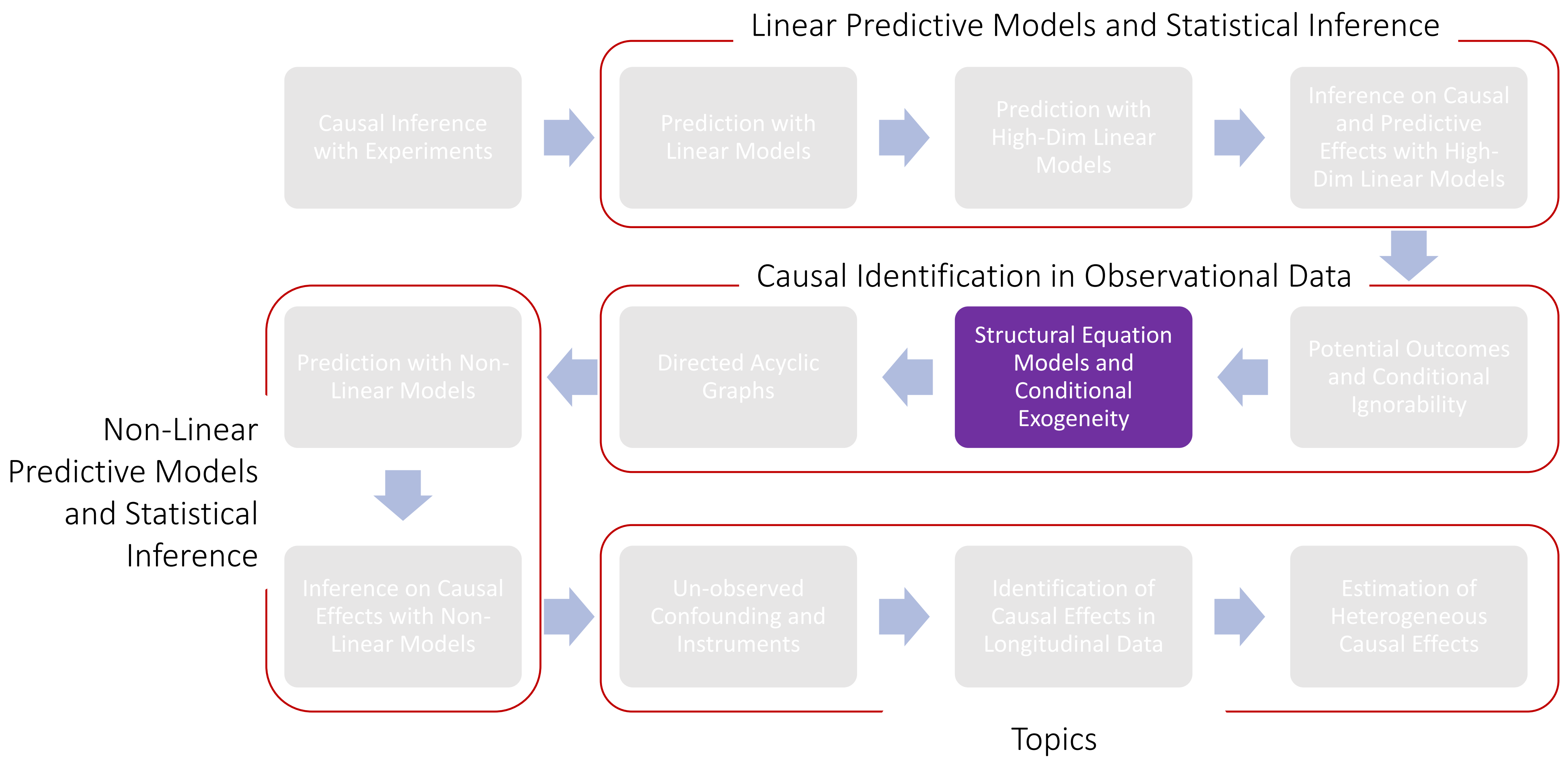
MS&E 228: Structural Equation Models

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
Prediction with Non-Linear Models
Directed Acyclic Graphs
Structural Equation Models and Conditional Exogeneity
Potential Outcomes and Conditional Ignorability
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

- Learn the “language” of (linear) structural equation models (SEMs)
- Connect to “potential outcome” and the “causal diagram” language
- Introduce “intervention” concepts
- Introduce some causal diagram concepts
- Examples of how to use SEMs (direct/indirect effects, collider bias, wage discrimination)

Structural Equation Models

[1] Philip G. Wright. *The tariff on animal and vegetable oils*. New York: The Macmillan company, 1928 (cited on pages 12, 15, 17, 23).

[2] Sewall Wright. 'Correlation and Causation'. In: *Journal of Agricultural Research* 20.7 (Jan. 1921), pp. 557–585 (cited on pages 12, 15, 17, 23).

[3] Jan Tinbergen. 'Bestimmung und Deutung von Angebotskurven Ein Beispiel'. In: *Zeitschrift für Nationalökonomie* 1.5 (1930), pp. 669–679 (cited on pages 12, 23).

[4] Trygve Haavelmo. 'The probability approach in econometrics'. In: *Econometrica: Journal of the Econometric Society* 12 (1944), pp. iii–vi+1–115 (cited on pages 12, 23).

The Language of SEMs

Structural Equation Models (SEMs)

- A more “mechanistic” definition of causality
- Mathematically equivalent to potential outcomes
- SEMs define “mechanisms” or “structures” of how the world works
- These “mechanisms” allow one to understand what would be the effects of “interventions”; the causal effect
- Parameters that go into these mechanisms are “structural” parameters

- Goal: identify “structural” parameters from observed data

Example: Demand for Gasoline

- Economic theory tells us that under a log-linear (Cobb-Douglas) model of demand that (log-)demand Y as a function of (log-)price p should follow a law
$$Y(p) \coloneqq \delta p$$
- δ is the price elasticity of demand
- In practice, observed demand is random across house-holds due to un-observed factors (shocks) U
- We can then write
$$Y(p) \coloneqq \delta p + U, \qquad E[U] = 0$$

SEMs and Potential Outcomes

- The random variable $Y(p)$ that is defined by the SEM is equivalent to the "potential outcome" defined by the potential outcome framework
- The stochastic function

$$p \to Y(p)$$

- Encodes the random outcome if we intervene and set the price to p
- Expected log-demand if we intervene and set price to p is

$$E[Y(p)]$$

- The average causal effect of switching from p_1 to p_2

$$E[Y(p_1)] - E[Y(p_2)]$$

More Reasonable SEM

- Part of the shock U corresponds to observable variables X
$$Y(p) \coloneqq \delta p + \beta' X + \epsilon_Y, \qquad \epsilon_Y \perp\!\!\!\perp X$$
- The above is not a “regression” equation
- It is a model the describes all counter-factual predictions of price
- If we offer to a household a price of p, and the house-hold happens to have characteristics X, then we expect to see demand
$$\delta p + \beta' X$$
- p is better thought as an “index” describing potential values of the price; not as a random variable

Identification Question

- What data (Y, P, X) of quantities, prices and characteristics should we collect to allow us to learn the structural parameter δ?

Conditional Exogeneity

- Suppose that the observed variables (Y, P, X) satisfy that
- The outcome was generated from the structural model:

$$Y = Y(P), \qquad \text{(consistency)}$$

- The observed price P is determined "exogenously", independently of ϵ_Y conditional on X

$$P \perp\!\!\!\perp \epsilon_Y \mid X, \qquad \text{(conditional exogeneity)}$$

- Note that under the equivalence of $Y(p)$ with the potential outcomes

$$P \perp\!\!\!\perp \epsilon_Y \mid X \Rightarrow P \perp\!\!\!\perp \{Y(p)\}_{p \in R} \mid X, \qquad \text{(conditional ignorability)}$$

Conditional Exogeneity Interpretation

- Conditional on characteristics X the price of gasoline demand should be independent of the remaining house-hold shocks that determine demand per house-hold
- Believable, assuming that gasoline prices are set based on aggregate supply and demand conditions and not at the household level
- Especially if we control for geographic regions; within region price fluctuations should be independent of house-hold shocks

- Wouldn't be a good assumption if gasoline suppliers "communicate with" local households and observe signals that they use for price setting.

Identification under Conditional Exogeneity

- Under conditional exogeneity we have
$$Y = Y(P) = \delta P + \beta' X + \epsilon_Y, \qquad \epsilon_Y \perp (P, X)$$

- Structural parameters δ, β can be identified as the BLP parameters of Y using (P, X)

Structural Model of Price

- We can also further describe the mechanism that sets the (log)-price
$$P(x) \coloneqq \nu' x + \epsilon_P, \qquad E[\epsilon_P] = 0$$
- Again $P(x)$ is the "counterfactual" stochastic process that describes the price if we set the characteristics to take value "x". It is the "potential outcomes" of the price

- If we further assume that observed X are independent of shock ϵ_P
$$X \perp\!\!\!\perp \epsilon_P$$
- Then "structural" parameters ν are identified as the BLP of P using X

Triangular Structural Equation Model (TSEM

$$\boxed{Y} := \delta P + \beta' X + \boxed{\epsilon_Y}$$
$$\boxed{P} := \nu' X + \boxed{\epsilon_P} \qquad , \qquad \epsilon_Y, \epsilon_P, X \text{ mutually independent}$$
$$\boxed{X,}$$

Endogenously determined by the structural model

Exogenously determined "outside" of the model

What do we mean by Structural

- The equations are “structural” if they allow us to answer:
  - **Comparative static questions:** how endogenous variables change in response to changes in exogenous variables
  - **Counterfactual questions:** how endogenous variables change if we “set” or “fix” some of the Right-Hand-Side variables to particular values
- To answer “setting” or “fixing” queries we just write:

$$Y(p,x) \coloneqq \delta p + \beta' X + \epsilon_Y$$
$$P(x) \coloneqq \nu' X + \epsilon_P$$

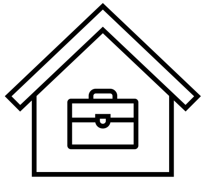
- The model is invariant to changes in the distributions of the exogenous variables $\epsilon_Y, \epsilon_P, X$

SEMs: *more "mechanistic" version of potential outcomes*

Allow us to determine the *explicit structural mechanisms* that give rise to the stochastic potential outcomes.

Parse out which parts of the generative process is *endogenous* and which part is *exogenously* driven.

Structural component motivated by a “model” and *invariant to shifts* in the exogenous components

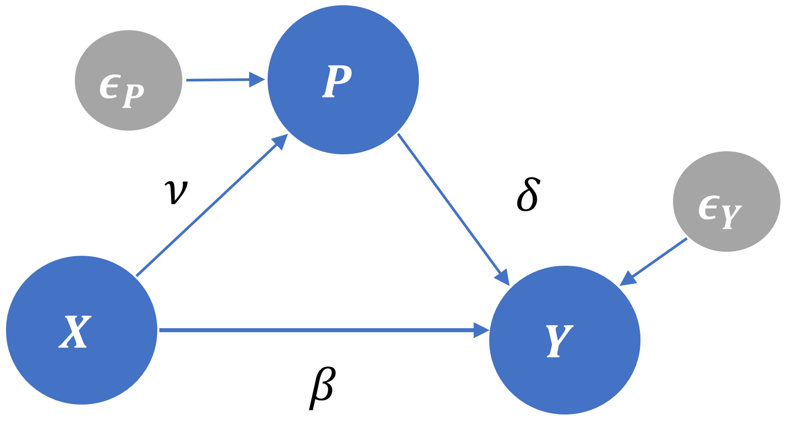


Structural component allows to answer *counterfactual queries*, under *interventions on endogenous variables*

Visualizing SEMs

SEMs as Causal Diagrams

- The graph starts at $\epsilon_P, \epsilon_Y, X$, aka root nodes
- The absence of links between the root nodes and any other node implies their un-correlation/exogeneous determination
- X, ϵ_P are parents of P
- P, X, ϵ_Y are parents of Y
- Y is a "collider" in the path $P \rightarrow Y \leftarrow X$ as it has two arrows pointing towards Y

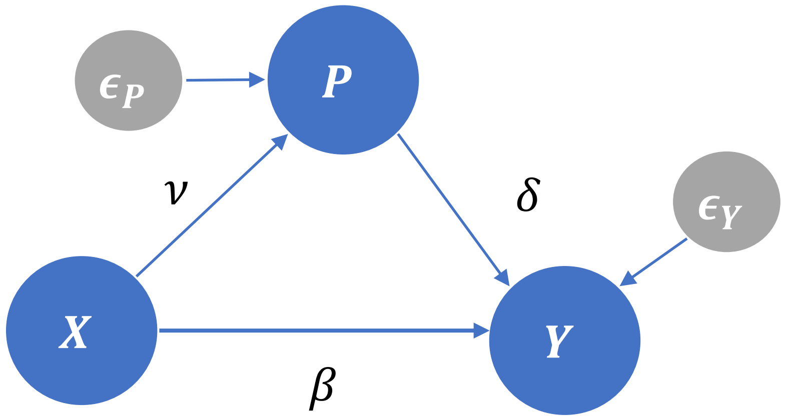


Back-door paths

- The quantity of interest is δ, the "structural causal effect" of P on Y
- The are two paths from P to Y

$$P \to Y, \qquad P \leftarrow X \to Y$$

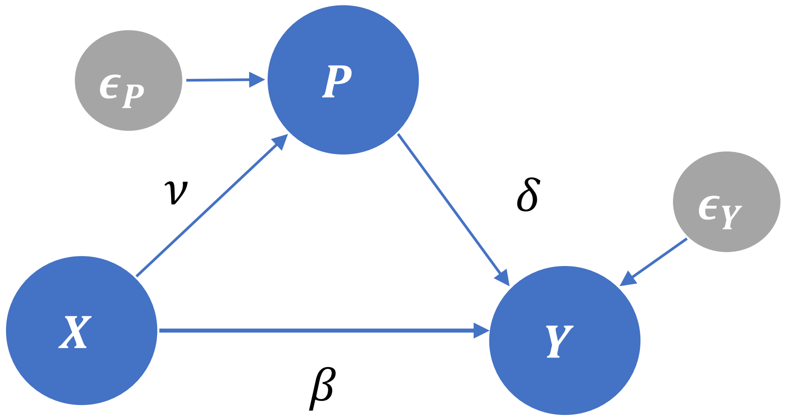
- The second path is called a "back-door path"; there is an arrow pointing back to P from X
- X is a "common cause" of P, Y
- Controlling/Adjusting for X "closes the back-door path"



Direct and Indirect Effects

- X affects Y in two ways
- Direct effect β from path $X \rightarrow Y$
- Indirect effect $\nu\delta$ from path $X \rightarrow P \rightarrow Y$
- Indirect effect is "mediated" by P
- The total effect of X on Y is
$$\beta + \nu\delta$$

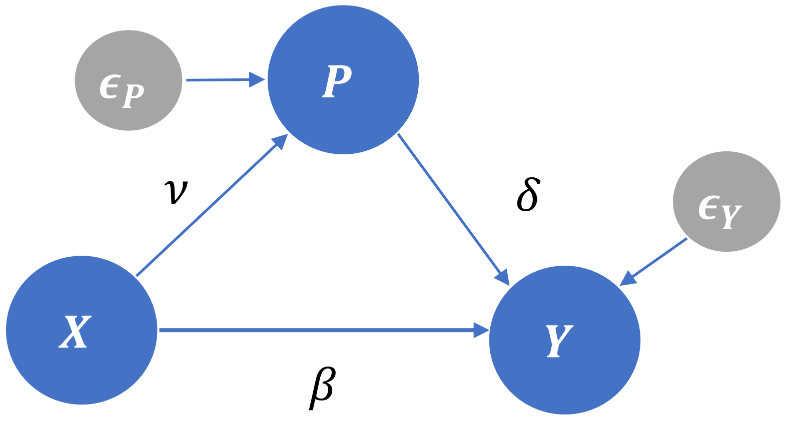
- δ, β is identified by $Y \sim P, X$
- ν is identified by $P \sim X$
- $\beta + \nu\delta$ is identified by $Y \sim X$



Direct and Indirect Effects

- The total effect of X on Y is
$$\beta + \nu\delta$$
- $\beta + \nu\delta$ is identified by $Y \sim X$

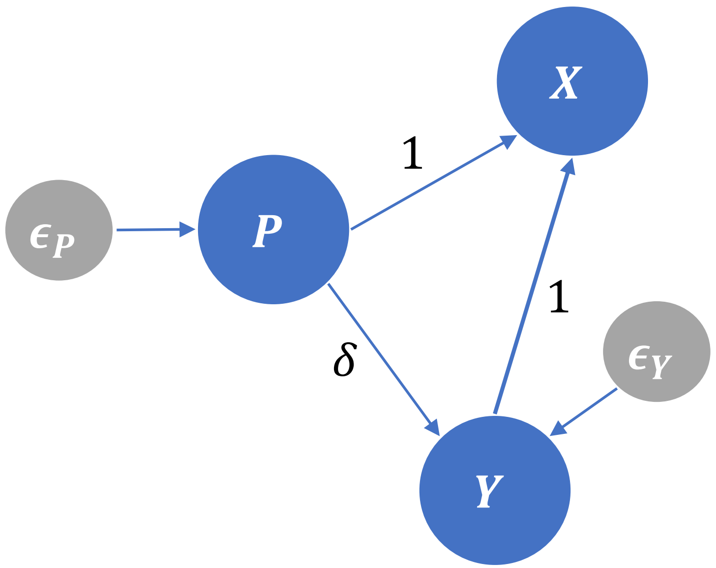
- By TSEM:
$$Y = \delta P + \beta' X + \epsilon_Y = (\delta\nu + \beta)' X + \delta\epsilon_P + \epsilon_Y$$
- Both errors are orthogonal to X
$$\delta\epsilon_P + \epsilon_Y \perp X$$
- Visual verification: no common-cause of X, Y in the Causal Diagram



Causal Diagram representations of a SEM allow us to understand the different “paths of influence”, and how to identify different structural parameters

Conditioning Gone Wrong

Colliders



Colliders

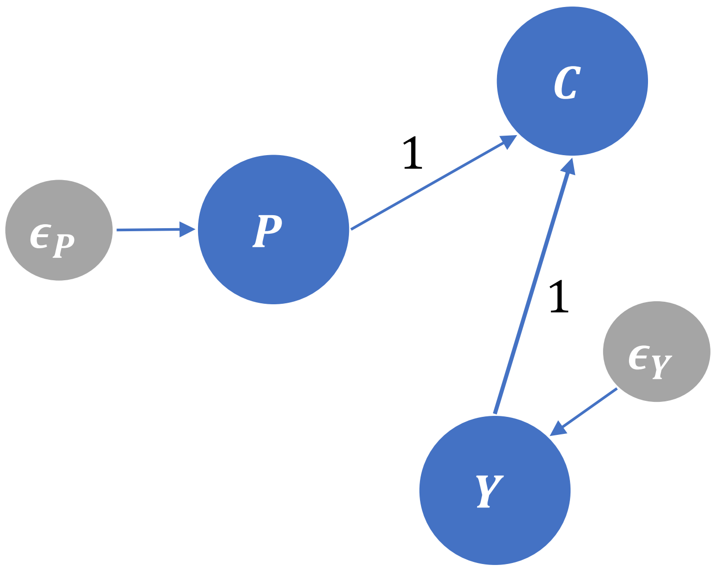
- Consider the SEM

$$Y := \epsilon_Y$$
$$P := \epsilon_P$$
$$C := Y + P + \epsilon_C$$

- All exogenous shocks drawn $N(0,1)$
- $Y \perp\!\!\!\perp P$ and BLP coefficient of $Y \sim P$ is zero
- If we condition on C

$$E[Y|P,C] = \frac{C-P}{2}$$

- BLP of $Y \sim P, C$ has coefficient $-½$, which is wrong



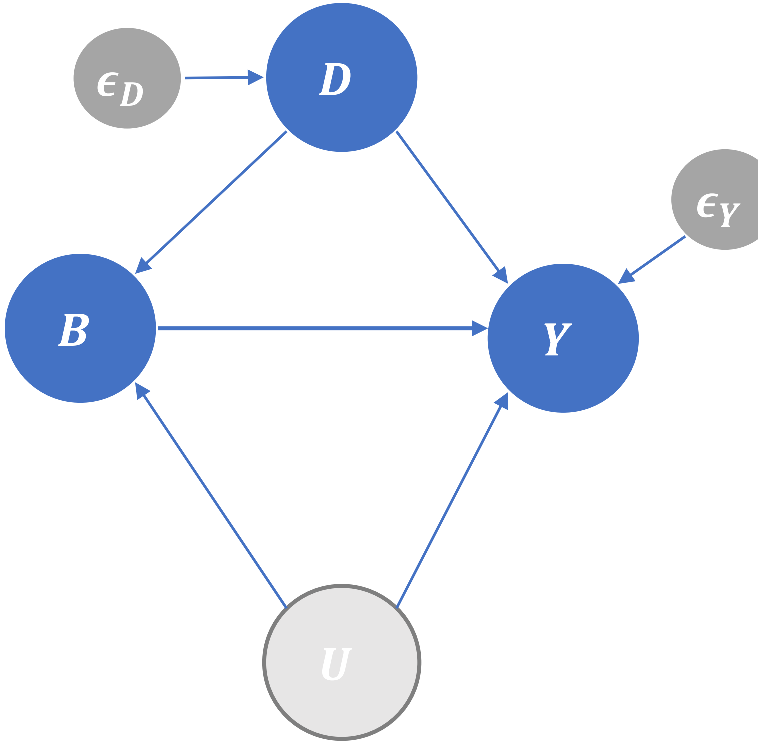
Birth Length Paradox

- Infants born to smokers have higher infant mortality than non-smokers
- If one conditions on infants with low birth-weight, effect is reversed!
- Many controversies in epidemiology [Hernandez et al, Am. J. of Epidemiology, 06]

- Let’s look at it from a SEM/Causal Diagram perspective

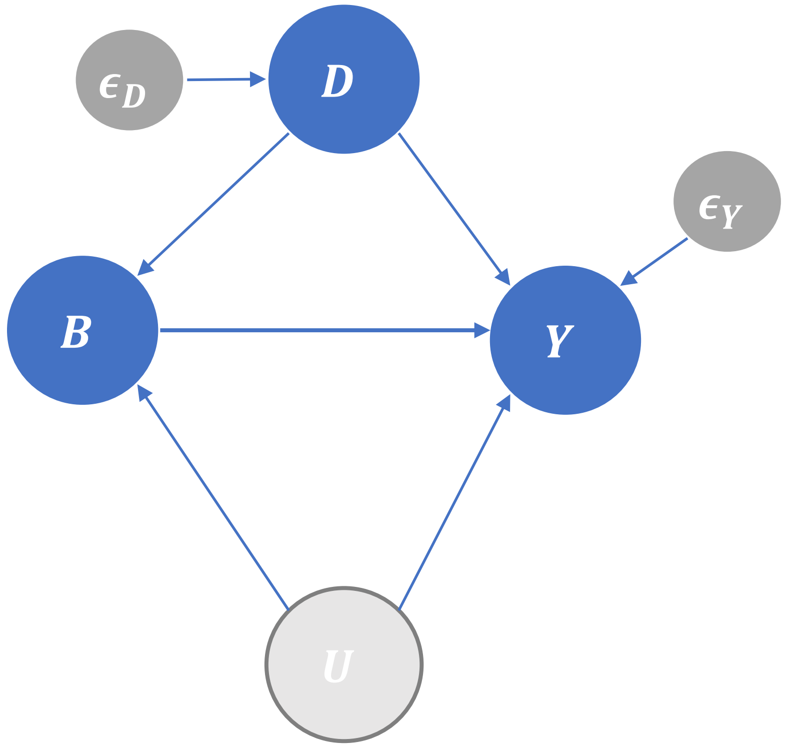
Birth Length Paradox

- Y=infant mortality, D=smoking parent, B=birthweight
- B is a variable that arises after D, and can be influenced by it
- B is an "outcome" or "endogenous" variable
- B is also affected by "other competing risks"
- Those competing risks also have an effect on outcome; increase mortality



Birth Length Paradox

- Y=infant mortality, D=smoking parent, B=birthweight
- If B is low and D is smoking, then most probably U does not occur (i.e. there are no other competing risks that lead to low birthweight)
- But if B is low and D is non-smoking, then most probably U occurs (i.e. the low birthweight was most probably caused by competing risks)
- Conditioning on B=low, is as if we are comparing infants from smokers with no other competing risks, to infants from non-smokers with competing risks



Birth Length Paradox

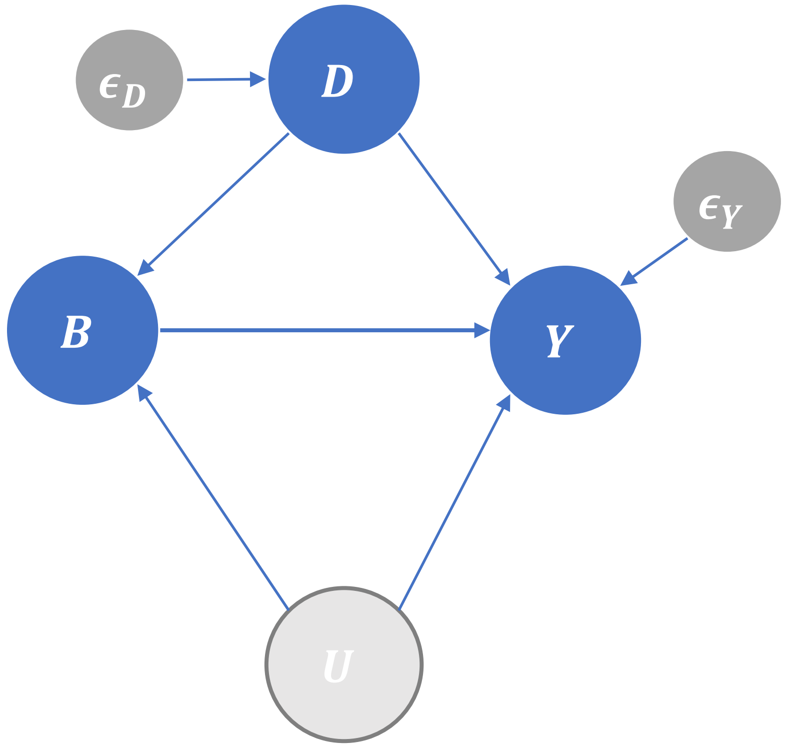
- A simple SEM

$$Y := D + B + \kappa U + \epsilon_Y$$
$$B := D + U + \epsilon_B$$
$$D := \epsilon_D$$

- Suppose all ϵ's are $N(0,1)$

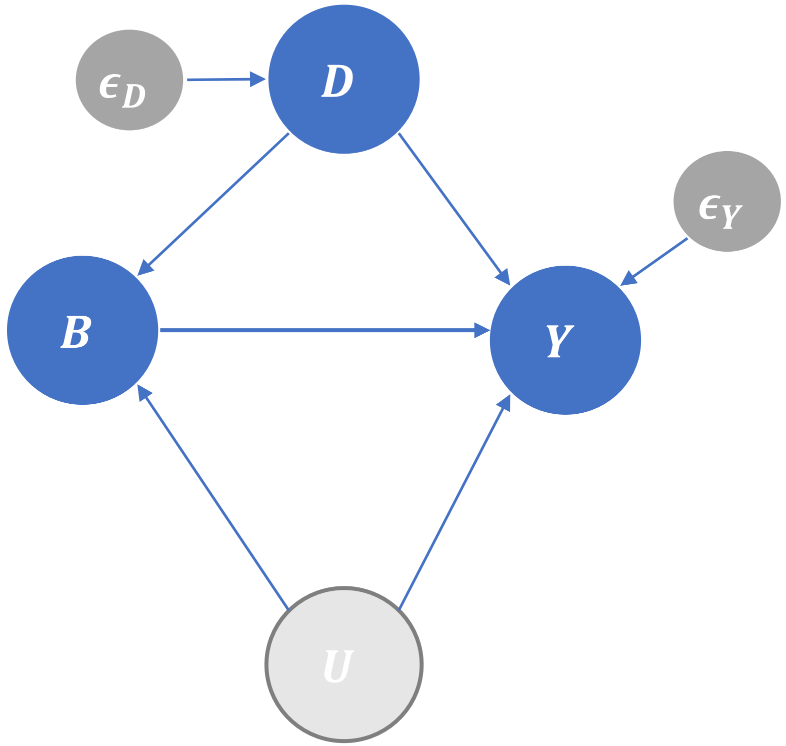
$$E[Y|B,D] = D + B + E[U|B,D]$$
$$= D + B + \kappa \frac{B - D}{2}$$

- If κ is large then coefficient of D reverses sign
- Smoking "decreases" infant mortality



Birth Length Paradox: Visually

- Conditioning on B "opens" a collider path from D to U (now D and U are correlated)
- Since U is also connected to Y
- There is an "open path of influence" from D to Y other than the direct path
- It is the path $D \to B \leftarrow U \to Y$
- The effect we are measuring is the "total effect" from these two paths
- Not the direct effect we were looking for



We should not always be conditioning/adjusting for any variable that we have access to.

Some variables (especially those that are “outcomes”) can introduce heavy biases

SEMs and Causal Diagrams is the tool that lets us deduce what is the right adjustment set

Direct and Indirect Effects for Analysis of Wage Discrimination

A Structural Equation Model of Wage Discrimination

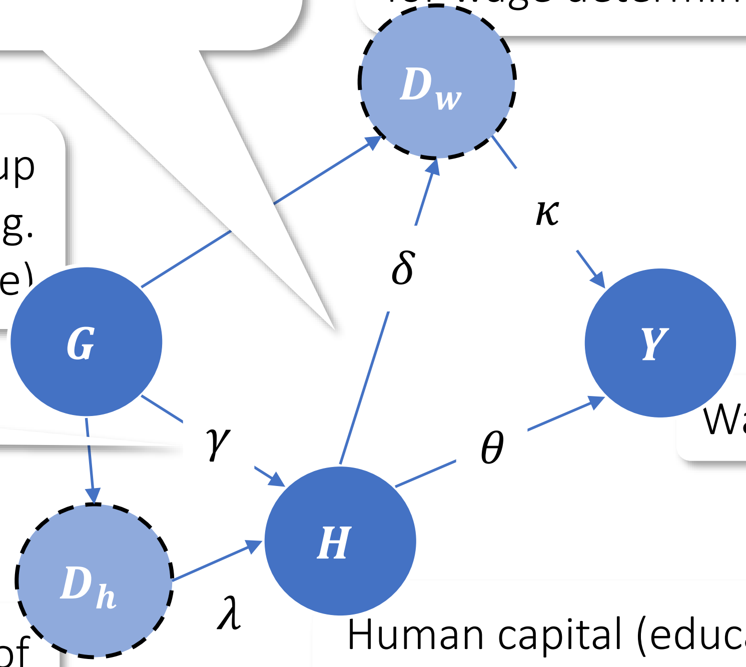
$$
\begin{gathered}
Y := \kappa D_w + \theta H + \epsilon_Y \\
D_w := G + \delta H + \epsilon_{D_w} \\
H := \gamma G + \lambda D_h + \epsilon_H \\
D_h := G + \epsilon_{D_h} \\
G,
\end{gathered}
$$

Absence of arrow $G \rightarrow Y$ indicates there is no inherent differentiation in productivity level driven by G; only due to discrimination or differences in human capital

Abstraction Variable of "Discriminatory Behavior" for wage determination

Sensitive Group Attribute (e.g. sex, race)

Existence of arrow $G \rightarrow H$ indicates there is potential heterogeneity in preferences for human capital acquisition (e.g. inherent preferences over different occupations)



Wage

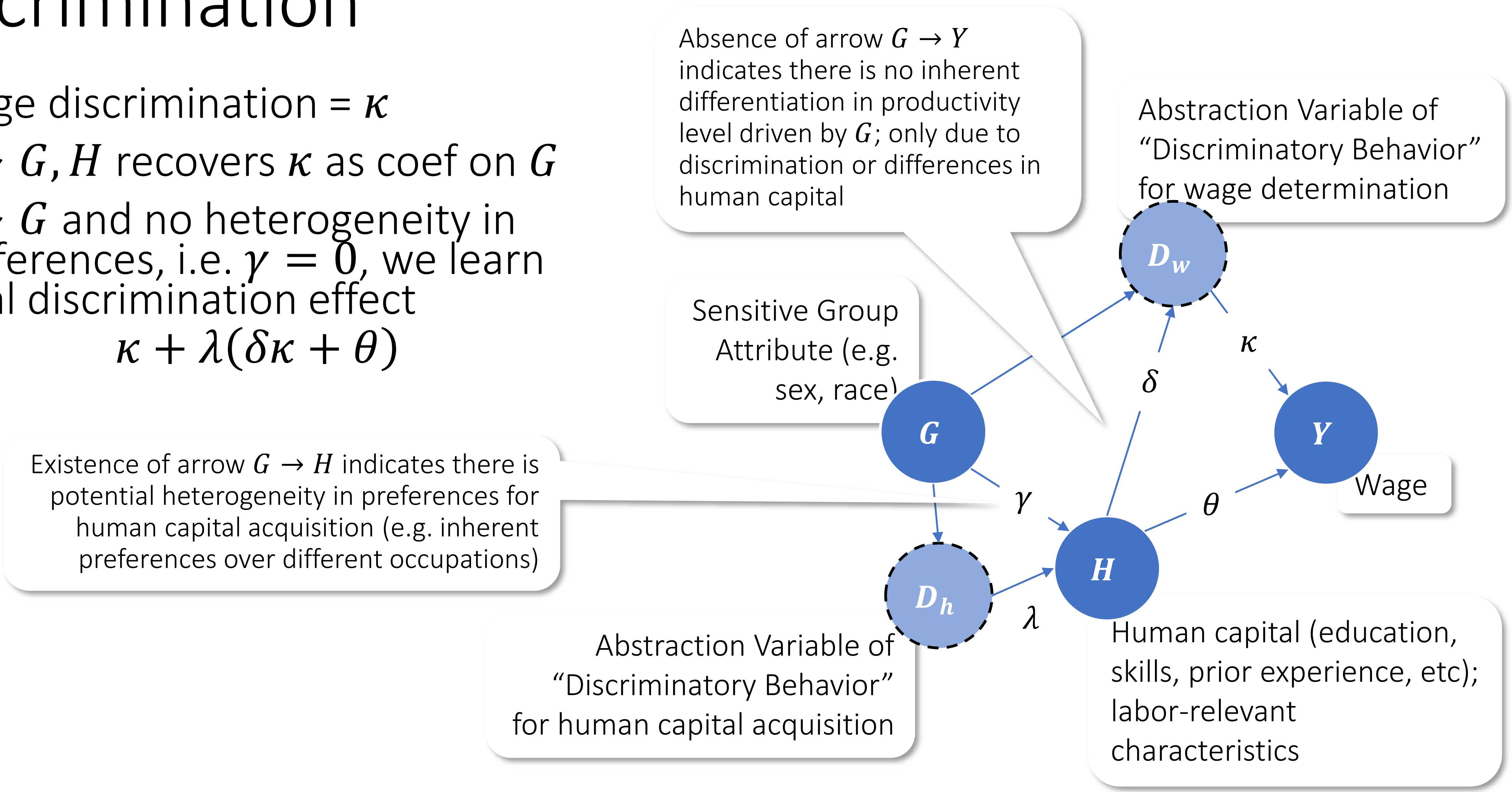
Abstraction Variable of "Discriminatory Behavior" for human capital acquisition

Human capital (education, skills, prior experience, etc); labor-relevant characteristics

A Structural Equation Model of Wage Discrimination

- Wage discrimination = κ
- $Y \sim G, H$ recovers κ as coef on G
- $Y \sim G$ and no heterogeneity in preferences, i.e. $\gamma = 0$, we learn total discrimination effect

$$\kappa + \lambda(\delta\kappa + \theta)$$



SEMs and Causal Diagrams let us disentangle the difference flows of “causal influence” and focus on the ones that we care to identify.