Visualizing Environmental Econometrics

Sharing DAGs to communicate causal inference logic

Pierce Donovan

Department of Economics, Connecticut College

EEA Annual Meeting (AERE Sessions)
February 2023

When we lack the facilities to present causal knowledge, we end up approaching causal inference blindfolded.

We should share our mental models of our DGPs to help justify whether or not our regression output deserves a causal blessing.

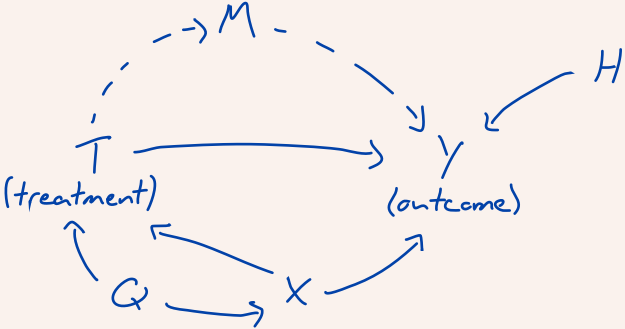
[†] Netflix, *Bird Box* (2018)

DAGs help us share causal info

identifying variation:

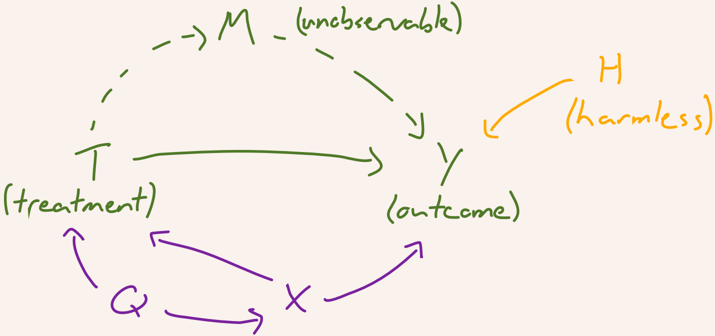
- variables and relationships of interest
- setting in which data are collected
- research design (variation selection)

A Directed Acyclic Graph is a visualization of an observed, structured process. It represents our best understanding of the various causal and spurious links between variables of interest.



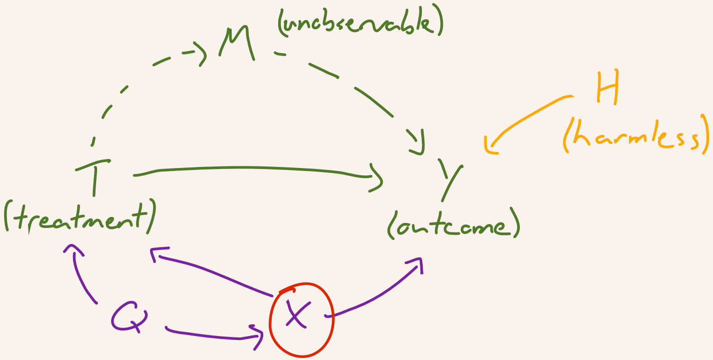
We want to focus on the *links* (edges) between nodes. They tell a story about our DGP, inform research design and presentation, and **validate claims of causal identification in regression output**.

Any directed paths that “leave” the independent variable of interest (T) are potentially-causal {green}. A path that “enters” the independent variable is a non-causal *back-door* path if it also links up with the dependent variable (Y) {purple}.



A confounding variable creates spurious correlations (non-causal paths) between two variables when omitted from a regression.

A path is *closed* once we control for/condition on/hold constant at least one variable along that path. When reporting regression output, $\hat{\beta}$ reflects all remaining *open* paths from T to Y.

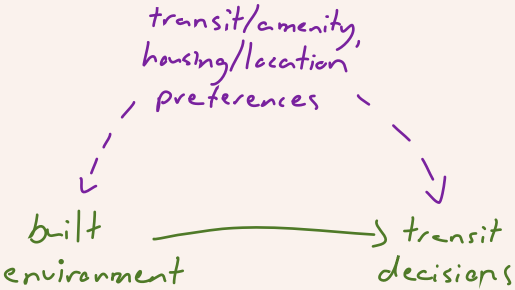


Can you determine a viable control strategy for measuring $\underline{T \rightarrow Y}$?

The “Big Four” identification methods

(1) Matching Estimators: “closing your backdoors”

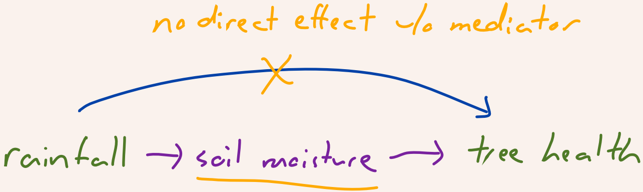
Regression automatically isolates the non-overlapping variation between regressors. It is our responsibility to determine whether or not this variation generates a measurement of a causal effect.



Illustrative example: Does the local built environment induce changes in transportation modality? (With a naïve regression, wouldn't transit and amenity preferences be confounders?)

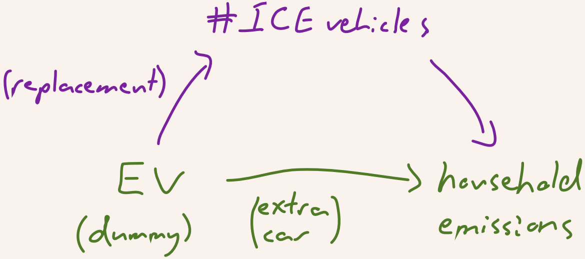
Matching is more complicated than we think. Some researchers include "controls" only because they are available. Many of these decisions are made without any causal logic to back them up.

Illustrative example: *How* do droughts impact tree health?



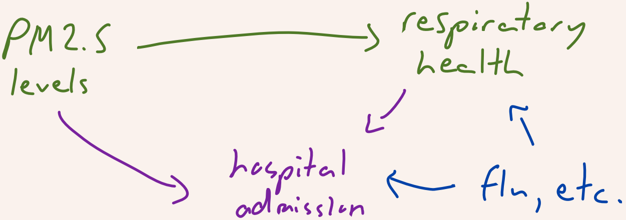
Over-conditioning: eliminating the variation that you wanted to exploit. Rainfall can't "reach" tree health without impacting soil moisture—but controlling for soil moisture closes this path.

Another mismatched example: Will adopting electric vehicles decrease household emissions? A gut reaction might be to control for the number of ICE cars in the household, but this blocks us from measuring the effect of replacing an ICE car with an EV.



Controlling for a *mediator* closes a causal path (replacement); our regression misses out on [the core] part of the treatment effect.

Some control strategies can inadvertently bias estimators. For example, what if we want to estimate the impact of high PM2.5 levels on respiratory health, but use data from hospital admissions?

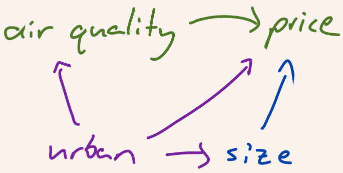


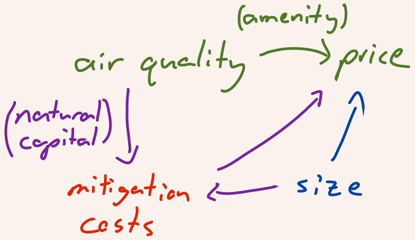
selection, in this case.

Some variables create spurious correlations between treatment and control only when they are controlled for. Stratifying on a *collider* variable will open an otherwise closed path.

Your story (and its DAG representation) will inform whether or not a proposed control strategy is viable, and if regression coefficients are causally-interpretable.

OLS: $house\_price_i = \alpha + \beta \cdot air\_quality_i + \gamma \cdot \underset{MC}{\cancel{urban}}_i + \delta \cdot size_i + \varepsilon_i$





Controlling for "urban" is sufficient.

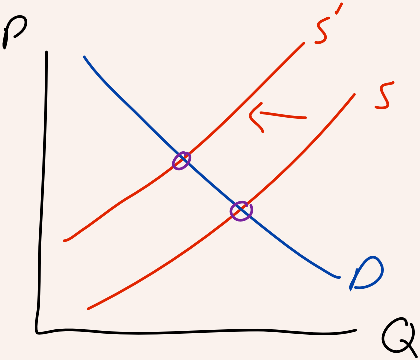
Controlling for mitigation pathway relates air quality to size of home.

Statistical significance is meaningless without a causal story. A better robustness check would test different *causal* hypotheses.

(2) Instrumental Variables: “being weird with colliders”

How do you measure the price elasticity of demand for fish (Graddy, 2006)? You observe several [counterfactual] price-quantity pairs along a single demand curve.

i.e. $\varepsilon = \left.\frac{\%\Delta Q}{\%\Delta P}\right|_D$

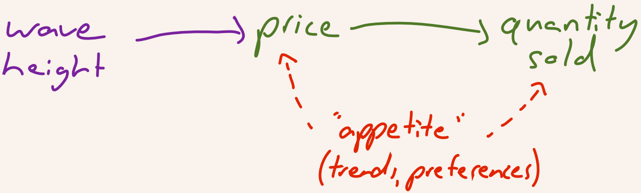


Two regressions, one fishy, one sound.

Dependent variable:	log(quantity sold)	
Independent variable	OLS	2SLS
log(price)	-0.549 (0.184)	-0.960 (0.406)
Monday	-0.318 (0.227)	-0.322 (0.225)
Tuesday	-0.684 (0.224)	-0.687 (0.201)
Wednesday	-0.535 (0.221)	-0.520 (0.219)
Thursday	0.068 (0.251)	0.106 (0.232)
Time trend	-0.001 (0.003)	-0.003 (0.003)
First stage for log(price)		
Wave height (feet)		0.103 (0.022)
F statistic for IV		22.638

Graddy used the maximum wave height during recent fishing trips as an *instrument* for the price of fish at the Fulton fish market. More difficult fishing conditions drive a supply-side price shock.

$$IV : \begin{cases} price_{dt} = \pi + \underline{\lambda} \cdot wave\_height_{dt} + \gamma_d + \delta_t + \vartheta_{dt} \\ quantity\_sold_{dt} = \alpha + \underline{\eta} \cdot wave\_height_{dt} + \theta_d + \phi_t + \varepsilon_{dt} \end{cases}$$

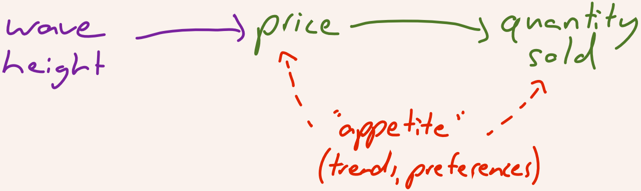


Our DAG turns price into a *collider*, which closes the spurious path between price and quantity! The coefficients $\hat{\lambda}$ and $\hat{\eta}$—and $\underline{\hat{\beta}_{IV}} = \hat{\eta}/\hat{\lambda}$—now have a causal interpretation. ($\hat{\beta}_{2SLS}$ works too.)

The IV DAG reveals three identification assumptions which can be communicated transparently and evaluated with a cursory glance.

Our instrument must satisfy three assumptions:

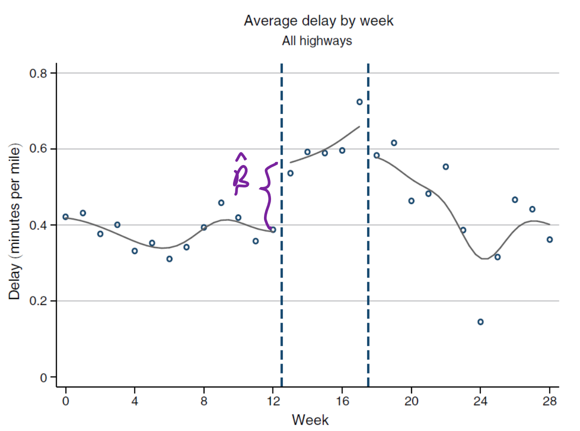
1. *Relevancy* (i.e. have a causal impact on price)
2. *Independence* (i.e. no relation to demand shifts)
3. *Validity*[*] (i.e. no impact on quantity sold except via price)



[*] What do we think about the exclusion restriction here?

(3) Regression Discontinuity: “finding your offensive linemen”

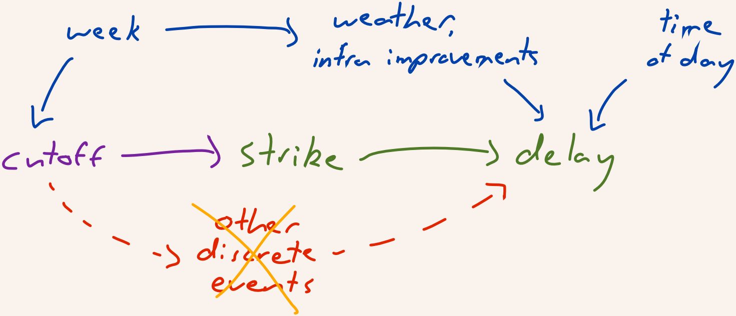
An LA transit worker strike resulted in huge increases in congestion on roadways (Anderson, 2014). What makes us so sure?



(because nothing else caused a jump in delays around the same time)

A running variable controls for the "smooth" variation in delay. Since the strike date is a function of *week*, all *continuous* [time-varying] BDPs are closed, leaving the discontinuity for $\hat{\beta}$.

OLS: $delay_t = \alpha + \beta \cdot strike_t + \gamma \cdot f(week_t) + \delta \cdot strike_t \cdot f(week_t) + \varepsilon_t$

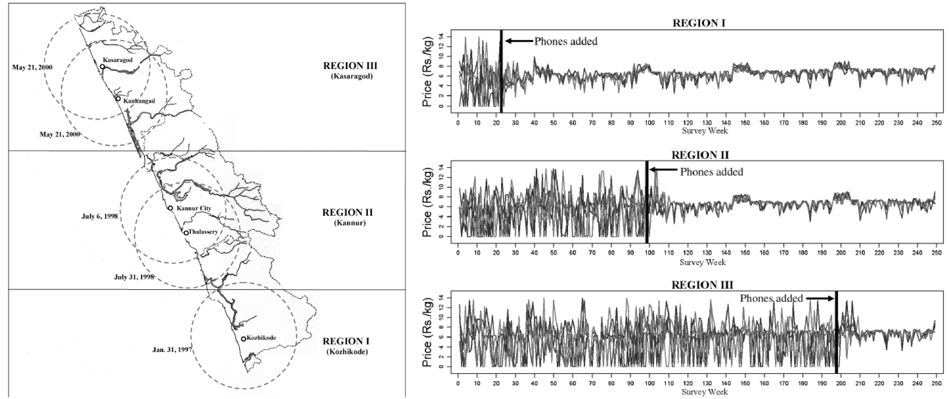


Because we want to attribute $\hat{\beta}$ solely to the strike, we assume that there are no other causal paths concurrent with the strike.

(4) Differences-in-Differences: "parallel trends and synecdochic control"

(fixed-effects)

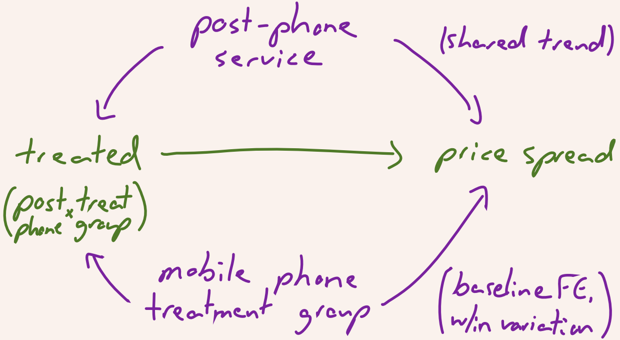
The introduction of cell phones in Kerala seriously reduced price dispersion in fish markets (Jensen, 2007). How do we know?



(b/c these coastal towns were on similar trajectories w/o intervention)

When treatments perfectly correlate with time, we cannot control for time without eliminating all variation in the treatment variable. Our solution? Give the treatment group a friend to compare to.

OLS: $spread_{it} = \alpha + \beta \cdot treat_i + \gamma \cdot post_t + \delta \cdot treat_i \cdot post_t + \varepsilon_{it}$


post-phone service
(shared trend)
treated
(post phone × treat group)
price spread
mobile phone treatment group
(baseline FE, w/in variation)

The DiD DAG suggests a *parallel trends* assumption: w/o treatment, the two groups would move in-step (effect attribution).

Thank you!

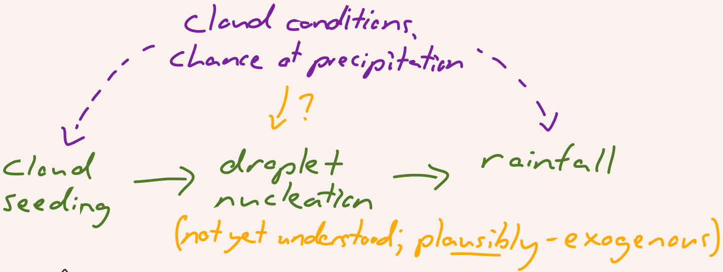
pdonovan@conncoll.edu

[†] Sony Pictures, *Snatch* (2000)

(5?) The Front-Door Criterion: “causality and the chain rule”

In the presence of confounders, we can decompose a non-causal relationship into two separate causally-identified measurements.

$$SUR: \begin{cases} \mathit{nucleation}_i = \pi + \lambda \cdot \mathit{seeded}_i + \vartheta_i \\ \mathit{rainfall}_i = \alpha + \gamma \cdot \mathit{nucleation}_i + \delta \cdot \mathit{seeded}_i + \varepsilon_i \end{cases}$$



For $\underline{\hat{\beta}_{FDC}} = \underline{\hat{\lambda} \cdot} \hat{\gamma}$ to be causal, the FDC DAG suggests three new assumptions: (1) seeding only impacts rainfall via nucleation, (2) seeding and nucleation are not confounded, (3) rainfall and nucleation are not confounded after controlling for seeding.