A statistical method for estimating (specific) causation in the law

Maria Cuellar
University of Pennsylvania
Department of Criminology

May 19, 2021

Speaking at: Northwestern University
Department of Statistics

Outline

1. **Legal example**: Monsanto's herbicide, Roundup
2. **Causal parameter**: Probability of causation
3. **Estimation**: Nonparametric influence-function-based projection
4. **Application**: Public health, bacteria in water in Kenya

Legal example: Roundup

1. Dewayne Johnson, born in 1972, worked as a school groundskeeper from 2012 to 2014, where he was **exposed to Roundup, an herbicide**.

2. Dewayne developed Non-Hodgkins Lymphoma (**terminal cancer**) in 2014.

3. Dewayne sued Monsanto. He wants to prove that his **cancer was caused by his exposure** to Roundup.

Photograph: Reuters, August 10, 2018

What statement should Dewayne's attorney make?

What statement should Dewayne's attorney make?

1. Given that a man is exposed to Roundup, it is likely that he will develop cancer. **Forecasting**

2. Given that a man developed cancer, it is likely that he was exposed to Roundup. **Backcasting**

3. Given that a man was exposed to Roundup and developed cancer, it is likely that the exposure to Roundup, and not something else, caused the cancer. **Attribution**

Dawid et al. (2016)

Attribution formalized as the probability of causation

Y: binary outcome (e.g. 1: cancer, 0: no cancer)

A: binary exposure (e.g. 1: exposed to Roundup, 0: not exposed.)

Y^a: potential outcome if we had set A=a.

X: vector of covariates about individual (e.g. gender, age, etc.)

The **probability of causation** is defined as

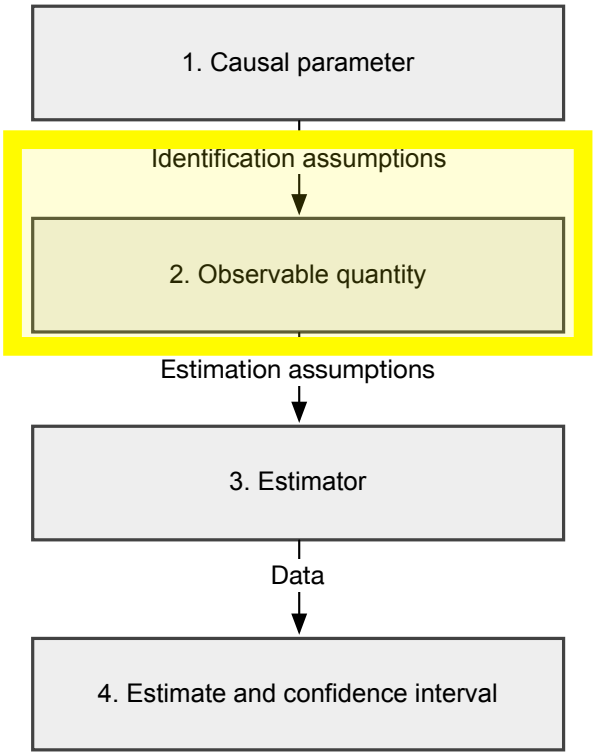
$$PC(\mathbf{x}) = P(Y^0=0 \mid Y=1, A=1, \mathbf{X=x}).$$

Robins et al. (1988, 89)
Pearl et al. (2009)
Dawid et al. (2016)

How to estimate PC?

Maria Cuellar: A nonparametric estimator for the probability of causation

Estimating the probability of causation



Identification assumptions

1. Binary outcome and treatment: $Y \in \{0, 1\}, A \in \{0, 1\}$
2. Consistency: $A = a \implies Y = Y^a$
3. No unobserved confounders: $Y^a \perp\!\!\!\perp A|X$
4. Monotonicity: $Y^1 \geq Y^0$

Identification assumptions

1. Binary outcome and treatment: $Y \in \{0,1\}, A \in \{0,1\}$
2. Consistency: $A = a \implies Y = Y^a$
3. No unobserved confounders: $Y^a \perp\!\!\!\perp A|X$
4. Monotonicity: $Y^1 \geq Y^0$

Under these assumptions:

$$PC \underbrace{=}_{\text{Law of tot. prob}} \frac{P(Y^0=0,Y^1=1|A=1,X)}{P(Y^1=1|A=1,X)} \underbrace{=}_{1,4} \frac{P(Y^1-Y^0=1|A=1,X)}{P(Y^1=1|A=1,X)} \underbrace{=}_{1} 1 - \frac{\mathbb{E}(Y^0|A=1,X)}{\mathbb{E}(Y^1|A=1,X)} \underbrace{=}_{2,3} 1 - \frac{\mathbb{E}(Y|A=0,X)}{\mathbb{E}(Y|A=1,X)} = 1 - RR.$$

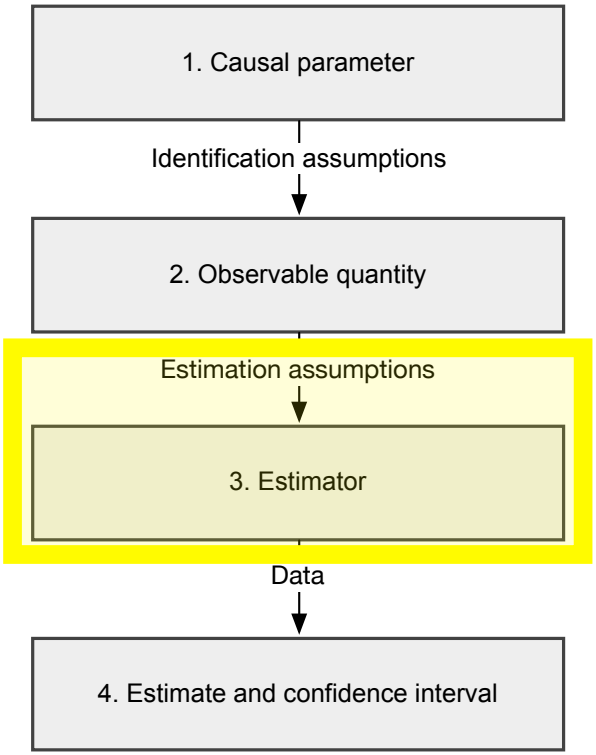
Identification assumptions

1. Binary outcome and treatment: $Y \in \{0, 1\}, A \in \{0, 1\}$
2. Consistency: $A = a \implies Y = Y^a$
3. No unobserved confounders: $Y^a \perp\!\!\!\perp A|X$
4. Monotonicity: $Y^1 \geq Y^0$

Under these assumptions:

$$PC(\boldsymbol{x}) = 1\text{–}RR(\boldsymbol{x}).$$ **Observable quantity**

Estimating the probability of causation



How should we estimate risk ratio here?

Plugin estimator

$$\widehat{PC}_{\text{PI}}(x) = 1 - \widehat{RR}(x) = 1 - \frac{\widehat{\mathbb{E}}(Y \mid A = 0, X = x)}{\widehat{\mathbb{E}}(Y \mid A = 1, X = x)}$$

1. **Parametric plug-in estimator**: Pro: Allows for confidence intervals. Con: Wrong parametric assumptions could have **serious consequences** for PC.

2. **Nonparametric plug-in estimator**: Pro: No parametric assumptions. Con: No valid confidence intervals in general and slow convergence rates.

3. **Influence-function-projection-based estimator**: Pro: No parametric assumptions, *and* valid confidence intervals under weak structural conditions (Cuellar and Kennedy, 2020).

Cuellar–Kennedy estimation

- **Want an influence function** for parameter of interest, but for nonparametric case, 1–RR does not have an influence function because it is not pathwise differentiable.
- **1. Project the data** onto a parametric model.
- **2. Derive influence function** of projection.
- **3. Derive estimator** by using this influence function.
- **4. Derive valid confidence intervals (e.g. 95%)**.
  - *Requirement: The errors of the nuisance estimators must converge at a certain (slow) rate, but can be estimated nonparametrically.*

Cuellar–Kennedy estimation

- **1. Projection —** Weighted least squares projection of $\gamma(\boldsymbol{X}) = 1 - RR(\boldsymbol{X})$ onto a parametric model $g(\boldsymbol{X}; \beta)$:

$$\beta = \text{argmin}_{\beta} E\left\{w(\mathbf{X})(\gamma(\mathbf{X}) - g(\mathbf{X};\beta))^2)\right\},$$

  where β: vector of parameters, $w(\boldsymbol{X})$: user-specified weight function, $\gamma(\boldsymbol{X})$: true function.

- **2. Influence function —** Next slide.

- **3. Estimator —** Our proposed estimator is defined as the value $\hat{\beta}$ that satisfies $\hat{\Psi}(\hat{\beta}) = 0$, where

$$\hat{\Psi}(\hat{\beta}) = P_n\left\{\varphi(Y, A, X; \beta, \mu_0, \mu_1, \pi)\right\}.$$

- **4. Confidence intervals —** In two slides.

Influence function

Theorem 3.1. *Under a nonparametric model, the (uncentered) efficient influence function for the moment condition $\Psi(\beta^*)$ at any fixed β^* is given by*

$$\varphi(X, A, Y; \beta^*, \mu_0, \mu 1; g) = h(X; \beta^*) \left[\frac{1}{\mu_1(X)} \left(\frac{\mu_0(X)}{\mu_1(X)} \frac{A(Y - \mu_1(X))}{\pi(X)} - \frac{(1 - A)(Y - \mu_0(X))}{(1 - \pi(X))} \right) + 1 - \frac{\mu_0(X)}{\mu_1(X)} - g(X; \beta^*) \right], \tag{11}$$

where $h(x; \beta) = \frac{\partial g(x;\beta)}{\partial \beta} w(x)$, and $\eta = (\pi, \mu_0, \mu_1)$ are the nuisance parameters.

Influence function of the projection

Confidence intervals

If $\left(\|\hat{\pi} - \pi\| + \|\hat{\mu}_1 - \mu_1\|\right) \sum_a \|\hat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1/\sqrt{n})$. **Requirement for valid CIs**

Then the proposed estimator attains the nonparametric efficiency bound, and is asymptotically normal with

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow N\left(0, M^{-1}\mathbb{E}(\varphi\varphi^T)(M^T)^{-1}\right), \qquad \text{where } \hat{M} = \mathbb{P}_n(\tfrac{\partial\hat{\varphi}}{\partial\beta})$$

and similarly for any fixed x we have

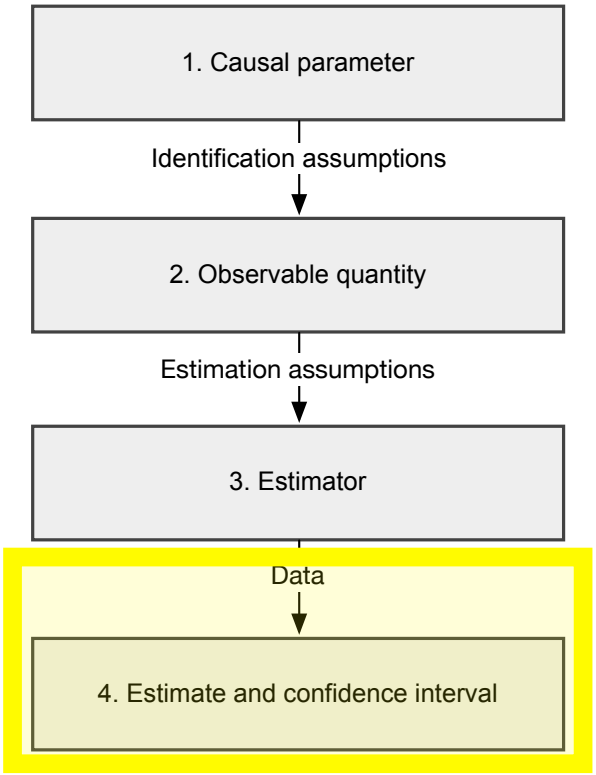
Variance σ²

$$\sqrt{n}\left(g(x;\hat{\beta}) - g(x;\beta)\right) \rightsquigarrow N\left(0, \left(\frac{\partial g(x;\beta)}{\partial\beta}\right)^T M^{-1}\mathbb{E}\left(\varphi\varphi^T\right)(M^T)^{-1}\left(\frac{\partial g(x;\beta)}{\partial\beta}\right)\right)$$

Then a simple Wald-style 95% confidence interval for $g(x;\hat{\beta})$ is given by

$$\left[g(x;\hat{\beta}) - 1.96\frac{\hat{\sigma}(x)}{\sqrt{n}}, g(x;\hat{\beta}) + 1.96\frac{\hat{\sigma}(x)}{\sqrt{n}}\right]$$ **CIs**

Estimating the probability of causation



Randomized controlled trial in Kenya

- **Data:** From JPAL (more in next slide).
- **Location:** Rural Busia and Butere-Mumias districts of Kenya's Western Province.
- **Problem:** 90% of households access spring water, 72% of water collection trips are to springs, and water springs have high fecal contamination as measured by *E. coli* in collected water.
- **Their question:** Does protecting the water spring (sealing off the source of the spring) decrease bacteria in water and child diarrhea? (ATE).
- **Results from RCT:** Spring protection reduces fecal contamination at spring by 66% and diarrhea for children under age 3 (at baseline or born since the baseline survey) by 25%.

Data: Kremer, Michael; Leino, Jessica; Miguel, Edward; Peterson, Alix, 2015, "Replication data for: Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions", https://doi.org/10.7910/DVN/28063, Harvard Dataverse, V2
Photographs: https://www.povertyactionlab.org/evaluation/cleaning-springs-kenya

- **My question: (Attribution)** For the children who were exposed to high concentration of bacteria and had diarrhea, did the exposure cause the diarrhea, or was it something else?

Data analysis

Sample:

- **Participants —** *n* = 2933 children, selected by following same procedure as Kremer et al.
- **Ages —** [0, 3] with mean of 1.6 years old.
- **Gender —** 50.3% female and 49.6% male.
- **Randomization —** on 184 springs. 7-8 households per spring (representative sample that regularly used each spring), 3-4 children per household.

Estimation:

- **Parametric model for projection —** logistic model.
- **Nuisance estimation —** using random forests and data splitting.

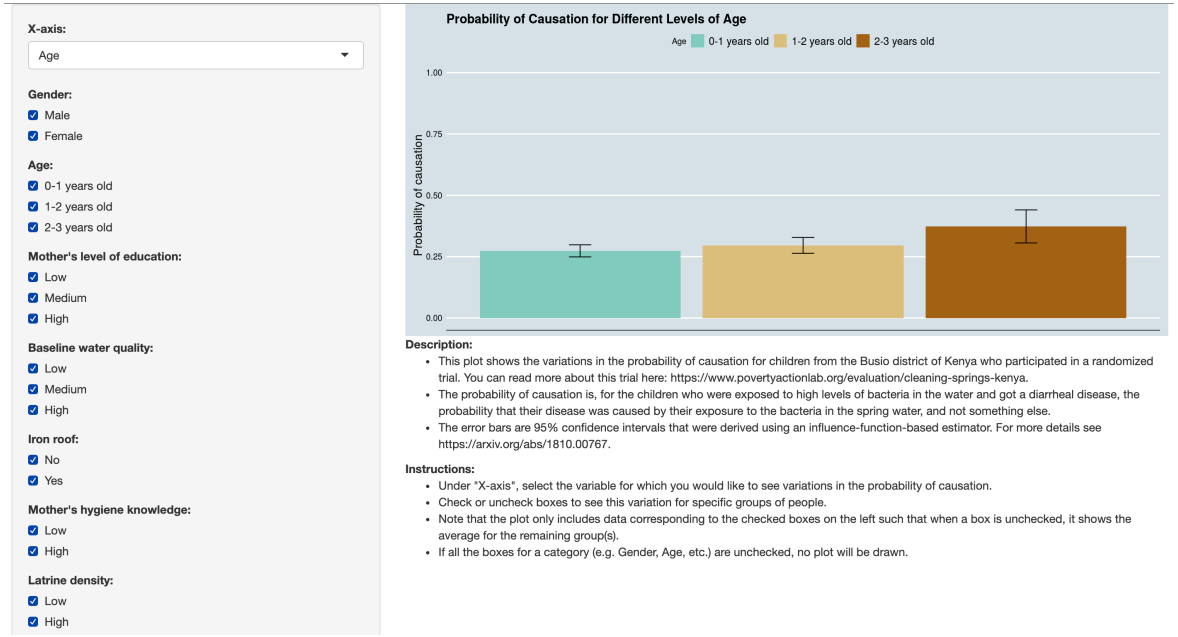
Variables:

- **Y: Outcome —** 1: if mother reported the child had diarrhea in past week. Diarrhea defined as three or more “looser than normal” stools within 24 hours at any time in the past week, 0: o.w.
- **X: Covariates —** Gender, age, mother’s years of education, (baseline ->) water quality, home has iron roof, mother’s hygiene knowledge, latrine density, number of children under 12 living at home.
- **A: Exposure —** 1: if spring is not protected, 0: o.w.

Identification assumptions:

- Binary exposure/outcome, consistency, no unobserved confounders, monotonicity.

Results:
An app to explore the heterogeneity of PC



https://mcuellar.shinyapps.io/apppcausation/

Note: 3 p's!

What happened with Dewayne Johnson?

https://www.theguardian.com/business/2018/aug/10/monsanto-trial-cancer-dewayne-johnson-ruling

Summary

1. Should use **probability of causation** for questions of attribution.

2. Should use a **nonparametric influence-function-based estimator** if you would like to avoid making parametric assumptions and obtain valid confidence intervals.

3. In a search for **new causal parameters**. The law has many causal claims that have been analyzed without using causal inference.

Questions?

R: `pcausation`

Background on Influence functions

- **What is an influence function?** Essentially the first derivative of a parameter, in a distributional/functional Taylor expansion. Exists for many but not all parameters because it requires some smoothness (pathwise differentiability).
- **Why do we care? In the 1980s,** they became well-known for being used in robust statistics in 1980s by Tukey and others: How to do estimation that isn't super sensitive to a specific point that has high influence?
- **Recently they have been used as a tool (recipe) to derive estimators that have useful properties** (e.g. in causal inference, just because of the types of parameters that are of interest in this field).
  - Main property of interest is they allow you to **generate estimators that converge to a normal distribution asymptotically** with mean zero and some variance (can reach efficiency bound - lowest possible theoretical error in some sense). Also called "root-n consistency".
  - **Asymptotic normality leads to being able to derive valid confidence intervals**, regardless of how you estimate the variance.
  - **But there is a catch**: The asymptotic variance will be a function of nuisance parameters (e.g., outcome regressions, propensity scores) that must be estimated. For the estimator to be asymptotically normal, **the errors in the nuisance estimators need to converge at a certain (relatively slow) rate**. The form of the error term depends on the specific parameter of interest.
  - This is where doubly robust estimators come from: Error term looks like $\| \hat{\pi} - \pi \| \| \hat{\mu} - \mu \|$, so you can use parametric models for estimating both nuisance functions and produce consistent estimators, even if one of the two models (π or μ) is misspecified.
  - In this research, we use nonparametric models for nuisance estimation, so we don't need double robustness to get asymptotic normality and CI's. Also, the nonparametric setting has a unique influence function.
- Influence-function-based estimators are **especially useful for high-dimensional data** (where parametric assumptions are more difficult to get right).

Background on influence functions

- An estimator $\hat{\psi}$ of the target parameter ψ has influence function φ if it is asymptotically linear such that

$$\hat{\psi} = \psi + \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right),$$

  where φ has zero mean and finite variance $\mathbb{E}(\varphi\varphi^T)$, and $Z_i = (X, A, Y)_i$.

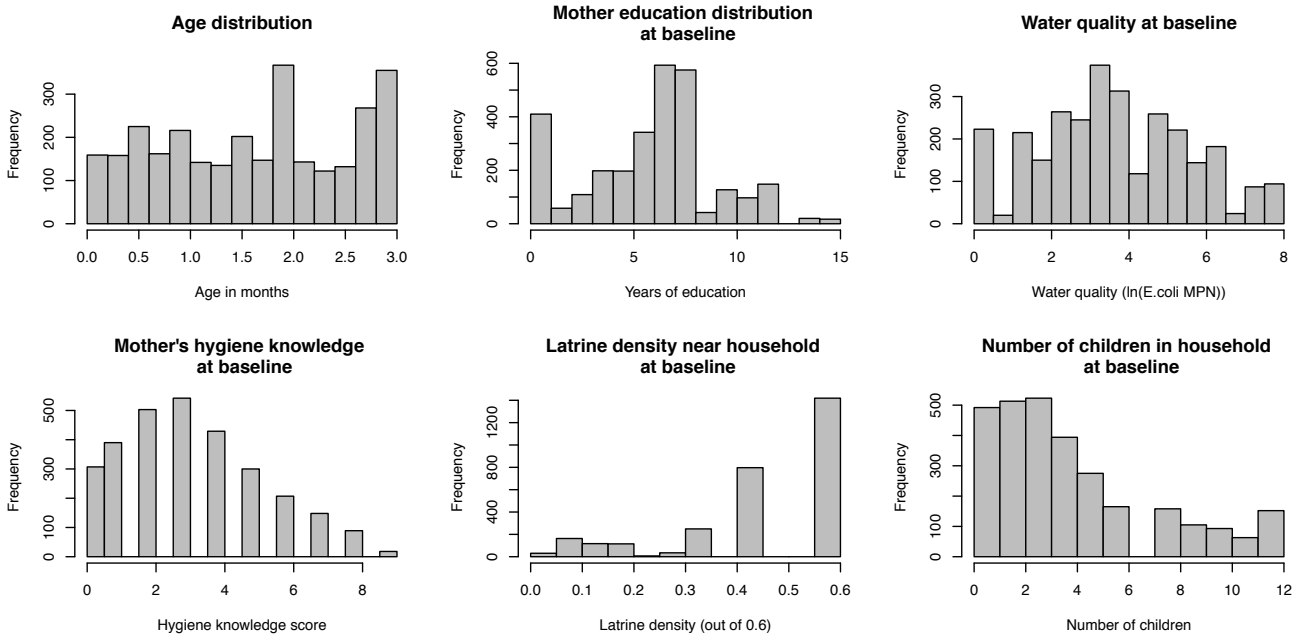
- By the CLT and Slutsky's theorem,

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, E(\varphi\varphi^T)), \text{ where } \varphi = \varphi(Z, \eta).$$

  van der Vaart 2000, Bickel et al. 1993, Kennedy 2016, others, and review by Cuellar, Mauro, and Kennedy 2017.

Back

Covariates histograms



Odds of causation

For a binary covariate X_1, the coefficient β_1 is the log-odds ratio between the group where $X_1 = 0$ and the group where $X_1 = 1$. As is usual, to translate to odds one can just exponentiate the log-odds. In our case, the odds refers to the "odds of causation." The odds of causation are

$$\text{Odds of causation} = \frac{P(Y^0 = 0 \mid Y = 1, A = 1, X = x)}{P(Y^0 = 1 \mid Y = 1, A = 1, X = x)}. \tag{22}$$

If the odds of causation are equal to three, for example, then we can say that it is three times more likely that the outcome Y was caused by exposure A than not. The odds ratio corresponding to covariate X_1, for example, is

$$\text{Odds ratio} = \frac{\text{Odds}(Y^0 = 0 \mid Y = 1, A = 1, X_1 = 1, X_2, \ldots, X_n)}{\text{Odds}(Y^0 = 0 \mid Y = 1, A = 1, X_1 = 0, X_2, \ldots, X_n)}. \tag{23}$$

Background on influence functions

- An estimator $\hat{\psi}$ of the target parameter ψ has influence function φ if it is asymptotically linear such that

$$\hat{\psi} = \psi + \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_i) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right),$$

  where φ has zero mean and finite variance $\mathbb{E}(\varphi\varphi^T)$, and $Z_i = (X, A, Y)_i$.

- By the CLT and Slutsky's theorem,

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, E(\varphi\varphi^T)), \text{ where } \varphi = \varphi(Z, \eta).$$

  van der Vaart 2000, Bickel et al. 1993, Kennedy 2016, others, and review by Cuellar, Mauro, and Kennedy 2017.

Back

Assumption: No unobserved confounders

$$Y^a \perp\!\!\!\perp A | X$$

- The outcome in the treatment group would have been the same as the outcome in the control group, had subjects in the treatment group received the control.
- The treated and the untreated are *exchangeable* (conditionally on X).
- No matter what treatment you actually get, if you'd been in A=0 then you get Y^0 and if you'd been in A=1 then you get Y^1.

A way to visualize path wise derivatives

- https://observablehq.com/@herbps10/one-step-estimators-and-pathwise-derivatives
- Herb Susmann PhD student in Biostatistics and Epidemiology at University of Massachusetts Amherst.