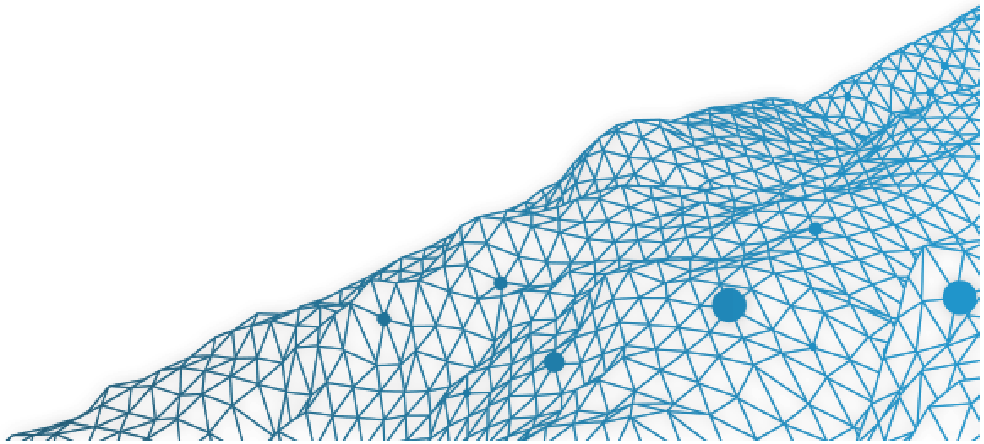
# # 11 Observations, Experiments & Causality

Arthur Charpentier (Université du Québec à Montréal)

Machine Learning & Econometrics

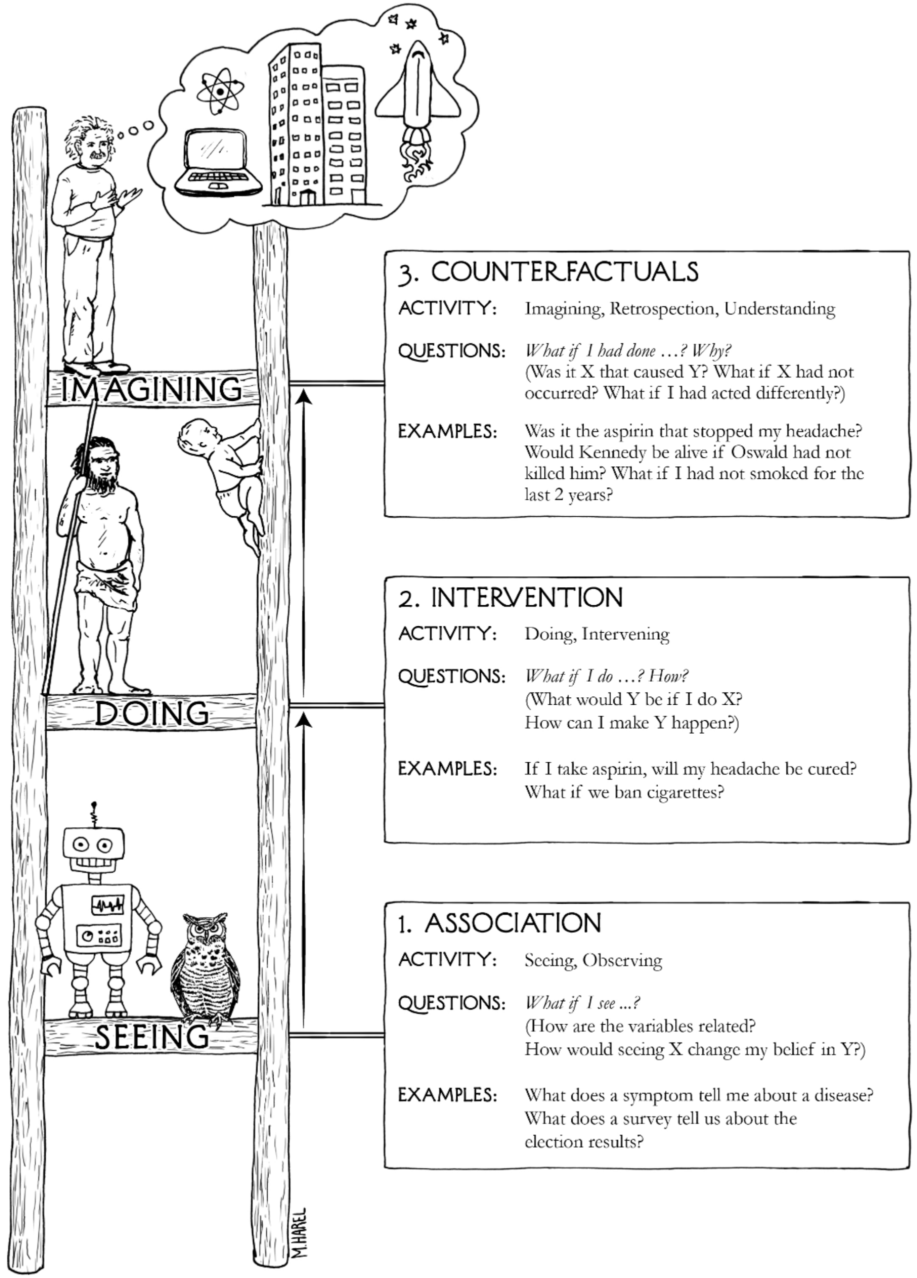
SIDE Summer School - July 2019



## Causal Inference

"*Social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult*", Grimmer (2015, We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together)

(source Pearl & Mackenzie (2018, The Book of Why))



## Causality and Counterfactuals

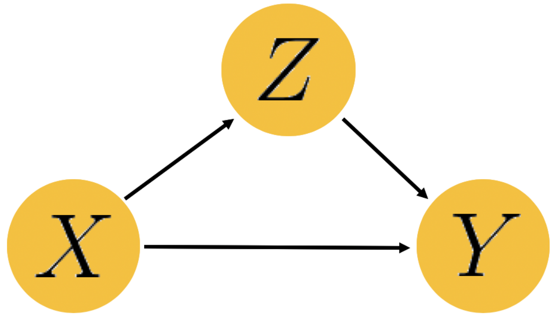
"*we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed*", Hume (1748, An Enquiry Concerning Human Understanding)

Classical conditional : if $X$ occurred, then $Y$ occurred

Counterfactual : if $X$ had not occurred, then $Y$ would not have occurred

"*no causation without manipulation*", Holland (1986, Statistics and Causal Inference)

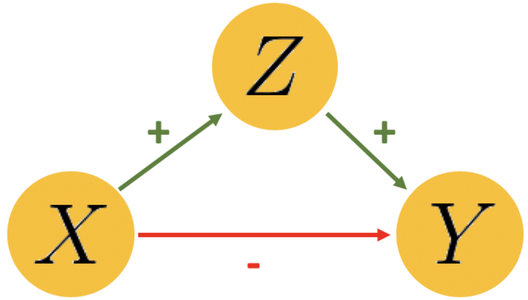
Importance of Directed Acyclic Graphs (DAG) to describe causal effects between variables



## Berkeley Gender Bias

Graduate admissions data from Berkeley, 1973

- men : 8442 applications, 44% admission rate
- women : 4321 applications, 35% admission rate




Discrimination towards women ?

| | | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|
| M | applied | 825 | 560 | 325 | 417 | 191 | 373 |
| | admitted | 62% | 63% | 37% | 33% | 28% | 6% |
| F | applied | 108 | 25 | 593 | 375 | 393 | 341 |
| | admitted | 82% | 68% | 34% | 35% | 24% | 7% |

see Bickel *et al.* (1975, Sex bias in graduate admissions)

## Simpson's Paradox & Ecological Fallacy

| | hosp. A | hosp. B |
|---|---|---|
| total | 1000 | 1000 |
| survivors | 800 | 900 |
| deads | 200 | 100 |
| rate (%) | 80% | 90% |

| | hosp. A | hosp. B |
|---|---|---|
| total | 600 | 900 |
| survivors | 590 | 870 |
| deads | 10 | 30 |
| rate (%) | 98% | 97% |

| | hosp. A | hosp. B |
|---|---|---|
| total | 400 | 100 |
| survivors | 210 | 30 |
| deads | 190 | 70 |
| rate (%) | 53% | 30% |

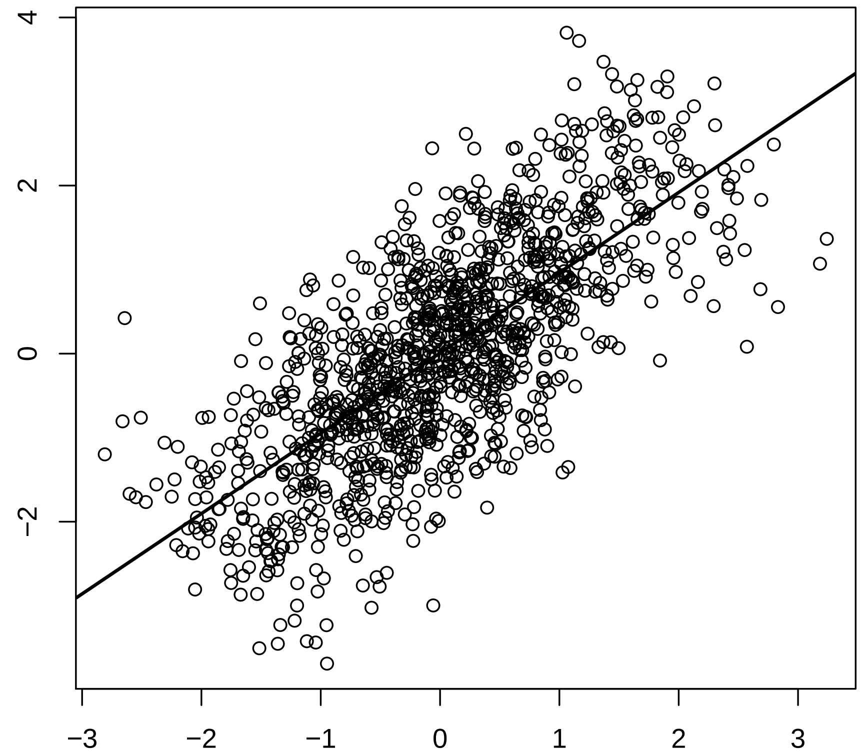
## Simpson's Paradox & Ecological Fallacy

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



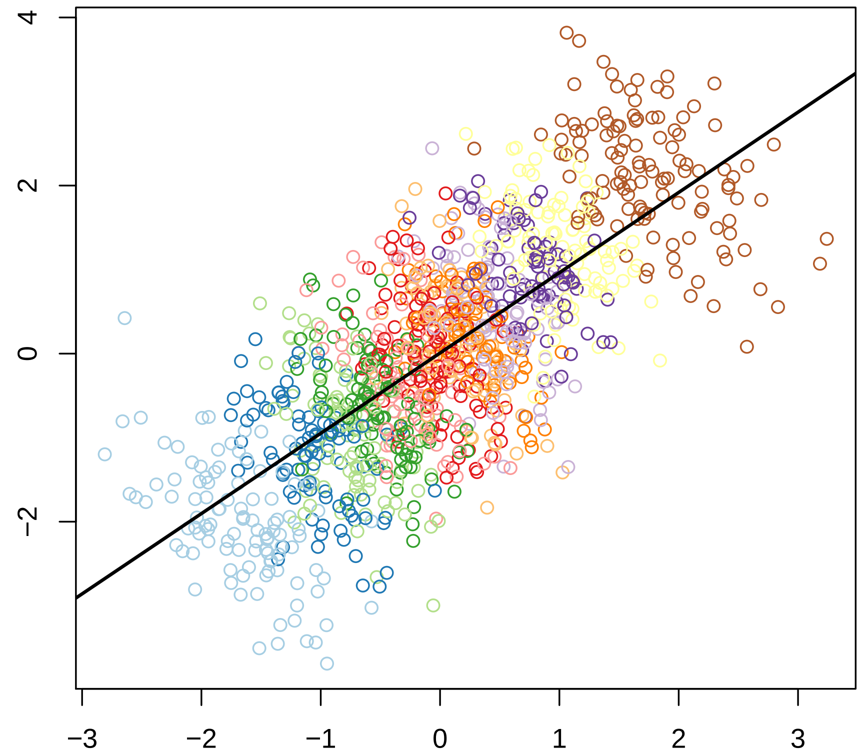
## Simpson's Paradox & Ecological Fallacy

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$



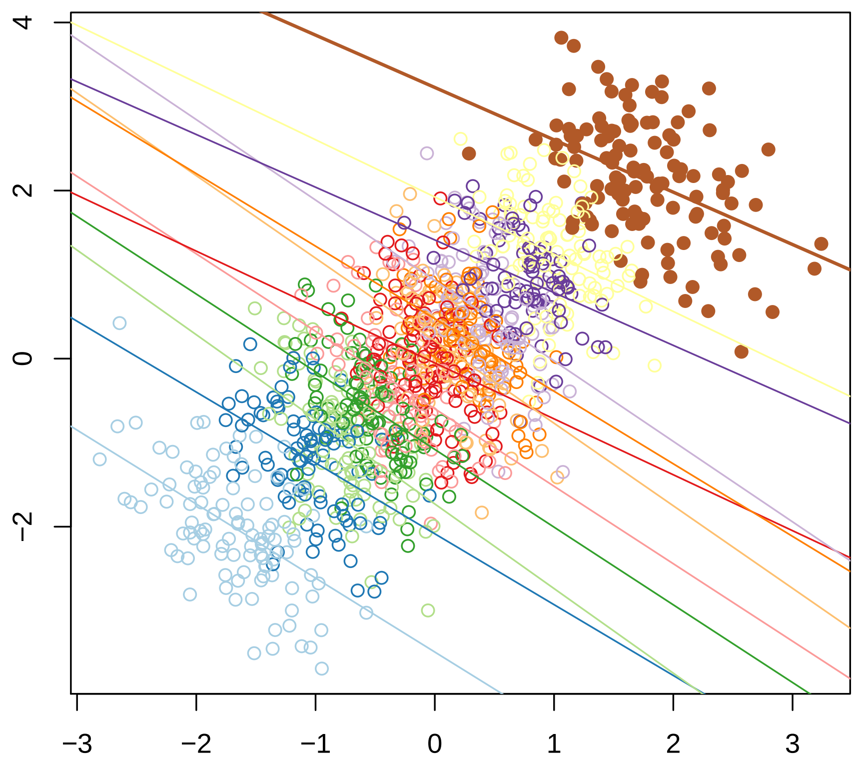
## Simpson's Paradox & Ecological Fallacy

Heuristically, it is possible to have

$$\frac{a}{A} \leq \frac{b}{B} \text{ and } \frac{c}{C} \leq \frac{d}{D}$$

and at the same time

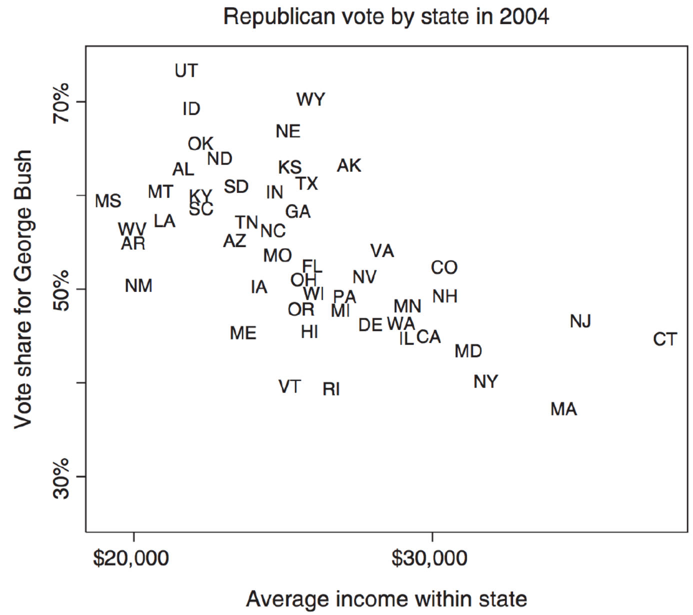
$$\frac{a+c}{A+C} \geq \frac{b+d}{B+D}$$

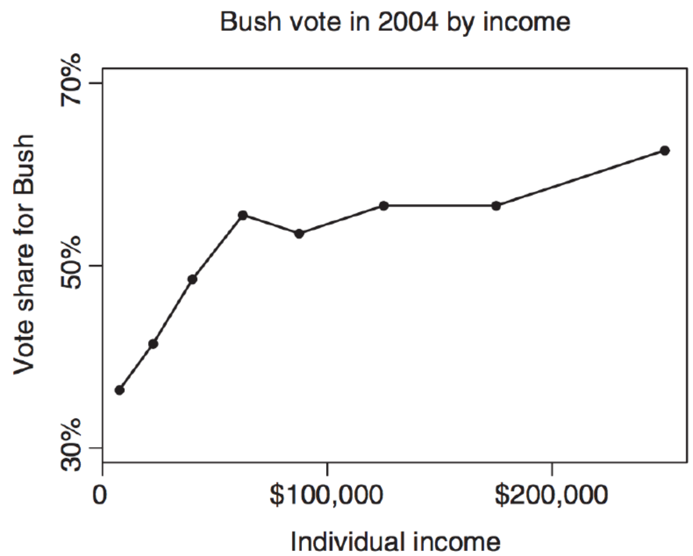


## Simpson's Paradox & Ecological Fallacy

Very important concept in political science

see Gelman (1986, Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do)



## Conditional Independence or *could a machine create a DAG* ?

### Conditional Independence

$X \perp\!\!\!\perp Y|Z$ if and only if

- $f(x,y|z) = f(x|z) \cdot f(y|z)$
- $f(x,y,z) \cdot f(z) = f(x,z) \cdot f(y,z)$
- $f(y|x,z) = f(y|z)$

i.e. once we know $Z$, learning the value of $x$ does not provide additional information about $X$

### Partial Correlation

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{(1-\rho_{xz}^2) \cdot (1-\rho_{yz}^2)}}$$

## Conditional Independence or *could a machine create a DAG* ?

$$\text{as } X \perp\!\!\!\perp Y \Longrightarrow \rho_{xy} = 0,\ \ X \perp\!\!\!\perp Y|Z \Longrightarrow \rho_{xy|z} = 0$$

and if $(X, Y, Z)$ is Gaussian, the converse is true

$$\text{as } \rho_{xy} = 0 \Longrightarrow X \perp\!\!\!\perp Y,\ \ \rho_{xy|z} = 0 \Longrightarrow X \perp\!\!\!\perp Y|Z$$

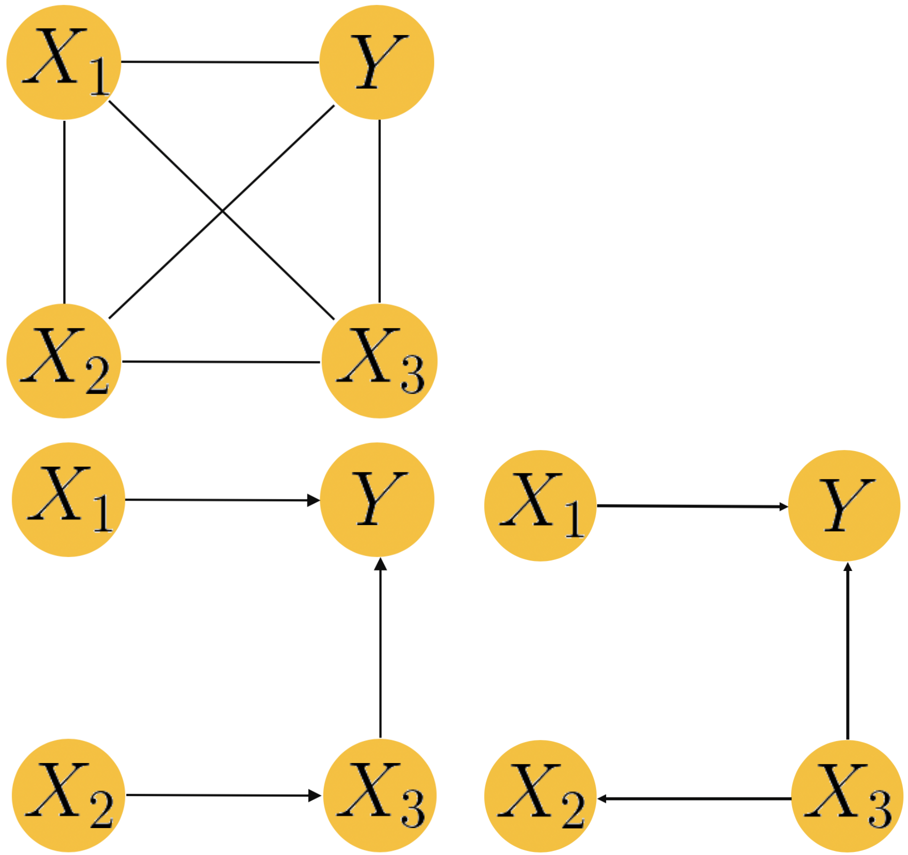
## Conditional Independence or *could a machine create a DAG* ?

Consider variables $X_1, X_2, X_3$ and $Y$

We have

- $X_1 \perp\!\!\!\perp X_2$
- $X_1 \perp\!\!\!\perp X_3$
- $X_2 \perp\!\!\!\perp Y | X_3$

Using conditional independence tests, remove edges, identify colliders and chains (might not be a unique solution...)



## Conditional Independence or *could a machine create a DAG* ?

Assuming a Gaussian setting, Spirtes *et al.* (2000, Causation, Prediction, and Search) suggested to test $\rho_{xy|\boldsymbol{z}} = 0$ using

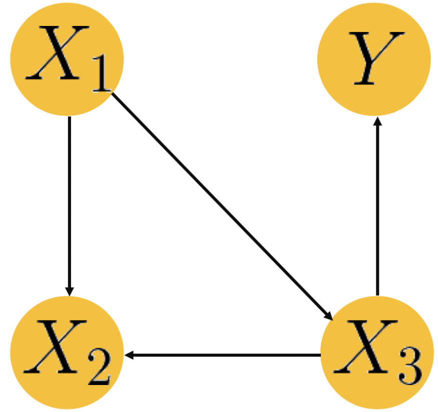
$$z_n = \frac{1}{2}\sqrt{n - \dim[\boldsymbol{z}] - 3}\log\frac{|1 + \rho_{xy|\boldsymbol{z}}|}{|1 - \rho_{xy|\boldsymbol{z}}|}$$

One can also consider a non-parametric test, based on the distance

$$d(x, y, \boldsymbol{z}) = \widehat{f}(x, y, \boldsymbol{z}) \cdot \widehat{f}(\boldsymbol{z}) - \widehat{f}(x, \boldsymbol{z}) \cdot \widehat{f}(y, \boldsymbol{z})$$

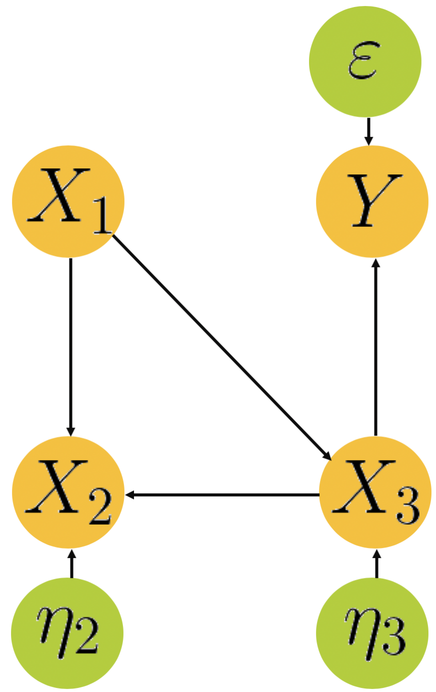
(main issue here: curse of dimensionality)

## From DAG to Structural Econometric Equations




There is a one-to-one mapping between a system of structural equations and a graphical model

- $X_2 = \beta_{23}X_3 + \beta_{21}X_1 + \eta_2$
- $X_3 = \beta_{31}X_1 + \eta_3$
- $Y = \beta_3 X_3 + \varepsilon$



## Observation and Experiment

Can we use observational data ?

Consider the following dataset (from Imai (2222) - resume.csv)

| | |
|---|---|
| `firstname` | first name of the fictitious job applicant |
| `sex` | sex of applicant, sex $\in$ {female,male} |
| `race` | race of applicant, race $\in$ {black,white} |
| `callback` | whether a call back was made, call $\in \{0,1\}$ |

See

```
  firstname    sex  race callback
1   Allison female white        0
2   Kristen female white        0
3   Lakisha female black        0
4   Latonya female black        0
5    Carrie female white        0
6       Jay   male white        0
```

## Observation and Experiment

It is experimental data, collected from an experimental research design, in which a treatment variable is manipulated in order to examine its causal effects on an outcome variable.

The treatment refers to the race of a fictitious applicant, implied by the name given on the résumé.

The outcome variable is whether the applicant receives a callback.

We are interested in examining whether or not the résumés with different names yield varying callback rates.

| | no call | call | total |
|---|---|---|---|
| black | 2278 | 157 | 2435 |
| white | 2200 | 235 | 2435 |
| total | 4478 | 392 | 4870 |

## Observation and Experiment

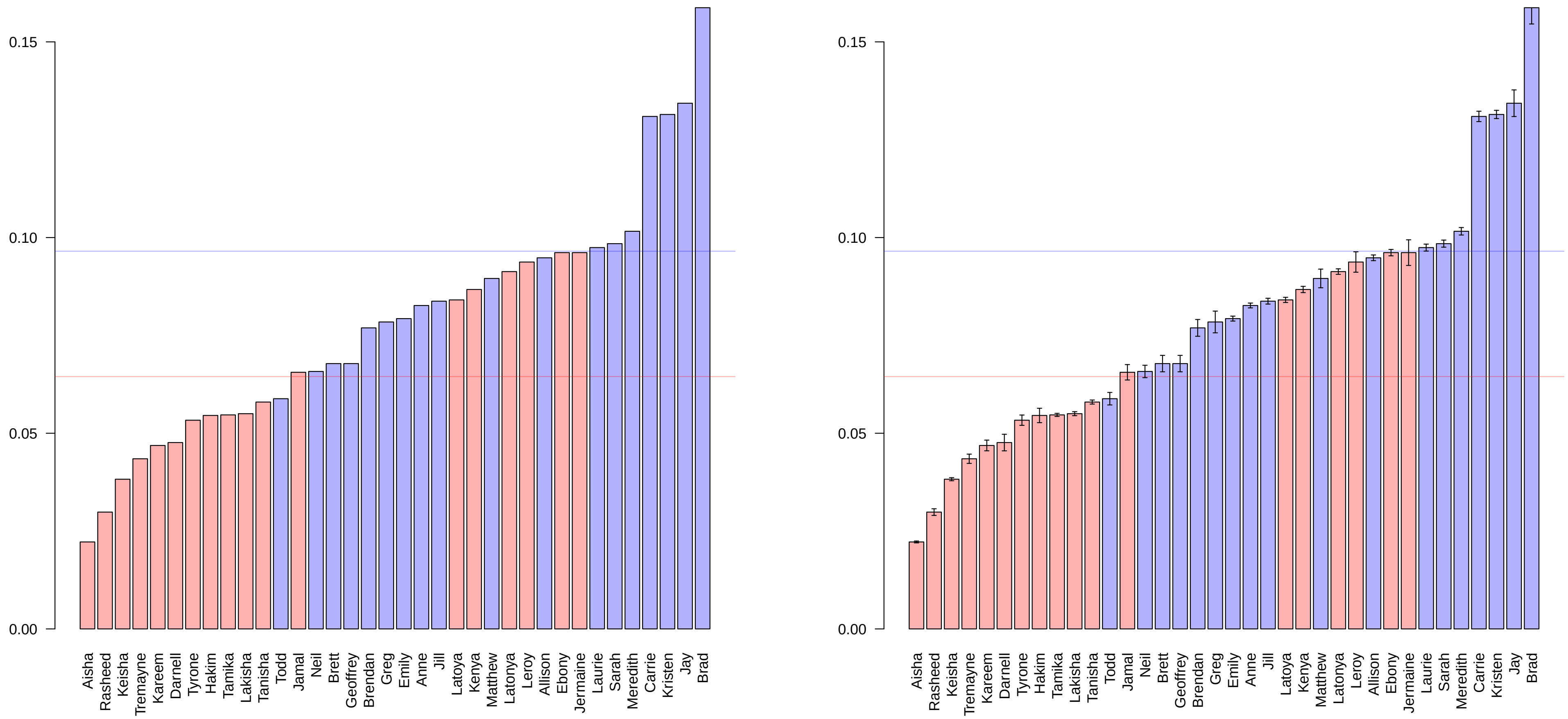
Conditional probabilities

| | no call | call | total |
|---|---|---|---|
| black | 50.87% | 40.05% | 50.00% |
| white | 49.13% | 59.95% | 50.00% |
| total | 100.00% | 100.00% | 100.00% |

| | no call | call | total |
|---|---|---|---|
| black | 93.55% | 6.45% | 100.00% |
| white | 90.35% | 9.65% | 100.00% |
| total | 91.95% | 8.05% | 100.00% |

3.2% points difference...

# Observation and Experiment



## Looking for a Conterfactual

```
  firstname    sex  race callback
1   Allison female white        0
```

Would the same employer would have called back if the applicant's name were instead a stereotypically African-American.

Unfortunately, we would never observe this counterfactual outcome, because the researchers who conducted this experiment did not send out the same résumé to the same employer using Lakisha as the first name.

Let $t$ denote the treatment (the first name0) which sound African-American $(t = 1)$ or not $(t = 0)$. Let $y(t)$ denote the response (call back) as a function of the treatment $t$. We do observe various covariate $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$ such as the age, the education, etc.

## Looking for a Conterfactual

The dataset is here

| | black-sounding | callback | | age | education |
|---|---|---|---|---|---|
| $i$ | name $t$ | $y(t=1)$ | $y(t=0)$ | $x_1$ | $x_2$ |
| 1 | 1 | 1 | ? | 20 | college |
| 2 | 0 | ? | 0 | 55 | high school |
| 3 | 0 | ? | 1 | 40 | graduate school |

We would like to quantify *ceteris paribus* $y(1) - y(0)$ (called causal effect)

Fundamental problem of causal inference : we cannot observe the counterfactual outcomes

## Looking for a Conterfactual

In observational studies, researchers do not conduct an intervention.

But they still want to quantify the impact of a policy. For instance the impact of an increase of minimum wage on unemployment (see Card & Krueger (1994), minwage.csv) )

In 1992, New Jersey (NJ) raised the minimum wage from $4.25 to 5.05$ (per hour).

```
chain
location
wage_before
wage_after
full_before
full_after
part_before
part_after
```

name of fast-food restaurant chain
location of restaurants (centralNJ, northNJ, PA, shoreNJ, southNJ)
wage before minimum-wage increase
wage after minimum-wage increase
number of full-time employees
(before and after)
number of part-time employees
(before and after)

## Looking for a Conterfactual

One can look at a counterfactual, with a neighboring state - Pennsylvania (PA) - that will be our control group

This would be a cross-section comparison design

| | NJ | | PA | |
|---|---|---|---|---|
| | mean | below $5.5 | mean | below $5.5 |
| before | $4.61 | 91.06% | $4.65 | 94.03% |
| after | $5.08 | 0.34% | $4.61 | 95.52% |

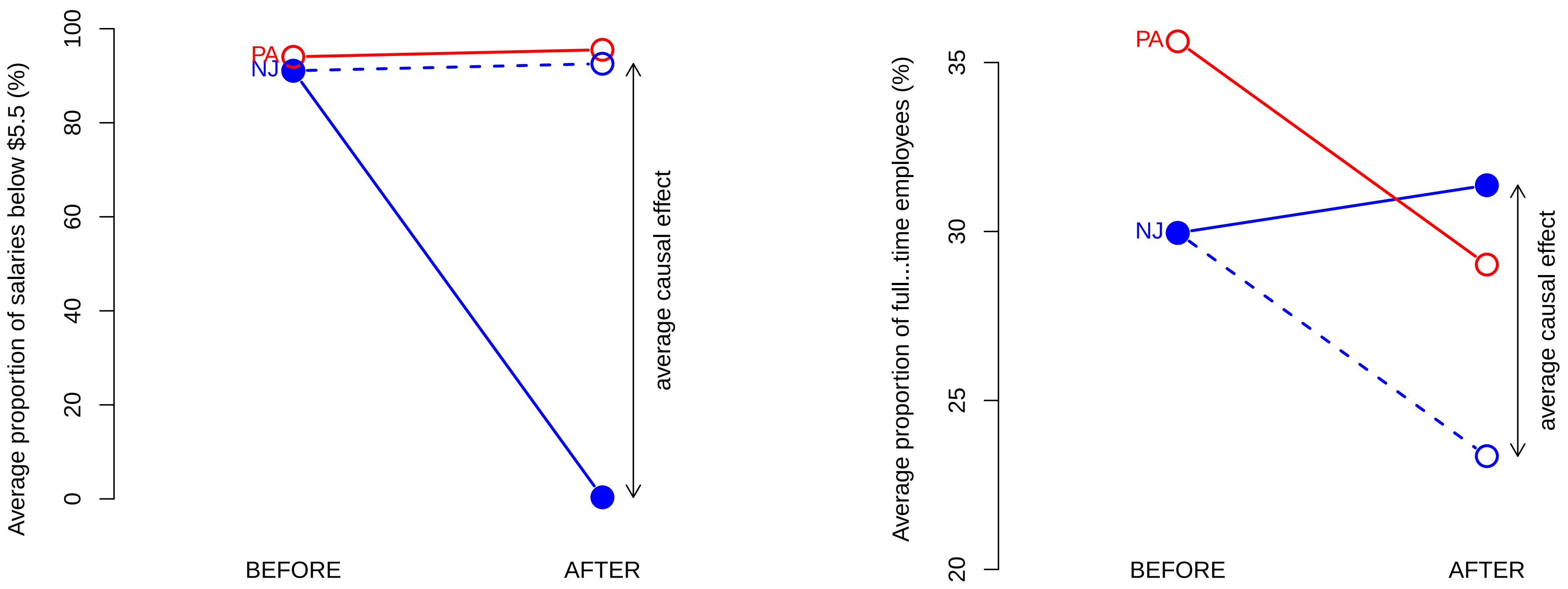
## Looking for a Conterfactual

| | NJ | | | PA | | |
|---|---|---|---|---|---|---|
| | partial | full | proportion | partial | full | proportion |
| before | 5423 | 2319 | 29.9% | 1291 | 714 | 35.6% |
| after | 5351 | 2446 | 31.4% | 1339 | 547 | 29.0% |

Difference-in-differences estimate,

$$DID = \underbrace{\overline{y}_{t=1}^{\text{after}} - \overline{y}_{t=1}^{\text{before}}}_{\text{difference in the treatment group}} - \underbrace{\overline{y}_{t=0}^{\text{after}} - \overline{y}_{t=0}^{\text{before}}}_{\text{difference in the control group}}$$

## Observation and Experiment




(the counterfactual outcome for the treatment group has a time trend parallel to that of the control group)

## Randomized Experiments

$n$ individuals, that are either treated ($t_i = 1$) or not ($t_i = 0$). We observe outcome $y_i$ for covariates $\boldsymbol{x}_i$.

We want to study potential outcomes $y_i(1)$ and $y_i(0)$

turnout

| | $y_i(1)$ | $y_i(0)$ | $t_i$ | $x_{1,i}$ | $x_{2,i}$ |
|---|---|---|---|---|---|
| 1 | $y_1$ | ? | 1 | $x_{1,1}$ | $x_{2,1}$ |
| 2 | ? | $y_2$ | 0 | $x_{1,2}$ | $x_{2,2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_n$ | ? | 1 | $x_{1,n}$ | $x_{2,n}$ |

The causal effect is $y_i(1) - y_i(0)$

Assume no simultaneity, no interference between individuals and same treatment

## Randomized Experiments

**ATE - Average Treatment Effect**

Average Treatment Effect $\mathbb{E}\big[Y(1) - Y(0)\big]$

Sample Average Treatment Effect $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\big[y_i(1) - y_i(0)\big]$

Assume randomization of the treatment assignment, $n_1$ treated, $n_0$ non treated

Assumption : $(Y(1), Y(0)) \perp\!\!\!\perp T$

Crude estimator (difference in means), $\displaystyle\widehat{\tau} = \frac{1}{n_1}\sum_{i:t_i=1} y_i - \frac{1}{n_0}\sum_{i:t_i=0} y_i$, or

$$\widehat{\tau} = \sum_{i=1}^{n} \frac{t_i y_i}{n_1} - \frac{(1-t_i)y_i}{n_0}$$

Then $\mathbb{E}\big[\widehat{\tau}\big] = \mathbb{E}\big[Y(1) - Y(0)\big]$

## Randomized Experiments

Intuition with a simple regression, $y_i(t_i) = \alpha + \beta t_i + \varepsilon_i$, where $\mathbb{E}\big[\varepsilon\big] = 0$.

Causal effect is measured by $y_i(1) - y_i(0) = \beta$

More generally, assume that $\varepsilon$ is $\varepsilon(t)$ - with $\mathbb{E}\big[\varepsilon(T)\big] = 0$, then $y_i(1) - y_i(0) = \beta + \varepsilon_i(1) - \varepsilon_i(0)$. Thus $\beta = \mathbb{E}[Y_i(1) - Y_i(0)]$ (and $\beta = \mathbb{E}[Y_i(1))$

and $\alpha = \mathbb{E}[Y_i(0)]$

One can consider a more general model, with $y_i(t_i) = \alpha + \beta t_i + \varepsilon(t_i)$ where $\mathbb{E}[\varepsilon(t)] = 0$.

Then $\beta = \mathbb{E}[Y_i(1) - Y_i(0)]$ and $\alpha = \mathbb{E}[Y_i(0)]$ (as previously)

## Randomized Experiments

In this model - introduced in Neyman (1923, Sur les applications de la théorie des probabilités aux expériences agricoles) - (called Rubin-Neyman) let $\sigma_t^2 = \text{Var}[\varepsilon(t)]$

Since $\widehat{\tau} = \dfrac{1}{n_1}\displaystyle\sum_{i:t_i=1} y_i - \dfrac{1}{n_0}\displaystyle\sum_{i:t_i=0} y_i = \dfrac{1}{n_1}\displaystyle\sum_{i=1}^n t_i \cdot y_i - \dfrac{1}{n_0}\displaystyle\sum_{i=1}^n (1-t_i)\cdot y_i$

The variance estimate of $\widehat{\tau}$ is usually biased

$$\mathbb{E}\left(\frac{\widehat{\sigma}^2}{\sum(t_i-\overline{t})^2}\right) - \left(\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}\right) = \frac{(n_1-n_0)(n-1)}{n_1 n_0(n-2)}(\sigma_1^2-\sigma_0^2)$$

so bias is null when either $n_1 = n_0$ or $\sigma_1^2 = \sigma_0^2$.

But generally biased (even asymptotically).

## (Cluster) Randomized Experiments

One can also consider some cluster randomized experiments : treatment is at cluster level, see Bland (2004, Cluster randomised trials in the medical literature) or Boruch *et al.* (2004, Estimating the effects of interventions that are deployed in many places: place-randomized trials)

Consider clusters $j = 1, \cdots, m$, outcome is $y_{i,j}$, with treatment $t_j \in \{0, 1\}$.

Assume random assignment, i.e. $(Y_{i,j}(1), Y_{i,j}(0)) \perp\!\!\!\perp T_j$.

And for convenience, assume that $n_j$'s are equal...

$$\widehat{\tau} = \frac{1}{m_1} \sum_{j:t_j=1} \overline{y}_j - \frac{1}{m_0} \sum_{j:t_j=0} \overline{y}_j \text{ where } \overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j}$$

## (Cluster) Randomized Experiments

As before, $\mathbb{E}\big[\widehat{\tau}\big]=\mathbb{E}\big[Y(1)-Y(0)\big]$

Exact variance is $\text{Var}[\widehat{\tau}]=\dfrac{\text{Var}[\overline{Y}(1)]}{m_1}+\dfrac{\text{Var}[\overline{Y}(0)]}{m_0}$,

where $\text{Var}[\overline{Y}(t)]=\dfrac{\sigma_t^2}{n}[1+(n-1)\rho_t]\leq\sigma_t^2$

(due to (possible) intracluster correlation $\rho_t$)

## Observational Studies & Propensity Score

In the regression discontinuity design, we want to find an (arbitrary) cutoff point $c$ that determines the treatment assignment, i.e. $t_i = \mathbf{1}_{x_i \geq c}$.

We want to estimate $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]$

Identification of the Average Treatment Effect :

Assumption 1: Overlap, $\forall x$, $\mathbb{P}[T = 1|X = x] \in (0, 1)$

Assumption 2: Ignorability $(Y(1), Y(0)) \perp\!\!\!\perp T|X = x$, $\forall x$

Set $\mu(t, x) = \mathbb{E}[Y_i(t)|T = t, X = x]$

Crude estimator (difference in means),

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(1, x) - \widehat{\mu}(0, x)$$

## Observational Studies & Regression Discontinuity

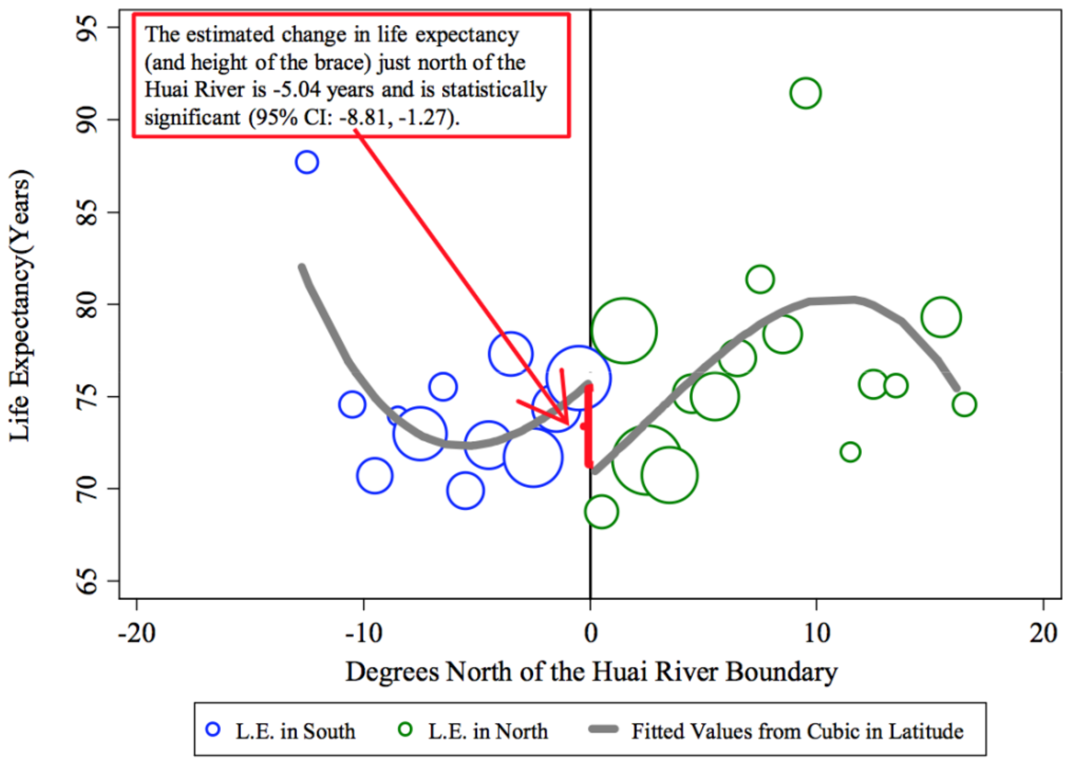



**Fig. 3.** The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

See Imbens & Lemieux (2008, Regression Discontinuity Designs).

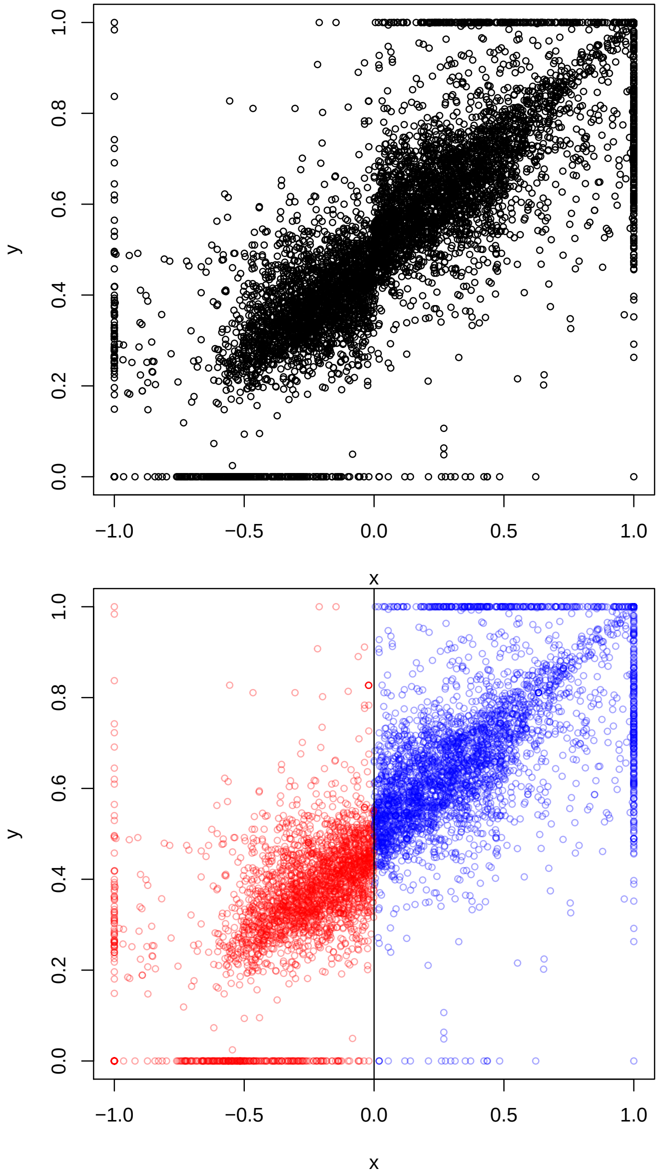
## Observational Studies & Regression Discontinuity

Consider the dataset from Lee (2008, Randomized experiments from non-random selection in U.S. House elections) - "*How would the Democratic party have performed in period 2, had they not held the seat (i.e. had the Democrats lost the election in period 1)*"

```
> library(RDDtools)
> data(Lee2008)
```

We want to test if there is a discontinuity (in 0, $x$ was rescaled here)

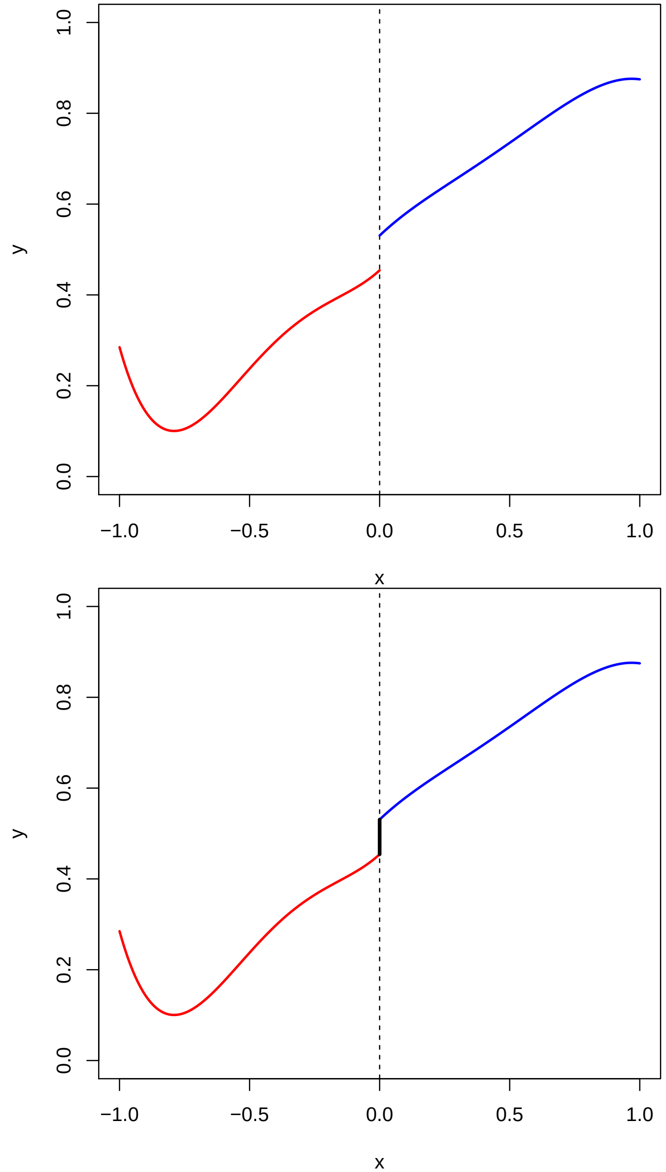
- with parametric tools
- with nonparametric tools



## Observational Studies & Regression Discontinuity

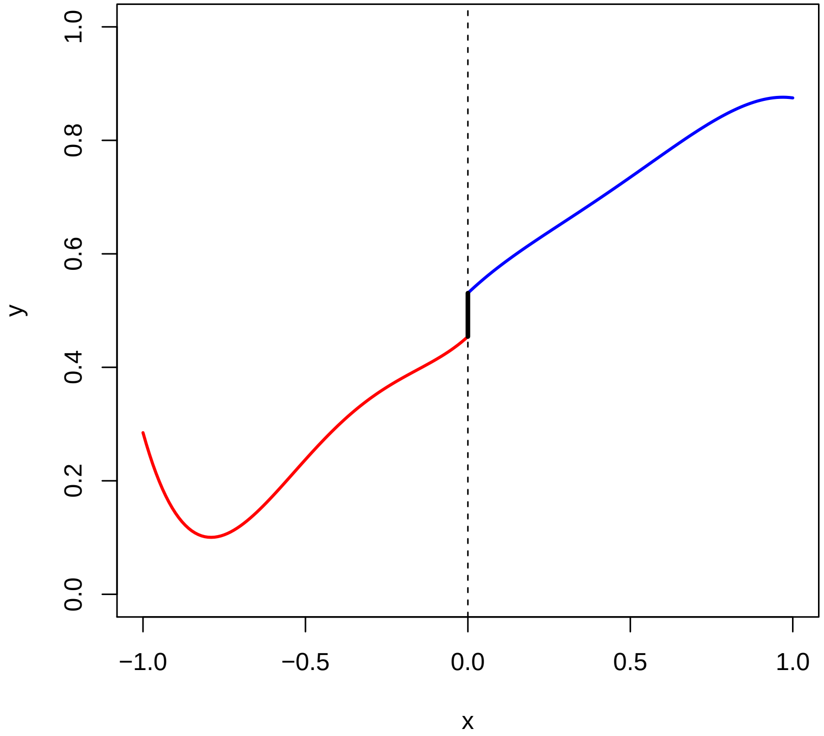
Use some 4th order polynomial, on each part

```
> idx1 = (Lee2008$x>0)
> reg1 = lm(y~poly(x,4),data=Lee2008[idx1,])
> idx2 = (Lee2008$x<0)
> reg2 = lm(y~poly(x,4),data=Lee2008[idx2,])
> s1=predict(reg1,newdata=data.frame(x=0))
> s2=predict(reg2,newdata=data.frame(x=0))
> abs(s1-s2)
          1
0.07659014
```



# Observational Studies & Regression Discontinuity

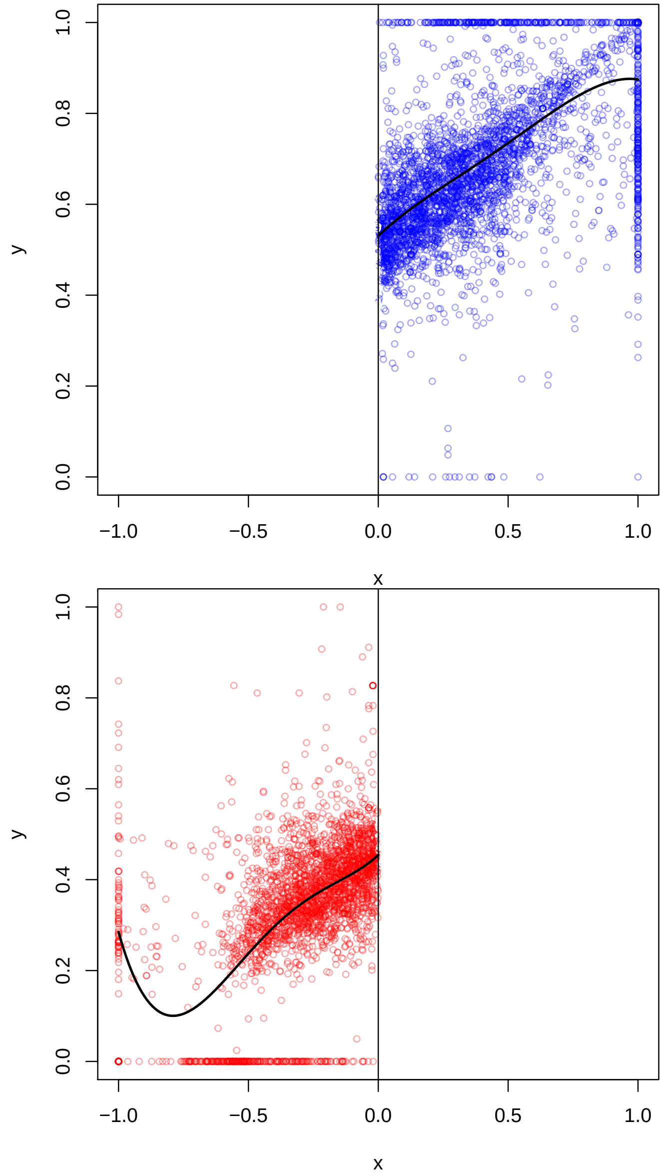
```
> reg_para <- RDDreg_lm(RDDdata(y = Lee2008$y, x = Lee2008$x, cutpoint = 0), order = 4)
> reg_para
### RDD regression: parametric ###
  Polynomial order:  4
  Slopes:  separate
  Number of obs: 6558 (left: 2740, right: 3818)

  Coefficient:
  Estimate Std. Error t value  Pr(>|t|)
D 0.076590   0.013239  5.7851 7.582e-09
    ***
```



## Observational Studies & Regression Discontinuity

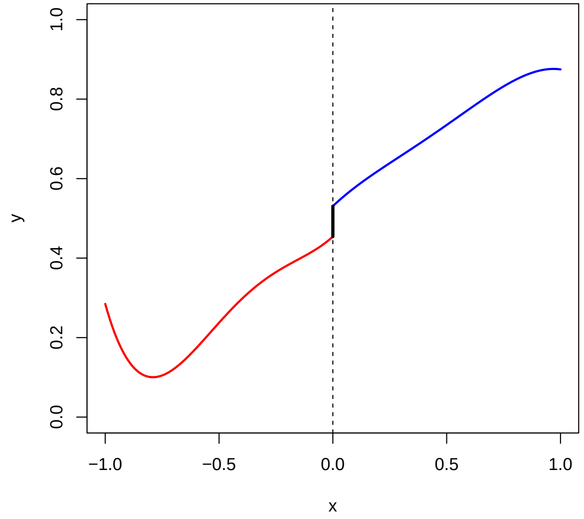
or use a simple local regression, see Imbens & Kalyanaraman (2012).

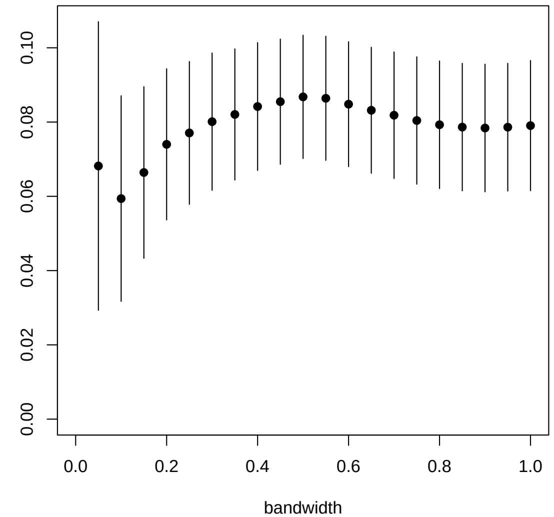
```
> reg1 = ksmooth(Lee2008$x[idx1],
    Lee2008$y[idx1], kernel = "normal",
     bandwidth = 0.1)
> reg2 = ksmooth(Lee2008$x[idx2],
    Lee2008$y[idx2], kernel = "normal",
     bandwidth = 0.1)
> s1 = reg1$y[1]
> s2 = reg2$y[length(reg2$y)]
> abs(s1-s2)
[1] 0.09883813
```



## Observational Studies & Regression Discontinuity

```
> reg_nonpara <- RDDreg_np(RDDobject = Lee2008_rdd, bw = .1)
> print(reg_nonpara)
### RDD regression: nonparametric local linear
  Bandwidth:  0.1
  Number of obs: 1209 (left: 577, right: 632)

  Coefficient:
  Estimate Std. Error z value  Pr(>|z|)
D 0.059397   0.014119   4.207 2.588e-05 ***
```



## Observational Studies & Propensity Score

**Propensity Score**

The Propensity Score is the probability to receive the treatment $\pi(x) = \mathbb{P}[T = 1|X = x]$

Assumption: Balancing property $T \perp\!\!\!\perp X|\pi(X)$

Assumption: Exogeneity given the propensity score $(Y(1), Y(0)) \perp\!\!\!\perp T|\pi(x), \forall x$

$$\widehat{\tau} = \sum_{i=1}^{n} \frac{t_i y_i}{n_1} - \frac{(1-t_i)y_i}{n_0}$$

consider

$$\widehat{\tau} = \sum_{i=1}^{n} \frac{t_i y_i}{n\widehat{\pi}(x_i)} - \frac{(1-t_i)y_i}{n(1-\widehat{\pi}(x_i))}$$

## Observational Studies & Propensity Score

Balancing condition: $\mathbb{E}\left[\dfrac{TY}{\pi(X)}-\dfrac{(1-T)Y}{1-\pi(X)}\right]=0$

See Lunceford & Davidian (2004, Stratification and Weighting Via the Propensity Score)

Consider also Robins's estimator,

$$\widehat{\tau}=\left(\sum_{i=1}^{n}\frac{\widehat{\mu}(1,x_i)}{n}+\frac{t_i(y_i-\widehat{\mu}(1,x_i))}{n\widehat{\pi}(x_i)}\right)-\left(\sum_{i=1}^{n}\frac{\widehat{\mu}(0,x_i)}{n}+\frac{(1-t_i)(y_i-\widehat{\mu}(0,x_i))}{n(1-\widehat{\pi}(x_i))}\right)$$

## Observational Studies & Matching

Consider a simple linear regression model, $y_i = \alpha + \beta t_i + \varepsilon_i$, then

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{y}_i(1) - \widehat{y}_i(0)\big)$$

where

$$\widehat{y}_i(1) = \begin{cases} y_i & \text{if } t_i = 1 \\ \displaystyle\frac{1}{n_1}\sum_{j=1}^{n} t_j \cdot y_j & \text{if } t_i = 0 \end{cases}$$

$$\widehat{y}_i(0) = \begin{cases} \displaystyle\frac{1}{n_0}\sum_{j=1}^{n}(1 - t_j) \cdot y_j & \text{if } t_i = 1 \\ y_i & \text{if } t_i = 0 \end{cases}$$

## Observational Studies & Matching

Consider a simple fixed effect regression model, $y_{i,t} = \alpha_i + \beta t_{i,t} + \varepsilon_{i,t}$

$$\widehat{\beta} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\big(\widehat{y}_{i,t}(1) - \widehat{y}_{i,t}(0)\big)$$

where, if $n_{1,i} = \sum_{\tau=1}^{T} t_{i,\tau}$ and $n_{0,i} = T - n_{1,i} = \sum_{\tau=1}^{T}(1 - t_{i,\tau})$

$$\widehat{y}_{i,t}(1) = \begin{cases} y_{i,t} & \text{if } t_{i,t} = 1 \\ \displaystyle\frac{1}{n_{1,i}}\sum_{\tau=1}^{T} t_{i,\tau}\cdot y_{i,\tau} & \text{if } t_{i,t} = 0 \end{cases}$$

$$\widehat{y}_{i,t}(0) = \begin{cases} \displaystyle\frac{1}{n_{0,i}}\sum_{\tau=1}^{T}(1 - t_{i,\tau})\cdot y_{i,\tau} & \text{if } t_{i,t} = 1 \\ y_{i,t} & \text{if } t_{i,t} = 0 \end{cases}$$

## Observational Studies & Matching

$\widehat{\beta}$ is actually the weighted least square estimate,

$$\widehat{\beta} = \underset{(\boldsymbol{\alpha},\beta)}{\text{argmin}}\left\{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\omega_{i,t}\big(y_{i,t}-\alpha_i-\beta t_{i,t}\big)\right\}$$

with $\omega_{i,t} = T/n_{t,i}$

More generally, we can have heterogeneity, i.e. $y_{i,t} = \alpha_i + \beta t_{i,t} + \boldsymbol{\gamma}^\top \boldsymbol{x}_{i,t} + \varepsilon_{i,t}$

Let $\pi(\boldsymbol{x})$ denote the propensity score $\mathbb{P}[T = 1|\boldsymbol{X} = \boldsymbol{x}]$.

## Observational Studies & Matching

We have the following (ex-post) interpretation

$$y_{i,t} - \widehat{\boldsymbol{\gamma}}^\top \boldsymbol{x}_{i,t} = \alpha_i + \beta t_{i,t} + \varepsilon_{i,t}$$

so use $\widehat{\beta}$ the weighted least square estimate,

$$\widehat{\beta} = \underset{(\boldsymbol{\alpha},\beta)}{\text{argmin}} \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \omega_{i,t} \big(y^\star_{i,t} - \alpha_i - \beta t_{i,t}\big) \right\}$$

with $\omega_{i,t} = T/n_{t,i}$ and

$$y^\star_{i,t} = \begin{cases} \dfrac{1}{\sum_\tau t_{i,\tau} \widehat{\pi}(\boldsymbol{x}_{i,\tau})^{-1}} \displaystyle\sum_{\tau=1}^{T} t_{i,\tau} \widehat{\pi}(\boldsymbol{x}_{i,\tau})^{-1} \cdot y_{i,\tau} & \text{if } t_{i,t} = 1 \\ \dfrac{1}{\sum_\tau (1 - t_{i,\tau})(1 - \widehat{\pi}(\boldsymbol{x}_{i,\tau}))^{-1}} \displaystyle\sum_{\tau=1}^{T} (1 - t_{i,\tau})(1 - \widehat{\pi}(\boldsymbol{x}_{i,\tau}))^{-1} \cdot y_{i,\tau} & \text{if } t_{i,t} = 0 \end{cases}$$

## Observational Studies & Matching

This is just the general case of the simple 2 time-period difference in differences

$$y_{i,\tau}(t) = \alpha_i + \beta t + \gamma\tau + \varepsilon_{i,\tau}$$

where $\tau$ is the time, $\tau \in \{0,1\}$. Hence

$y_{i,0}(0) = \alpha_i + \varepsilon_{i,0}$

$y_{i,1}(0) = \alpha_i\gamma + \varepsilon_{i,1}$

$y_{i,1}(1) = \alpha_i\beta + \gamma + \varepsilon_{i,1}$

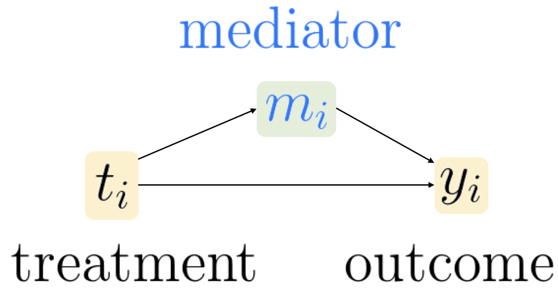
Assumption: $\mathbb{E}[y_{i,1}(0) - y_{i,0}(0)|T_{i,1} = t] = \gamma$

i.e.: $\mathbb{E}[\varepsilon_{i,1} - \varepsilon_{i,0}|T_{i,1} = t] = 0$

Imai *et al.* (2011, Unpacking the Black Box of Causality).

## Experimental / Observational Studies & Mediation

Assume that there is a possible mediator $m$ (possible indirect effects), see Baron & Kenny (1986, The Moderator-Mediator Variable Distinction)




Brader, Valentino & Suhay (2008, What Triggers Public Opposition to Immigration?)

Treatment $t \in \{\text{causasian}, \text{latino}\}$ - randomized experiment

Outcome $y$ is preference over immigration policy - measured

Mediation $m$ is anxiety - measured

Bertrand & Mullainathan (2008, Are Emily and Greg More Employable than Lakisha and Jamal?)

Treatment $t \in \{\text{white name}, \text{black name}\}$ - randomized experiment

Outcome $y\{\text{callback}, \text{no callback}\}$ - measured

## Experimental / Observational Studies & Mediation

Mediation $m$ is perceived qualifications of applicants

(1) Regress $y$ on $t$: $y_i = \alpha_1 + \beta_1 t_i + \boldsymbol{\xi}_1^\top \boldsymbol{x}_i + \varepsilon_{1,i}$

(2) Regress $m$ on $t$: $m_i = \alpha_2 + \beta_2 t_i + \boldsymbol{\xi}_2^\top \boldsymbol{x}_i + \varepsilon_{2,i}$

(3) Regress $y$ on $m$ and $t$: $y_i = \alpha_3 + \beta_3 t_i + \gamma_3 m_i + \boldsymbol{\xi}_3^\top \boldsymbol{x}_i + \varepsilon_{3,i}$

for some control variables $\boldsymbol{x}$

$\beta_1$ is ATE, and can be decomposed in $\beta_1 = \underbrace{\beta_3}_{\text{direct}} + \underbrace{\gamma\beta_2}_{\text{indirect}}$

## Experimental / Observational Studies & Mediation

Here $m$ might depend on $t$, $m_i(t_i)$ and $y$ depends on $m$ and $t$, $y_i(t_i, m_i(t_i))$.

- Indirect (mediation) causal effect is $\delta_i(t) = y_i(t, m_i(1)) - y_i(t, m_i(0))$ (identification pb)

- Direct causal effect is $\zeta_i(t) = y_i(1, m_i(t)) - y_i(0, m_i(t))$

Total causal effect is $\tau_i = y_i(1, m_i(1)) - y_i(0, m_i(0))$ and

$$\tau = \frac{[\delta_i(0) + \zeta_i(0)] + [\delta_i(1) + \zeta_i(1)]}{2}$$

$\delta_i(t)$ measures counterfactuals about treatment-induced mediator values

- Controlled direct effects: $\xi_i(t, m, w) = y_i(t, m) - y_i(t, w)$

- Interaction effects: $\xi_i(1, m, w) - \xi_i(0, m, w)$

See `mediation` library

## Experimental / Observational Studies & Mediation

In Brader, Valentino & Suhay (2008, What Triggers Public Opposition to Immigration?)

Randomized treatment: $t$ is randomized given $\boldsymbol{x}$, $(Y,M) \perp\!\!\!\perp T|X$

Sequential ignorability: $m$ is randomized given $\boldsymbol{x}$ and $t$, $Y \perp\!\!\!\perp M|T,X$

Then

$$\overline{\delta}(t) = \int \mathbb{E}[Y_i|m,t,\boldsymbol{x}_i]\big(d\mathbb{P}[m|1,\boldsymbol{x}_i] - d\mathbb{P}[m|0,\boldsymbol{x}_i]\big)d\mathbb{P}(\boldsymbol{x}_i)$$

$$\overline{\zeta}(t) = \int \big(\mathbb{E}[Y_i|m,1,\boldsymbol{x}_i] - \mathbb{E}[Y_i|m,0,\boldsymbol{x}_i]\big)d\mathbb{P}[m|t,\boldsymbol{x}_i]d\mathbb{P}(\boldsymbol{x}_i)$$

Use linear structural equations,

$$\begin{cases} m_i = \alpha_2 + \beta_2 t_i + \boldsymbol{\xi}_2^\top \boldsymbol{x}_i + \varepsilon_{2,i} \\ y_i = \alpha_3 + \beta_3 t_i + \gamma_3 m_i + \boldsymbol{\xi}_3^\top \boldsymbol{x}_i + \varepsilon_{3,i} \end{cases}$$

Separate least square to estimate $\widehat{\beta}_2$ and $\widehat{\gamma}$ and then Sobel test for significance

## Experimental / Observational Studies & Mediation

In Bertrand & Mullainathan (2008, Are Emily and Greg More Employable than Lakisha and Jamal?), crossover desgin

(1) standard randomized experiment: send Jamal's CV, record $y$

(2) change treatment to the opposite status, keep mediator fixed: send CV (same qualification) as Greg, record $y$

similar to Hainmueller & Hiscox (2010, Attitudes toward Highly Skilled and Low-skilled Immigration)

Baron-Kenny Procedure

(1) regress $y$ on $t$ (significant relationship)

(2) regress $m$ on $t$ (significant relationship)

(3) regress $y$ on $m$ and $t$ (significant relationship between $y$ and $m$)

## Experimental / Observational Studies & Mediation

E.g. binary mediator $m \in \{0,1\}$,

$$\text{logit}[\mathbb{E}[M_i|t_i,\boldsymbol{x}_i]] = \alpha_2 + \beta_2 t_i + \boldsymbol{\xi}_2^\top \boldsymbol{x}_i$$

$$\text{logit}[\mathbb{E}[Y_i|m_i,t_i,\boldsymbol{x}_i]] = \alpha_3 + \beta_3 t_i + \gamma_3 m_i + \boldsymbol{\xi}_3^\top \boldsymbol{x}_i +$$

Cannot use $\beta_2\gamma$, or $\beta_1 - \beta_3$ if

$$\text{logit}[\mathbb{E}[Y_i|t_i,\boldsymbol{x}_i]] = \alpha_1 + \beta_1 t_i + \boldsymbol{\xi}_1^\top \boldsymbol{x}_i$$

Assume that $\rho_{2,3} = \text{Corr}[\varepsilon_2,\varepsilon_3]$. Sequential ignorability means $\rho_{2,3} = 0$.

$$\overline{\delta}(0) = \overline{\delta}(1) = \frac{\beta_2\sigma_1}{\sigma_2}\left(\rho_{1,2} - \rho_{2,3}\sqrt{\frac{1-\rho_{1,2}^2}{1-\rho_{2,3}^2}}\right)$$

Thus, $\overline{\delta}(t) = 0$ means $\rho_{2,3} = \rho_{1,2}$.

## Experimental / Observational Studies & Mediation

```
> fit_m <- lm(mediator ~ treat + x)
> fit_y <- lm(y ~ treat + mediator +x)
```

For the mediation analysis use

```
> med <- mediation::mediate(fit_m, fit_y, treat="treat", mediator="
    mediator")
```

and for the sensitivity analysis

```
> mediation::medsens(med)
> plot(medsens(med), "rho")
```

Problem, sometimes $m$ is difficult to manipulate.

Use instrumental variables

## Causal Trees

Prediction with Machine Learning techniques ?

Get the smallest mean-squared error in a test set...

Given a candidate $\widehat{\mu}(\boldsymbol{x})$, use $\displaystyle\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu}(\boldsymbol{x}_i))^2$

With a tree, $\widehat{\mu}(\boldsymbol{x})$ is the sample mean of $y_i$'s within leaf $\ell(\boldsymbol{x})$

The in-sample goodness of fit measure is the mse, $\displaystyle\text{mse}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{\mu}(\boldsymbol{x}_i))^2$

Try to minimize $\text{mse}+\lambda\cdot\text{number of leaves}$

Select $\lambda$ with lowest out-of-sample goodness of fit measure, or use cross-validation.

```
> tree <- rpart(y ~ x, method="anova")
> i <- which.min(tree$cptable[,"xerror"])
> tree_2 <- prune(tree, cp=tree$cptable[i,"CP])
```

## Causal Trees

Consider our causal model, an let $\tau_i$ denote the treatment effect $\tau_i = y_i(1) - y_i(0)$

Consider possible heterogeneity, set $\mu(t, \boldsymbol{x}) = \mathbb{E}[Y(t)|T = t, \boldsymbol{X} = \boldsymbol{x}]$ and $\tau(x) = \mu(1, \boldsymbol{x}) - \mu(0, \boldsymbol{x})$

To estimate $\tau(x)$ a partition tree can be more interesting than a linear predictor.

Approach 1 : analyze the two groups separately

- estimate $\widehat{\mu}(1, \boldsymbol{x})$ on the sub-dataset where $t_i = 1$
- estimate $\widehat{\mu}(0, \boldsymbol{x})$ on the sub-dataset where $t_i = 0$
- use propensity score weighting
- (use within group cross-validation to tune parameters)
- prediction is $\widehat{\tau}(x) = \widehat{\mu}(1, \boldsymbol{x}) - \widehat{\mu}(0, \boldsymbol{x})$

Approach 2 : estimate $\mu(t, x)$ on both covariates

## Instruments

Recall that in a regression model, $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \varepsilon_i$, with endogeneous variables $\mathbb{E}(\boldsymbol{x}_i^\top\varepsilon_i) \neq \boldsymbol{0}$ and least squares estimators are not convergent ($\text{plim}\widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$ as $n \to \infty$)

Classical motivations

• variable omission : the true model is $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \boldsymbol{z}_i^\top\boldsymbol{\gamma} + \varepsilon_i$, then

$$\text{plim}\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta} + \mathbb{E}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\mathbb{E}(\boldsymbol{X}^\top\boldsymbol{Z})\boldsymbol{\gamma} \neq \boldsymbol{\beta}$$

See Mincer equation, where $y$ is the wage, $x$ the number of years of study, but there might be some ability bias

• measurement error : the true model is $y_i = \boldsymbol{x}_i^{\star\top}\boldsymbol{\beta} + \varepsilon_i$ where variables are measured with error, $\boldsymbol{x}_i = \boldsymbol{x}^\star + \boldsymbol{\eta}_i$ where $\mathbb{E}[\eta_i|\boldsymbol{x}_i^\star] = 0$ and $\eta_i \perp \varepsilon_i$. Thus, $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \nu_i$ where $\nu_i = \varepsilon_i - \boldsymbol{\beta}^\top\boldsymbol{\eta}_i$. Here, $\mathbb{E}[\nu_i\boldsymbol{x}_i] = -\boldsymbol{\beta}\text{Var}[\boldsymbol{\eta}] \neq \boldsymbol{0}$, and

$$\text{plim}\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta} + \mathbb{E}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\mathbb{E}(\boldsymbol{X}^\top\boldsymbol{\eta})\boldsymbol{\gamma} \neq \boldsymbol{\beta}$$

• simultaneity bias

## Causal Trees

Assume here that $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + b\tilde{x}_i + \varepsilon_i$, where

- $\mathbb{E}(\boldsymbol{x}_i^\top \varepsilon_i) = \boldsymbol{0}$
- $\mathbb{E}(\tilde{x}_i^\top \varepsilon_i) \neq 0$

We will use instruments $\boldsymbol{z}$ that

- should be exogeneous, $\mathbb{E}(\boldsymbol{z}_i^\top \varepsilon_i) = \boldsymbol{0}$
- should add information to $\boldsymbol{x}$, in the sense that in the regression

$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \varepsilon_i'$, $\boldsymbol{\gamma} \neq \boldsymbol{0}$

Set

$$\hat{x} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \tilde{x}$$

and observe that

$$\hat{\boldsymbol{x}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{x} = \boldsymbol{x}$$

## Causal Trees

Thus, define the filtrated model

$$y_i = \hat{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + b\hat{x}_i + \hat{\varepsilon}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + b\hat{x}_i + \hat{\varepsilon}_i$$

where $\hat{\varepsilon} = \varepsilon + (\tilde{x} - \hat{x})$.

From the rank condition, $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{X}$ exists

The filtrated model is exogeneous, in the sens that $\text{plim} \dfrac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{x}}_i \hat{\varepsilon}_i - -$

The IV estimate is here $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$

More generally,

$$y_i = \underbrace{\boldsymbol{x}_i^\top \boldsymbol{\beta}}_{\text{exogeneous}} + \underbrace{\tilde{\boldsymbol{x}}_i^\top \gamma}_{\text{endogeneous}} + \varepsilon_i$$

## Causal Trees

Consider instruments $\boldsymbol{z}_i$ and set $\boldsymbol{z}_i^\star = (\boldsymbol{z}_i, \boldsymbol{x}_i)$, such that they

• should be exogeneous, $\mathbb{E}[\boldsymbol{z}_i^{\star\top}\varepsilon_i] = \mathbf{0}$

• should add information to $\boldsymbol{x}$, in the sense that
$\text{rank}(\mathbb{E}[\boldsymbol{z}_i^{\star\top}\boldsymbol{x}_i]) = \dim(\boldsymbol{x}) + \dim(\tilde{\boldsymbol{x}})$ (rank condition)

Note that if rank $< \dim(\boldsymbol{x}) + \dim(\tilde{\boldsymbol{x}})$ the model is under-identified

Note that if rank $> \dim(\boldsymbol{x}) + \dim(\tilde{\boldsymbol{x}})$ the model is over-identified

indirect least square

The rank condition means that $\mathbb{E}[\boldsymbol{Z}^{\star\top}\boldsymbol{X}]$ can be inverted, and

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{Z}^{\star\top}\boldsymbol{X}]^{-1}\mathbb{E}[\boldsymbol{Z}^{\star\top}Y]$$

and the empirical version is

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{Z}^{\star\top}\boldsymbol{X}]^{-1}\boldsymbol{Z}^{\star\top}\boldsymbol{y}$$

## Causal Trees

More generally, we can have some over-identified model

We need to find a matrix $A$ such that $A\mathbb{E}[\boldsymbol{Z}^{\star\top}\boldsymbol{X}]$ can be inverted, and then

$$\boldsymbol{\beta} = \mathbb{E}[A\boldsymbol{Z}^{\star\top}\boldsymbol{X}]^{-1}\mathbb{E}[A\boldsymbol{Z}^{\star\top}Y]$$

and the empirical version is

$$\widehat{\boldsymbol{\beta}}_A = [A\boldsymbol{Z}^{\star\top}\boldsymbol{X}]^{-1}A\boldsymbol{Z}^{\star\top}\boldsymbol{y}$$

Then, for any $A$ such that this estimator exists $\widehat{\boldsymbol{\beta}}_A$ is convergent and asymptotically Gaussian.

## Causal Trees

There is an optimal matrix $A^\star$ such that the asymptotic variable of $\widehat{\boldsymbol{\beta}}_{A^\star}$ is minimal, and it is

$$A^\star = \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{Z}^\star]^{-1}\mathbb{E}[\boldsymbol{Z}^{\star\top}\boldsymbol{Z}^\star]$$

Set $A_n = \boldsymbol{X}^\top \boldsymbol{Z}^\star(\boldsymbol{Z}^{\star\top}\boldsymbol{Z}^\star)^{-1}$, then $A_n \to A^\star$, and then, the empirical estimator is the so-called double-least-square estimate

$$\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\top \boldsymbol{Z}^\star(\boldsymbol{Z}^{\star\top}\boldsymbol{Z}^\star)^{-1}\boldsymbol{Z}^{\star\top}\boldsymbol{X}]^{-1}\boldsymbol{X}^\top \boldsymbol{Z}^\star(\boldsymbol{Z}^{\star\top}\boldsymbol{Z}^\star)^{-1}\boldsymbol{Z}^{\star\top}\boldsymbol{y}$$

or

$$\widehat{\boldsymbol{\beta}} = [\Pi_{\boldsymbol{Z}^\star}\boldsymbol{X}^\top\Pi_{\boldsymbol{Z}^\star}\boldsymbol{X}]^{-1}\Pi_{\boldsymbol{Z}^\star}\boldsymbol{X}^\top\boldsymbol{y} = (\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}\boldsymbol{y}$$

which is called double-least-square estimate since

- we consider the regression of $\boldsymbol{x}$ on $\boldsymbol{z}^\star$
- we consider the regression of $y$ on $\tilde{\boldsymbol{x}}$

Sometimes, instruments are poor proxys for endogeneous variables, and are called weak instruments.

## Appendix: Regression Trees

One can use the within variance (within a leaf)

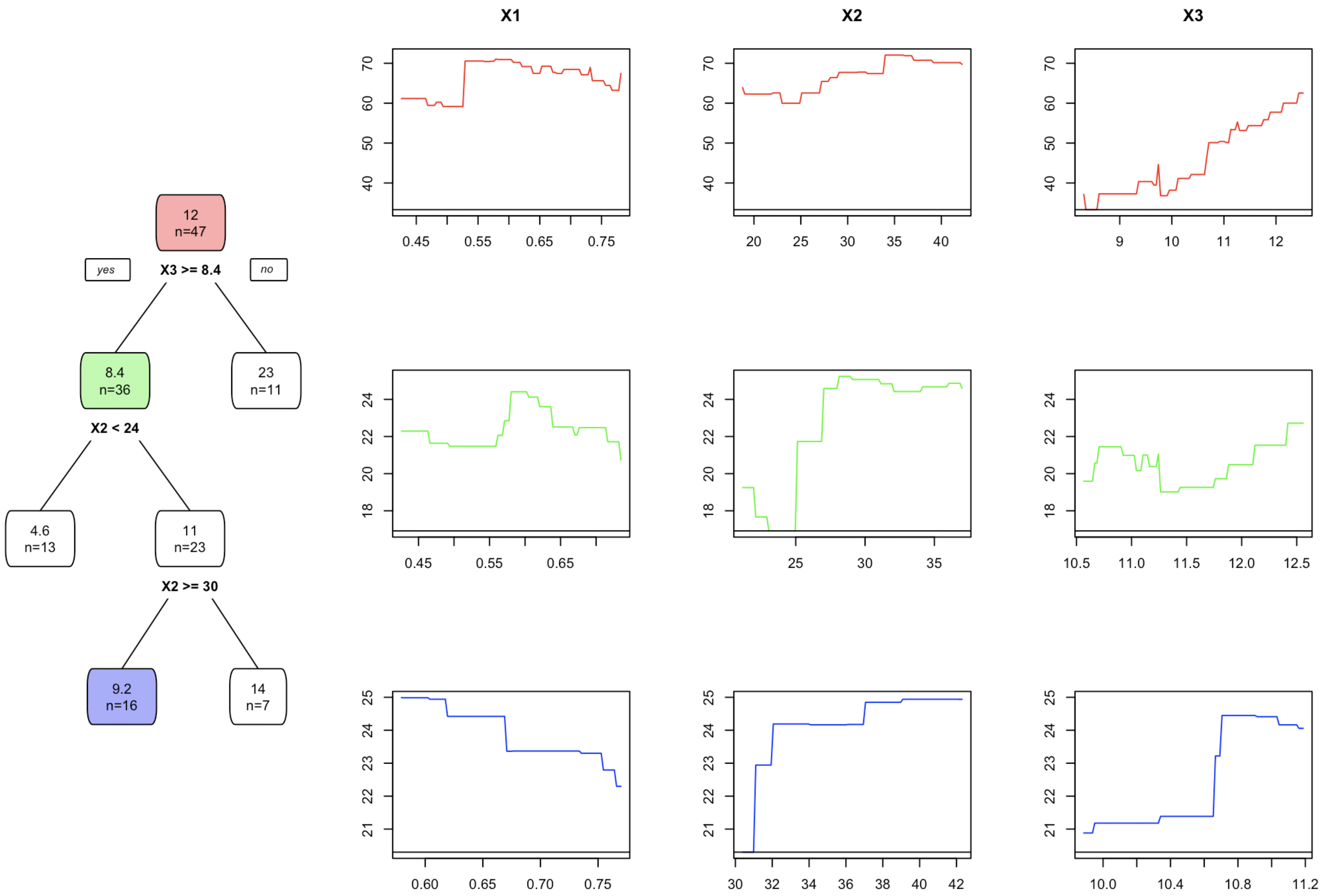
- within a leaf the total sum of squares is $\sum_i (y_i - \overline{y})^2$
- if we split in two parts, $\sum_{i:L} (y_i - \overline{y}_L)^2 + \sum_{i:R} (y_i - \overline{y}_R)^2$

The spluit is chosen to maximize $\sum_i (y_i - \overline{y})^2 - \sum_{i:L} (y_i - \overline{y}_L)^2 - \sum_{i:R} (y_i - \overline{y}_R)^2$

(one can use Student $t$-test for pruning)

See the application on the Chicago dataset, with 3 explanatory variables

# Appendix: Regression Trees



## Appendix: Regression Trees

One can use the entropy (that can be related to Kullback-Leibler distance)

**Entropy**

For a random variable $X$ the entropy is $H(X) = -\mathbb{E}_X\left[\log f(X)\right]$

Natural extensions are the joint entropy $H(X,Y) = -\mathbb{E}_{X,Y}\left[\log f(X,Y)\right]$

and the conditional entropy $H(Y|X) = -\mathbb{E}_{X,Y}\left[\log f(Y|X)\right]$

Since $H(X) \neq H(Y)$, $H(Y|X) \neq H(X|Y)$.

**Mutual Information**

Mutual information of $(X,Y)$ is $I(X,Y) = \mathbb{E}_{X,Y}\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right]$

$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$

## Appendix: Regression Trees

$I(X,Y)\geq 0$, and $I(X,Y)=0$ if and only if $X \perp\!\!\!\perp Y$

thus $H(X)\leq H(X|Y)$ and $H(X)=H(X|Y)$ if and only if $X \perp\!\!\!\perp Y$

**Kullback-Leibler**

For two distributions $f,g$ $KL(f\|g)=\mathbb{E}_f\left[\log\frac{f(X)}{g(X)}\right]=\int\log\frac{f(x)}{g(x)}f(x)dx$

Hence, $I(X,Y)=KL(f\|f^{\perp})$

Thus, Kullback-Leibler divergence can be called relative entropy.

## Appendix: Regression Trees

For a Gaussian vector $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, the joint entropy is

$$h(\boldsymbol{Z})=\frac{1}{2}\log\left[(2\pi)^d|\boldsymbol{\Sigma}|\right]$$

If $\boldsymbol{Z}^\star\in\text{argmax}\{H(\boldsymbol{Z})\}$ s.t. $\text{Var}[\boldsymbol{Z}]=\boldsymbol{\Sigma}$, then $\boldsymbol{Z}^\star\sim\mathcal{N}(\mathbb{E}[\boldsymbol{Z}^\star],\boldsymbol{\Sigma})$

**Cross entropy**

For distributions $f,g$, $CE(f|g)=-\mathbb{E}_f\left[\log g(X)\right]=-\int\log[g(x)]f(x)dx$

$CE(f|g)=H(f)+KL(f\|g)$