Date: Thursday, May 9th, 2019

WHO Example

y_i = age of death of individual

Covariates: - Location (US, UK, Canada)

- Type of smoking (Cigars/pipes, Cigarettes)
- Demographic info

Aim: Determine whether smoking leads to an earlier death.

Correlation vs Causation

- Without control over allocation of experimental units we can no longer can directly obtain causal effects.
- Could try running a regression of y_i on covariates but results only represent correlation, not causation.

Matching general idea: Compare treated vs not treated (smoking vs not) on individuals that are similar according to their covariate information.

WHO Example:

Data:

Smoker (?) (treated ?) ← factor of interest	Age of death ← y_i	Local ← Characteristics →
1	65	US
0	68	UK
0	52	Canada
0	⋮	⋮
1		
0		

- Matching partitions groups of individuals based on shared characteristics (not the metric of interest nor response)

y_i^1 = response (age of death) of individual i if they smoked

y_i^0 = response (age of death) of individual i if they did not smoke

We only see one of these! The other hidden response is a counterfactual!

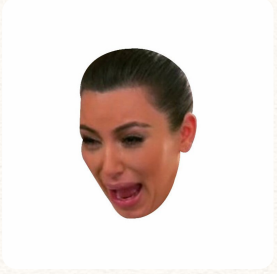
δ_i: $y_i^1 - y_i^0$ (effect on individual i of smoking)

Overall point: We can never (ever in experimentation) calculate individual-level effects. Counterfactuals prevent us from doing so.

* we can however calculate average treatment effects over a population.

$$\delta = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$$

* To calculate δ w/experimentation, we randomly allocate individuals to treatment and control while controlling for possible confounding variables.



↳ grouping units according to similar (same) factor level!

* In observational data treated units have no common factor levels guaranteed.

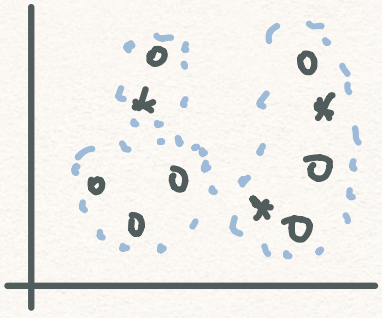
↳ Treatment & control groups may have widely variable characteristics.

* matching can (approximately) take care of confounding variables and mimic the homogeneity of groups from experimentation!

Exact Matching:

- Identify groups whose covariates X are exactly the same.
- Ensure that at least one treated & one non-treated are in each group

Neighborhood matching



* Treated
o Not-treated

Distance are based on covariate similarity

K = 2
for each * identify two closest O's.

propensity score matching:

- Most commonly used form of matching.
- Defined as propensity from logistic regression of D_i on X_i.
- Ex: D_i = smoker or not

  X_i = covariates other than D & y

Two steps:

1. Calculate $\pi_i = P(D_i = 1 | X_i)$ from logistic regression of D on X.
2. Cluster into five groups using k means or use a neighborhood matching approach.