Causal Inference in Social Science
An Elementary Introduction

Hal Varian
Google,Inc.

February 8, 2015

Philosophy of this talk

- Everything should be made as simple as possible, but no simpler. (attributed to Albert Einstein)

Philosophy of this talk

- Everything should be made as simple as possible, but no simpler. (attributed to Albert Einstein)
- But that's too hard, so we'll just make things as simple as possible. (attributed to Robert Wilensky)

A motivating problem from marketing

- y_c = sales in city c
- x_c = advertising in city c
- e_c = error in city c (cumulative effect of omitted predictors)
- For simplicity, center the data to eliminate the constant
- model: $y_c = bx_t + e_t$

Simple approach

- Run a regression of y_c on x_c
- When do we get a good estimate of b?
  - $b^{LS} = \text{cov}(x, y)/\text{cov}(x, x)$
  - $b^{LS} = \text{cov}(x, bx + e)/\text{cov}(x, x) = b + \text{cov}(x, e)/\text{cov}(x, x)$
  - Need $\text{cov}(x, e) = 0$
- But that is very unlikely when x is chosen by someone
- They will generally base their choice of x on other factors that affect y they observe but we don't

Honolulu and Fargo

- Product: movie about surfing will open in two cities
  - Fargo: ad spend per capita 10 cents, revenue \$1
  - Honolulu: ad spend per capita \$1, revenue \$10
- $y_c = 10x_c$ fits the data perfectly

Honolulu and Fargo

- Product: movie about surfing will open in two cities
  - Fargo: ad spend per capita 10 cents, revenue \$1
  - Honolulu: ad spend per capita \$1, revenue \$10
- $y_c = 10x_c$ fits the data perfectly
- But do you really think that increasing Fargo ad spend by a factor of 10 will increase revenue there by a factor of 10?

What's wrong?

- Revenue also depends on other factors, e.g., "interest in surfing"
  - "Interest in surfing" will affect revenue directly
  - ... and indirectly, via ad spend choices
- "Interest in surfing" is an example of a confounding variable.
- Confounding variables are omitted variables that are also correlated with the predictors, x.
- Very common in models involving human choice since decision makers observe important factors that the analyst cannot observe

Examples

How does fertilizer affect crop yields? Farmers choose fertilizer application.

How does education affect income? Wealthy parents or high ability students tend to acquire both more education and more income.

How does health care affect income? Those who have good jobs tend to have health care.

Contemplated policy choices

Want to estimate effects to evaluate some proposed policies, e.g.,

- What would happen if we apply more fertilizer?
- What would happen if we offered more scholarships?
- What would happen if we offered cheaper health insurance?

Ideally would run an experiment, but this is expensive. What can we learn from observational data (where there is no explicit experimentation)?

- Critical question: is the treatment *imposed* on the population, or *chosen* by members of population?
  - Impact of treatment on population?
  - Impact of treatment on those who choose to be treated?

Fundamental identity of causal inference

Outcome for treated − outcome for untreated

= [Outcome for treated − Outcome for treated if not treated]
+ [Outcome for treated if not treated − Outcome for untreated]
= Impact of treatment on treated
+ selection bias

If treatment is randomly assigned

Fundamental identity of causal inference

Outcome for treated − outcome for untreated

= [Outcome for treated − Outcome for treated if not treated]
+ [Outcome for treated if not treated − Outcome for untreated]
= Impact of treatment on treated
+ selection bias

If treatment is randomly assigned

- Selection bias is zero.

Fundamental identity of causal inference

Outcome for treated − outcome for untreated

= [Outcome for treated − Outcome for treated if not treated]
+ [Outcome for treated if not treated − Outcome for untreated]
= Impact of treatment on treated
+ selection bias

If treatment is randomly assigned

- Selection bias is zero.
- Treated are random selection from population, so impact on treated = impact on population

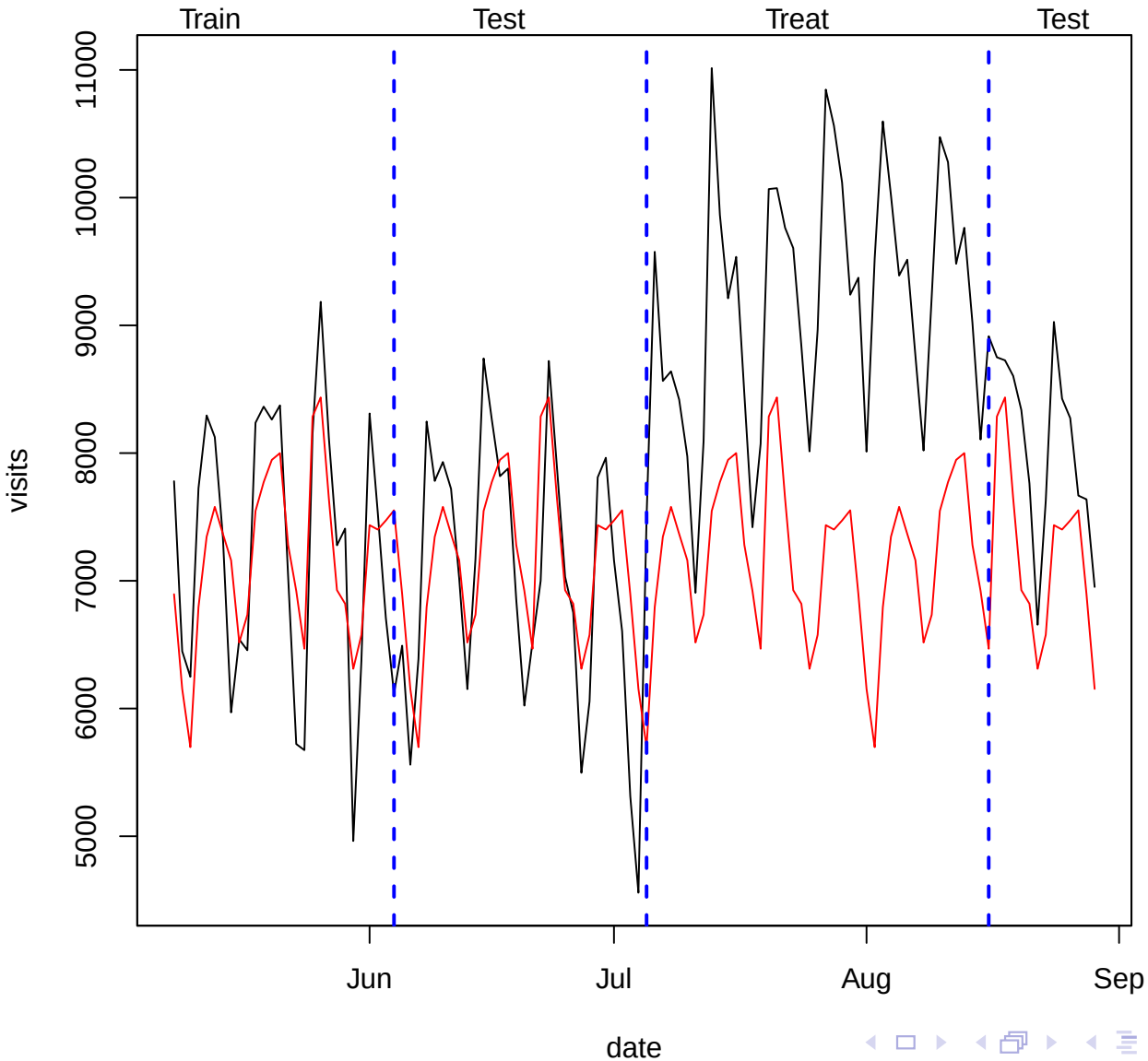
Ways to estimate causal effects from observational data?

1. Randomized Experiments
2. Natural Experiments
3. Instrumental Variables
4. Regression Discontinuity
5. Difference in Differences

Randomized experiments

- Want to treat subject at some (randomly chosen) times rather than others
- What to treat some (randomly chosen) subjects and not others
- Use pre-experiment behavior to build a predictive model for outcomes. Use this model to estimate counterfactual.
- Train, test, treat, compare
  - Train: a model on training data
  - Test: a model on holdout data (placebo experiment)
  - Treat: apply treatment to treatment group
  - Compare: treated to predicted counterfactual (model and/or control group)

Train-test-treat-compare paradigm



Natural experiments

- What happens if we don't have an actual experiment?
- Perhaps we can find a natural experiment
  - Something that assigns treatment in essentially random way

Super Bowl and ad impact

- Well known that the home cities of teams in Super Bowl see elevated viewership of about 10%
- Super Bowl ads are sold out by October
- From viewpoint of advertiser two "randomly" chosen cities see 10% more ad impressions
- Compare per capita sales in treated cities (home team + host) to sales in untreated cities
- Plausible to think that choice of these cities is independent of advertiser decisions

Instrumental variables

- $y_c = bx_c + e_c$
- Can we find an instrumental variable?
  - Something that moves x_c but is independent of e_c?
- Example from Super Bowl: z = home city of teams playing
- Ad impressions depend on home city: $x_c = az_c + d_t$
- $b^{IV} = \text{cov}(z, y)/\text{cov}(z, x)$

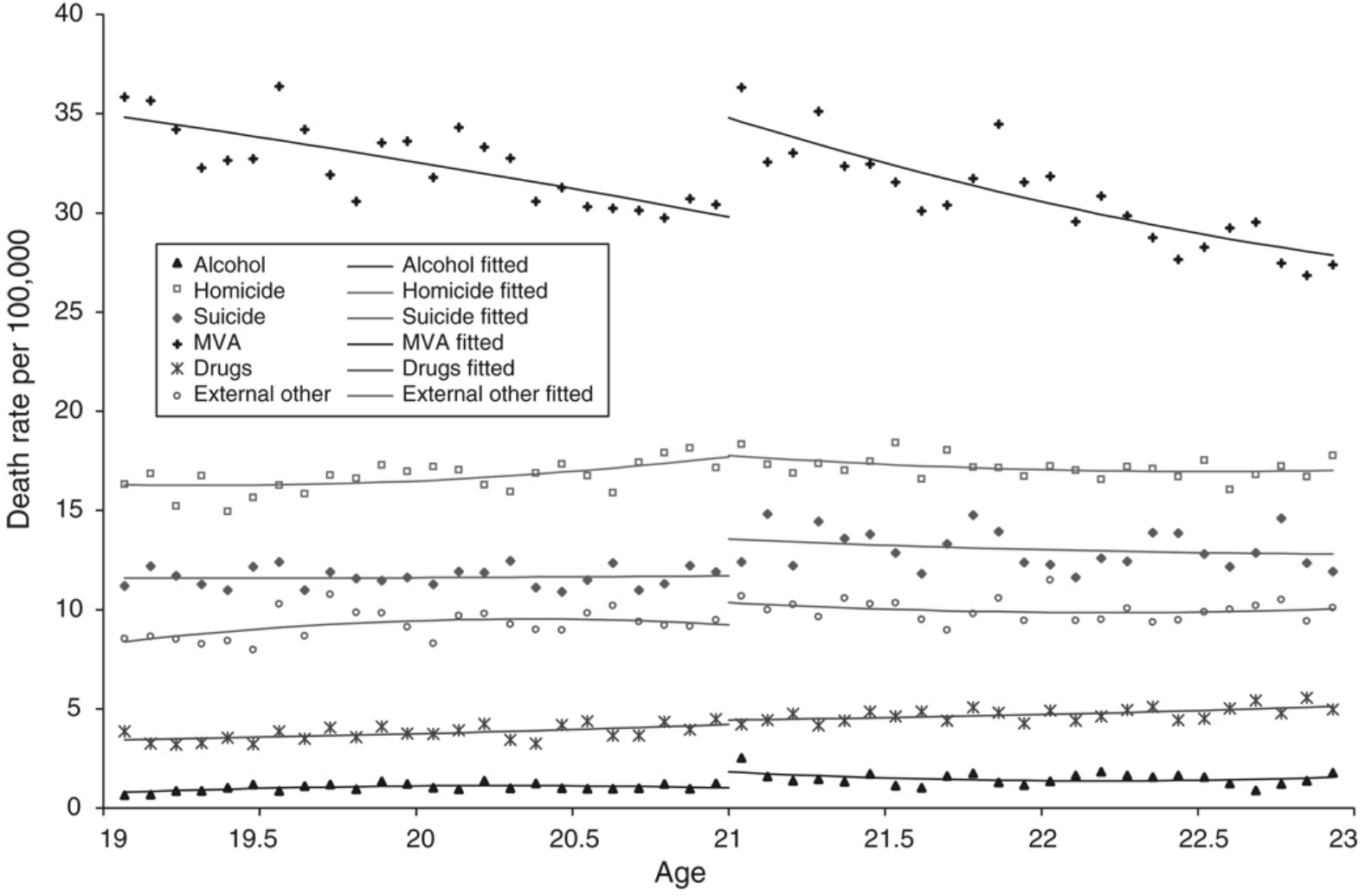
A simple example of IV

- Want to estimate how air travel revenue responds to change in price
- But price is chosen by airlines
  - When times are good, airlines choose high prices
  - But when times are good people travel a lot
  - When times are bad, airlines choose low prices
  - Then find high prices predict high demand, and low prices predict low demand
- "times are good" is a confounding variable
- Need an instrument that moves price but does not *directly* affect travel, such as a tax, unionization, etc.

Regression discontinuity

- What is impact of class size on student performance?
  - Observed data is problematic since schools in wealthy areas might have small class size . . .
  - In Israel, maximum class size is 40 students
  - So compare classes with an initial 40 students to those with an initial 41
- What is impact of ISP speed on housing values?
  - Valletti et al (2014): Compare people on borders of ISP service areas
- Impact of minimum legal drinking age on mortality
  - Compare 20.5 year olds to 21.5 year olds

Mortality by age and type



Carpenter and Dobkin (2009).

Difference in differences

- s_{TA} = sales after treatment in treated groups
- s_{TB} = sales before treatment in treated groups
- s_{CA} = sales after treatment in control groups
- s_{CB} = sales before treatment in control groups

	treatment	control	counterfactual
before	s_{TB}	s_{CB}	s_{TB}
after	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

- actual - counterfactual $= (s_{TA} - s_{TB}) - (s_{CA} - s_{CB})$
  - Ratio of ratios or use logs
  - Sampling distribution using bootstrap
  - Regression formulation with additional predictors
- D-in-D is just a simple model of the counterfactual
- Estimate of impact of treatment on the treated

What next?

- Other approaches
  - Structural models (economists): multi-equation IV
  - Propensity scores (Rubin): probability of treatment
  - Graphical models (Pearl): identification
- Further reading
  - Angrist and Pischke (2011): *Mastering 'Metrics*
  - Angrist and Pischke (2009): *Mostly Harmless Econometrics*