Causal Inference Meets Big Data

Hal Varian

Chief Economist,
Google
@halvarian

Causal Inference in Economics and Marketing: An Elementary Introduction

Hal Varian
Google,Inc.

Proc Nat Acad Sci 2016

February 3, 2017

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
- Linear model: $y_c = bx_c + e_c$
  - Center data to eliminate the constant
  - e_c = error in city c (cumulative effect of omitted predictors)
- Goal: estimate *causal* effect of changing x_c
  - E.g., what would happen if we spent 10% more in every city?

Naive approach

- Run a least-squares regression of y_c on x_c
- When do we get a “good” estimate of b?
  - $b^{LS} = \mathrm{cov}(x, y)/\mathrm{cov}(x, x)$
  - $b^{LS} = \mathrm{cov}(x, bx + e)/\mathrm{cov}(x, x) = b + \mathrm{cov}(x, e)/\mathrm{cov}(x, x)$
  - Need $\mathrm{cov}(x, e) = 0$
- But this is unlikely when x is chosen by decision makers
- They will choose x based on factors they observe but the analyst doesn’t. Many of these factors will affect y directly as well.

- Best client for the analyst is someone who is totally incompetent and makes choices randomly!

Example: 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
- Model "revenue $= 10 \times$ spend" fits the data perfectly

Example: 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
- Model "revenue $=$ 10 $\times$ spend" 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 movie revenue directly
  - . . . and indirectly, via ad spend chosen by distributor
- "Interest in surfing" is an example of a confounding variable.
- Confounding variables are relevant omitted variables (i.e. they help predict y) that are also correlated with the other predictors, x.
- Very common in models involving human choice since decision makers typically observe important factors that the analyst does not observe

Observational data v. designed experiments

- In true experiments, predictors are independent of error term by design
- In observational social science data, predictors are unlikely to be independent of error term
  - Predictors are usually correlated with each other
  - Everyone knows that including a new predictor in a regression on observational data will typically affect the coefficients of included predictors
  - Yet we blithely assume the error term—comprised of many omitted predictors—is independent of included predictors
- If we have "all" important predictors, this problem goes away
- But when does that ever happen?

Prediction v causal inference

- Suppose you have more observations from *the same data generating process*
- If your goal is predicting revenue in additional cities, regression is fine
  - Cities with little interest in surfing will likely have small ad expenditure *and* small sales
- If your goal is predicting what would happen if spend is increased by 10% across the board, you have a problem.
  - Why? Because this is a different data-generating process.
- If you are interested in using data for policy, you generally ask what would happen *if the data generating process changes.*
- Need causal inference.

Examples

How does fertilizer affect crop yields? Farmers choose fertilizer application based on land quality.

How does education affect income? Students with wealthy parents or high ability tend to acquire both more education and more income.

How does health care affect income? Those who have good jobs tend to have good 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. Challenge: what can we learn from observational data, where there is no explicit experimentation?

- Important consideration: is the treatment going to be *imposed* on the population, or *chosen* by members of population?
  - Imposed: impact of treatment on population
  - Chosen: impact of treatment on those who choose treatment

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

- LHS is observed, so you only need to estimate one of the terms on the RHS to infer the other term.
- 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

- LHS is observed, so you only need to estimate one of the terms on the RHS to infer the other term.
- 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

- LHS is observed, so you only need to estimate one of the terms on the RHS to infer the other term.
- If treatment is randomly assigned
  - Selection bias is zero.
  - Treated are random selection from population, so impact on treated = impact on population

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

- LHS is observed, so you only need to estimate one of the terms on the RHS to infer the other term.
- If treatment is randomly assigned
  - Selection bias is zero.
  - Treated are random selection from population, so impact on treated = impact on population
  - Otherwise need a model to estimate who was treated
  - See Angrist and Pischke (2011) for more

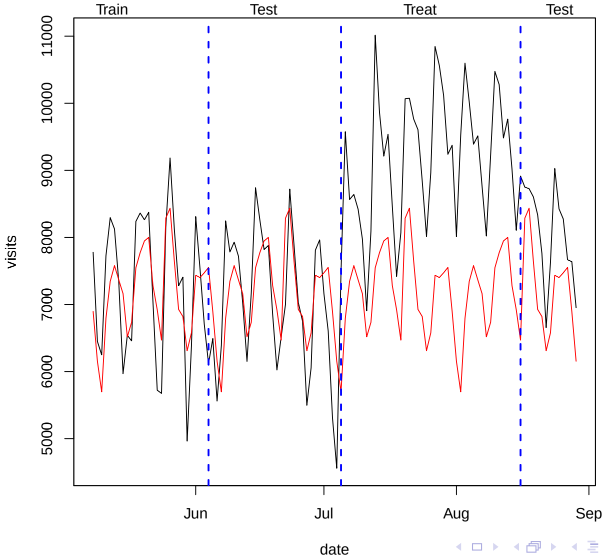
Ways to estimate causal effects from observational data

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

Randomized experiments

- Want to treat subjects (at randomly chosen) times; compare to nontreated times
- Want to treat (randomly chosen) subjects; compare to nontreated subjects
- Use pre-experiment behavior to build a predictive model for outcomes. Use this model to estimate counterfactual: what would have happened without treatment.
- Train, test, treat, compare
    - Train: a model on some of your data
    - Test: a model on holdout data (placebo experiment)
    - Treat: apply treatment to treatment group
    - Compare: outcome for treated to the counterfactual prediction

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 a way that is "as good as" random

Example: Super Bowl and ad impact

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

Instrumental variables

- A systematic way to think about natural experiments
- Have a model $y_c = bx_c + e_c$, but are worried about confounders
- Can we find an instrumental variable?
  - Something that moves x_c but is independent of e_c?
- Example from Super Bowl: $z =$ home cities of teams playing
- Ad impressions depend on home city: $x_c = az_c + d_t$
- Assume that $\text{cov}(z, e) = 0$. That is, the *only* way home cities affect ad spend is via fan effect
- Since ad spend is chosen months in advance, this is plausible
- $b^{IV} = \text{cov}(z, y)/\text{cov}(z, x)$

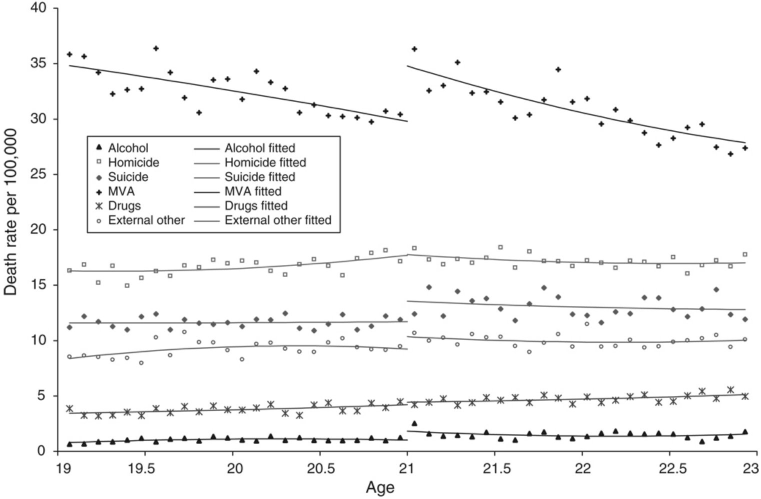
Another example of IV

- Want to estimate how air travel revenue responds to change in ticket 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
  - But when times are bad, people don't travel much
  - Then find high prices predict high demand, and low prices predict low demand
- "times are good" is a confounding variable
  - Find a good proxy (e.g., GDP) and/or
  - Find an instrument that moves price but does not *directly* affect travel, such as a tax, fuel cost, unionization, etc.

Regression discontinuity

- What is impact of class size on student performance?
  - Observed data is problematic since schools in wealthy areas tend to have small class size . . .
  - In Israel, maximum class size is 40 students
  - Angrist et al (1999): 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 house prices 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 $= (s_{TA} - s_{TB}) - (s_{CA} - s_{CB})$
  - Proportional model: ratio of ratios or D-in-D with logs
  - Sampling distribution: using bootstrap or regression
  - Regression formulation: can use additional predictors
- D-in-D is just a simple model of the counterfactual to estimate impact of treatment on the treated

Nonlinear estimation of counterfactual

- Let y_{TF} = counterfactual for the treated group: what would have happened if they had not been treated
- D-in-D rests on the model $y_{TF} - y_{TB} = y_{CA} - y_{CB}$. A natural generalization with additional predictors (x) is:

$$\begin{aligned} y_{TA} &= G(y_{TB}, x_{TB}) && (1) \\ y_{CA} &= G(y_{CB}, x_{CB}) && (2) \end{aligned}$$

- Train using subsets of control group, test using remainder of control group, then use estimated $G(\cdot)$ to predict counterfactual.
- Just like "train-test-treat-compare"

What next?

- Challenges
  - Use machine learning to estimate counterfactual
  - Extend ML methods to panel and time series (non-IID) data
- Additional topics
  - Structural models (economists): multi-equation IV
  - Graphical models (Pearl): identification and more
  - Propensity scores (Rubin): probability of treatment assignment
- Further reading
  - Angrist and Pischke (2011): *Mastering 'Metrics* [undergrad]
  - Angrist and Pischke (2009): *Mostly Harmless Econometrics* [grad]