Machine Learning and Econometrics

Hal Varian
Jan 2014

Definitions

Machine learning, data mining, predictive analytics, etc. all use data to predict some variable as a function of other variables.

- May or may not care about insight, importance, patterns
- May or may not care about *inference*---how y changes as some x changes

Econometrics: Use statistical methods for prediction, inference, *causal* modeling of economic relationships.

- Hope for some sort of insight, inference is a goal
- In particular, *causal* inference is goal for decision making

What econometrics can learn from machine learning

"Big Data: New Tricks for Econometrics"

- train-test-validate to avoid overfitting
- cross validation
- nonlinear estimation (trees, forests, SVGs, neural nets, etc)
- bootstrap, bagging, boosting
- variable selection (lasso and friends)
- model averaging
- computational Bayesian methods (MCMC)
- tools for manipulating big data (SQL, NoSQL databases)
- textual analysis (not discussed)

Scope of this talk: what machine learning can learn from econometrics

I have nothing to say about

- Computation
- Modeling physical/biological system (e.g., machine vision, etc.)

Focus is entirely on

- Causal modeling involving *human choices*
- Economic, political, sociological, marketing, health, etc.

What machine learning can learn from econometrics

- non IID data (time series, panel data) [research topic, not in textbooks]
- causal inference -- response to a treatment [manipulation, intervention]
  - confounding variables
  - natural experiments
  - explicit experiments
  - regression discontinuity
  - difference in differences
  - instrumental variables

Note: good theory available from Judea Pearl et al, but not widely used in ML practice. The techniques described above are commonly used in econometrics.

Non IID data

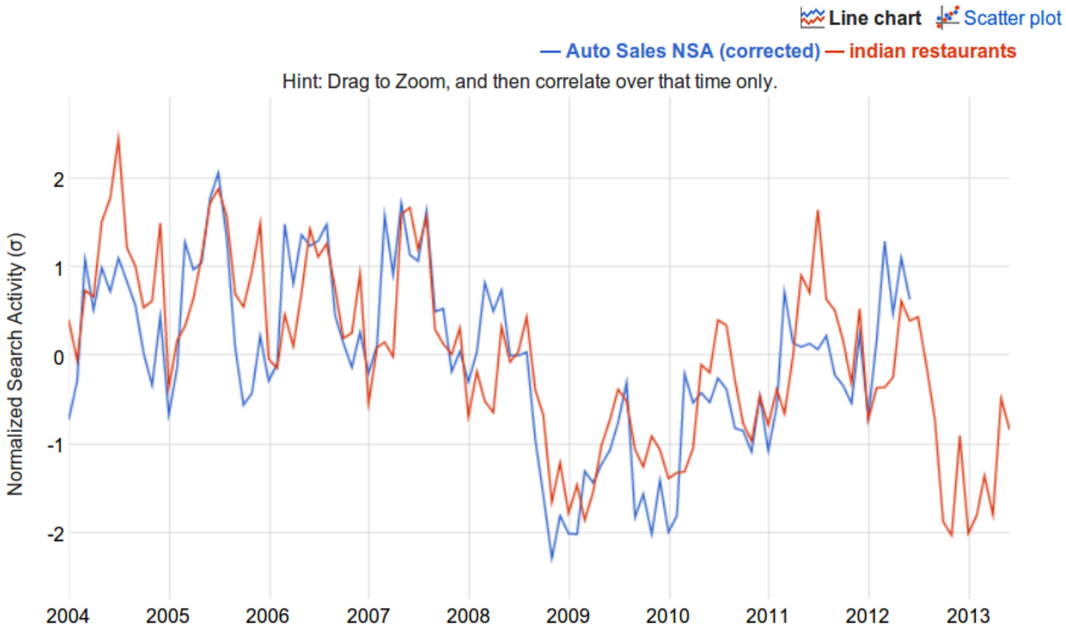
Time series: trends and seasonals are important; cross validation doesn't work directly; analog is one-step ahead forecasts; spurious correlation is an issue (**auto sales**); whitening data as a solution: decompose series into trend + seasonal components, look at deviations from expected behavior.

Panel data: time effects and individual effects.

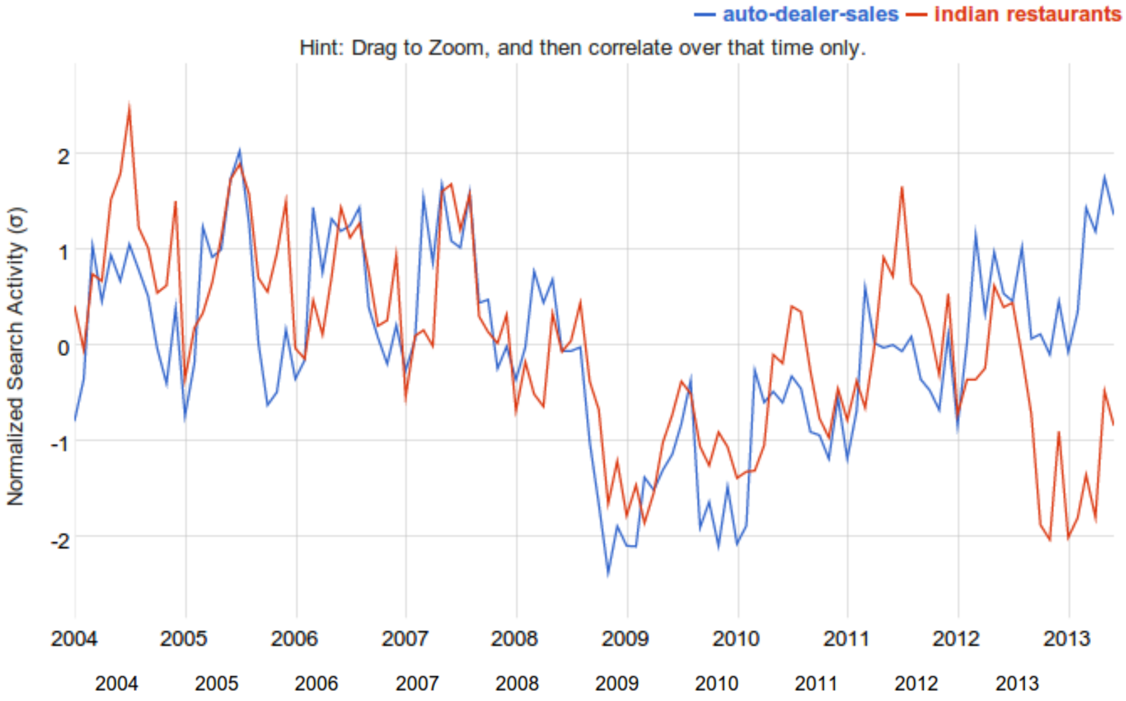
- Example: anomaly detection
- Simplest model: $y_{it} = F_i + bx_{it} + e_{it}$
  - Fixed effects
  - Random effects

NSA auto sales and Google Correlate to 2012

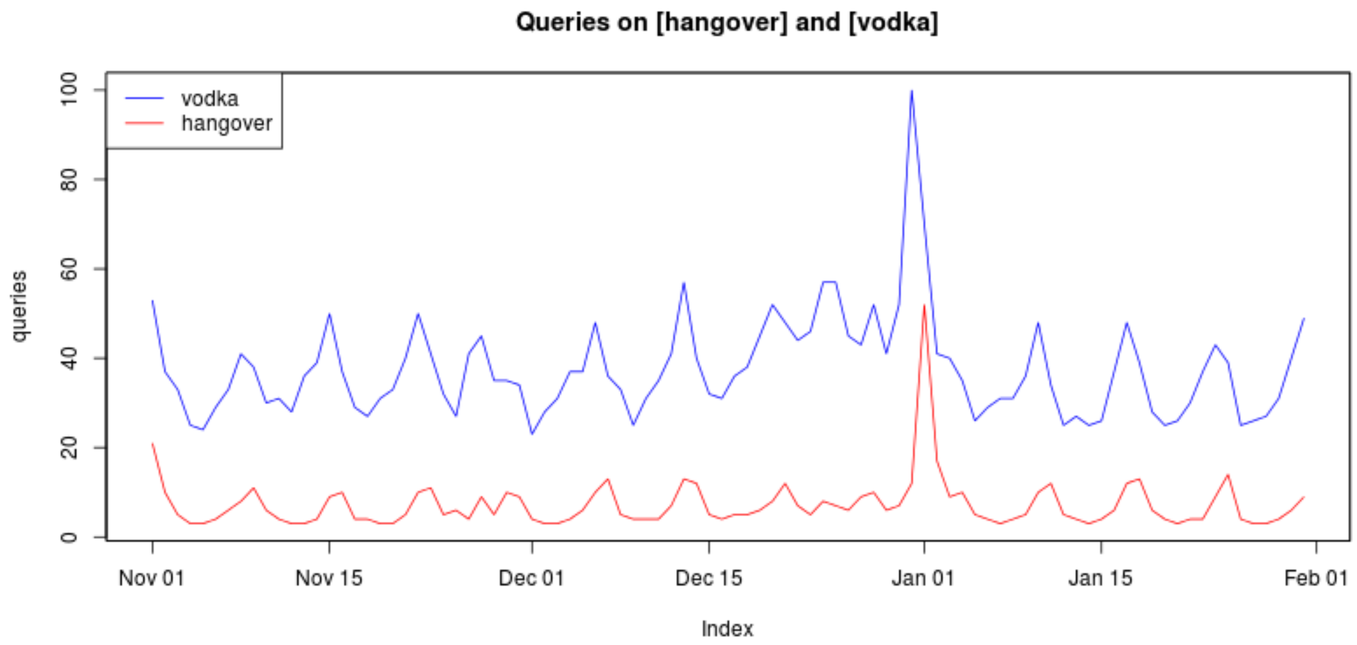
User uploaded activity for **Auto Sales NSA (corrected)** and United States Web Search activity for **indian restaurants** (r=0.7848)



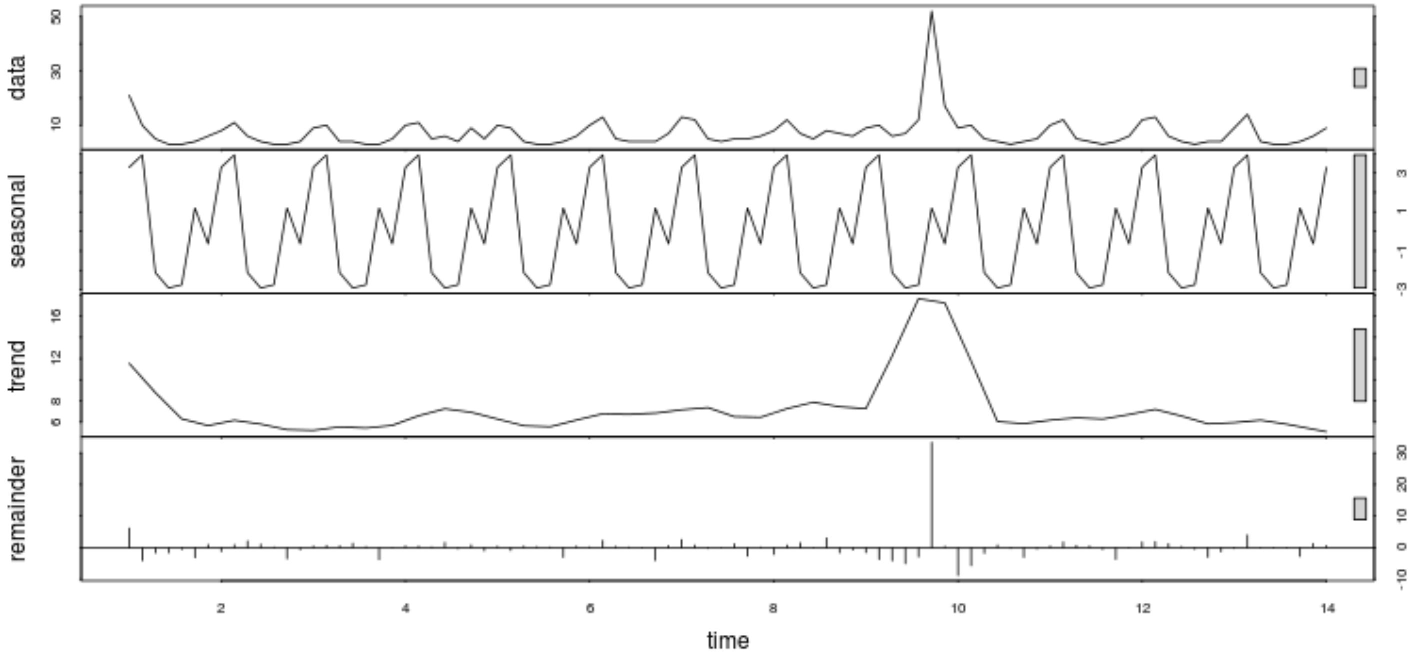
NSA auto sales and Google Correlate through 2013



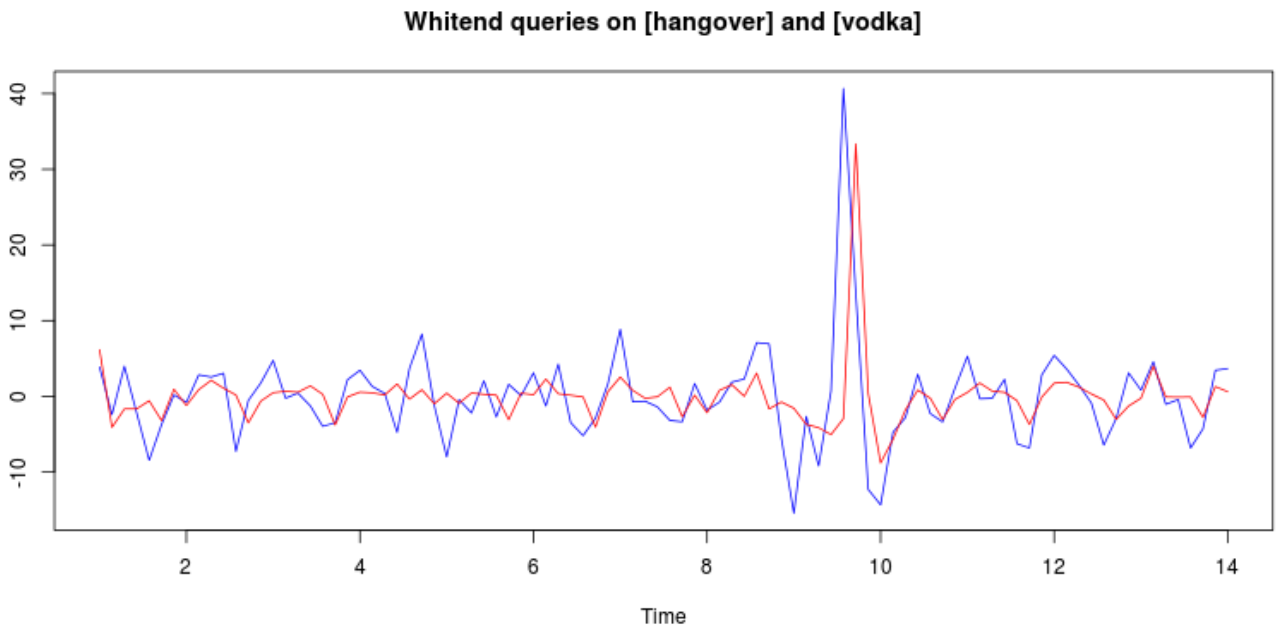
Queries on [hangover] and [vodka]



Seasonal decomposition of [hangover]



Does [vodka] predict [hangovers]?



Example of simple transformations for panel data

$y_{it} = F_i + bx_{it} + e_{it}$

$\overline{y_i} = F_i + b\overline{x_i} + \overline{e_i}$ average over time for each individual i

$y_{it} - \overline{y_i} = b\,(x_{it} - \overline{x_i}) + (e_{it} - \overline{e_i})$ subtract to get “within estimator”

Anomaly detection: look for deviations from typical behavior *for each individual.*

Also, panel data is helpful for causal inference as we will see below.

Causality

“More police in precincts with higher crime; does that mean that police cause crime?”   Policy decision: should we add more police to a given district?

“Lots of people die in hospitals, are hospitals bad for your health?” Policy decision: should I go to hospital for treatment?

“Advertise more in December, sell more in December.” But what is the *causal* impact of ad spending on sales? Policy decision: how much should I spend on advertising?

Important considerations: counterfactuals, confounding variables

Counterfactuals and causality

Crime. It is likely data was generated by a decision rule that said "add more police to areas with high crime." This may have reduced crime over what it ***would have been***, but these area may still have had high crime.

Hospital. If I go to hospital will be better off than I ***would have been*** if I didn't go?

Advertising. What would my sales be if I ***would have advertised*** less?

Confounding variables 1

Confounding variable: unobserved variable that correlates with both y and x.

sales = f(advertising) + other stuff

Xmas is a confounding variable but
there are potentially many others

In this case, the solution is easy: put Christmas (seasonality) in as an additional predictor. But there are many other confounding variables that the advertiser can observe that the analyst doesn't. (E.g., product quality.)

Confounding variables 2

Commonly arise when human choice is involved

- Marketing: advertising choice, price choice
- Returns to education: IQ, parents' income, etc. affect both choice of amount of schooling and adult earnings
- Health: compliance with prescription directions is correlated with both medication dosage and health outcome

Omitted variables that are not correlated with x just add noise, but confounders *bias* estimates

What do you want to estimate?

Causal impact: change in sales associated with change in advertising expenditure *everything else held constant*?

or

Prediction: Change in sales you would expect to observe when *advertising expenditure changes* ?

If you want to make a decision, the former is what is relevant. If you want to make a prediction the latter is relevant.

Ceteris paribus vs mutatis mutandis

- Ceteris paribus: causal effect with other things being held constant; partial derivative
- Mutatis mutandis: correlation effect with other things changing as they will; total derivative

- Passive observation: If I *observe price change* of dp, how do I expect quantity sold to change?
- Explicit manipulation: If I *explicitly change price* by dp, how do I expect quantity sold to change?

“No causation without manipulation” Paul Holland (1986)

Big data doesn’t help

You can have a great model of the relationship between police and crime, but won’t answer question of what happens if you intervene and add more police. Why?

- Data generating process is different.
- Observed data generated by a “more crime -> more police” rule but now want to know what happens to crime when you add more police
- When predictors are *chosen* by someone (as in economic examples), they will often depend on other omitted confounders. Xmas example

Estimating a demand function

Model: sales ~ price + consumer income + other stuff
Policy: if I manipulate price, what happens to sales?
Observe: historical data on sales and price

Possible data generating process

- When times are good (boom) people buy a lot and aren't price sensitive, so merchants raise prices.
- When times are bad (recession) people don't buy much and are price sensitive, so merchants cut prices.

Result: high prices associated with high purchases, low prices associated with low purchases. Problem: "income" is confounding variable. Solutions: 1) bring "income" into model (but what about other confounders?), 2) do a controlled experiment, 3) find a natural experiment (e.g., taxes, supply shocks).

One solution

Find other variables that affect price that are independent of confounding variables.

sales ~ price + consumer income + other stuff
price ~ markup x cost [markup is chosen, cost is exogenous]
price ~ pre-tax price + sales tax [price is chosen, sales tax exogenous]

Here changes in cost could be due to weather (coffee), global factors (oil), tech change (chips), etc. Sales tax could vary across time and state. As long as these variables are *independent* of the demand-side factors, we should be OK.

Variables like this are called *instrumental variables* since they are an "instrument" that moves predictor exogenously, similar to the manipulation you are considering.

What is the intended use of demand estimation?

Tell consumers what to expect prices to be in the future?

- Want to model historical relationship
- Estimate relationship “mutatis mutandis”
- Oren Etzioni, et al paper: “To buy or not to buy: mining airfare data to minimize ticket purchase price”

Tell managers what will happen if they *manipulate* price?

- Want to model causal relationship
- Ideally, run an experiment
- Alternatively, find a natural experiment and/or instrument (fuel price?)
- Estimate relationship “ceteris paribus”

You usually want the causal impact for policy

If you are using data to make decisions, you usually want the causal impact.

Examples from: James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning*, 2013

- Marketing
  - "What effect will changing the price of a product have on sales?"
  - Not just an inference problem, but a *causal* inference problem
  - Generally there will be confounding variables in such a problem
- Education
  - income ~ education + seniority (Mincer equation)
  - For policy (e.g., changing schooling requirements) you want a *causal* estimate of education effect, but you won't get that from historical data since people choose education and choice depends on ability, family income, etc.

Practical techniques for causal inference

Need some sort of exogenous change in x to estimate causal effect

1. Gold standard: true randomized treatment-control experiment
   a. Google, Bing, Yahoo, Facebook, etc
2. Natural experiments which may or may not be randomized
   a. Example: draft lottery, Oregon healthcare lottery, etc

May need to model: *who gets treated* (a prediction problem)

- Random, volunteers, chosen, invited …
- Question to ask: how does proposed policy relate to experiment?

Definitely need to model: *counterfactual* (prediction problem)

- What would have happened to the treated if they weren't treated?
  - Simple: they would look like the control on average
  - Complex: more elaborate predictive model

Role of counterfactual

Should I recuperate from an operation in hospital or home?
Lots of people die in hospitals!
health(went to hospital) - health(stayed home) looks bad
But correct comparison is:
health(went to hospital) - health(if they *had stayed* home)

Fundamental equation of causal analysis (Angris & Pischke)

health(went to hospital) - health(stayed home) =	[observed]
health(went to hospital) - health(if they had stayed home)	[treatment effect]
+ health(if they had stayed home) - health(stayed home)	[selection bias]

Fundamental equation in causal modeling

observed difference in outcome = average effect of treatment on treated + selection bias

Randomization: solves selection bias, so observed difference is average effect of treatment on (a random sample of) the population. So observed difference in controlled experiment gives you effect of treatment on population.

But you may be interested in impact of treatment on a subjects chosen for treatment in some other way (volunteers, selected, etc.)

Better prediction give you better causal inference

Predicting who is selected

- No problem if treatment random, otherwise an issue
- Probability selected for treatment ~ observables (test scores)
- Probability selected for treatment ~ unobservables (personality)

Predicting the counterfactual

- Can you predict outcome *before or after* treatment? (I.e., those who are going to be treated.)
- Can you predict outcome for those not treated (control)?

Both of these problems are prediction problems and can benefit from better predictive analytics

Example

Treatment: advertise more
Outcome: sales?

1. Predict company X sales using Google Trends category-level query data using time-series model
2. Compare actual sales to predicted sales prior to treatment
3. Compare actual sales during advertising treatment to predicted sales during that period
4. Difference is causal effect of advertising

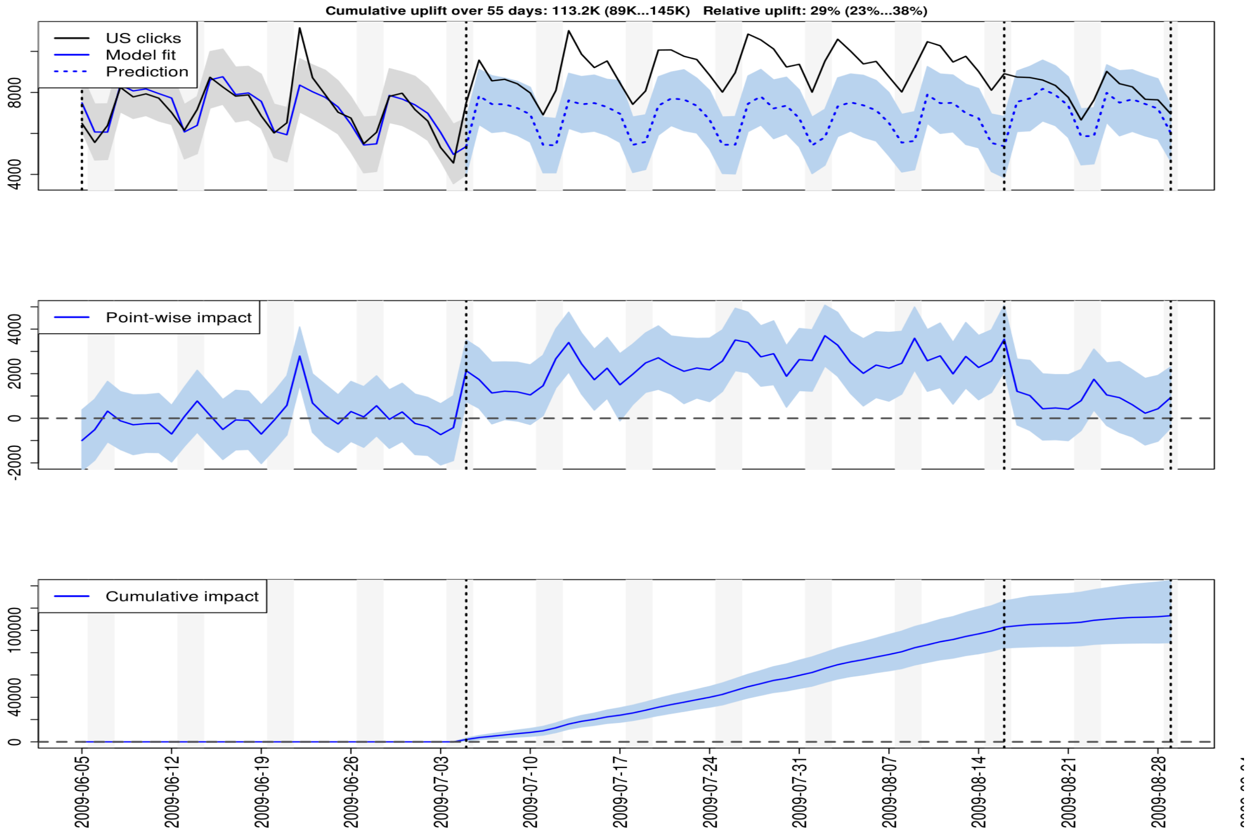
"Inferring causal impact using Bayesian structural time-series models"

Google

Observed + counterfactual using Trends data

Difference

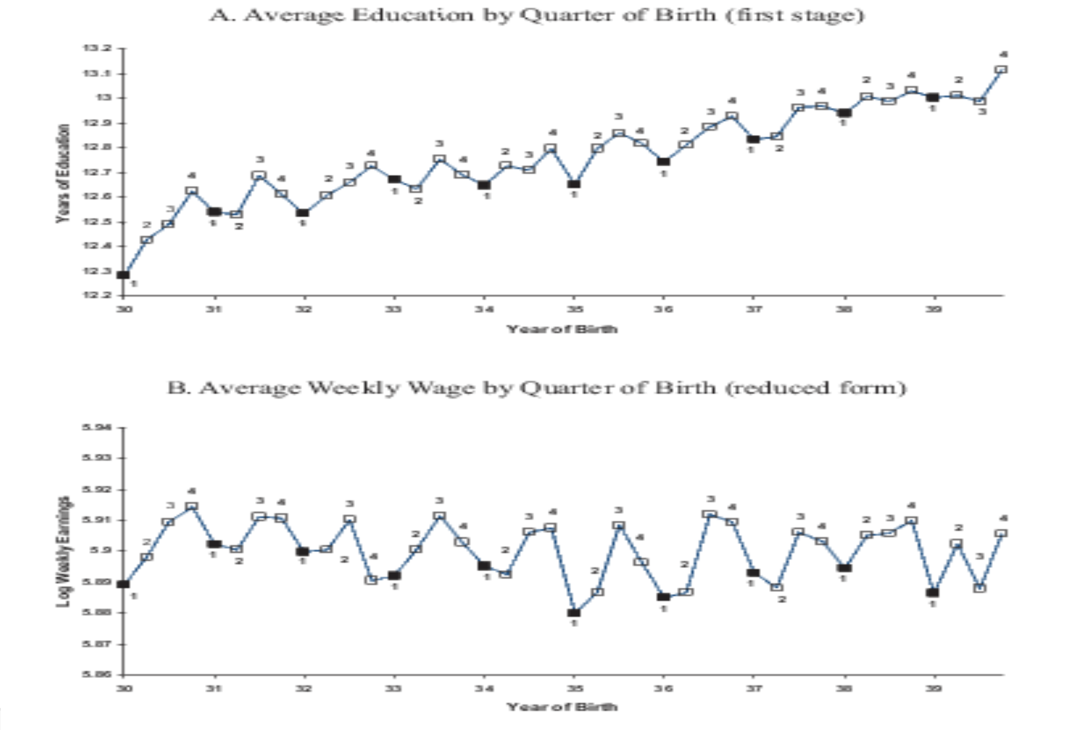
Cumulative difference



Natural experiments

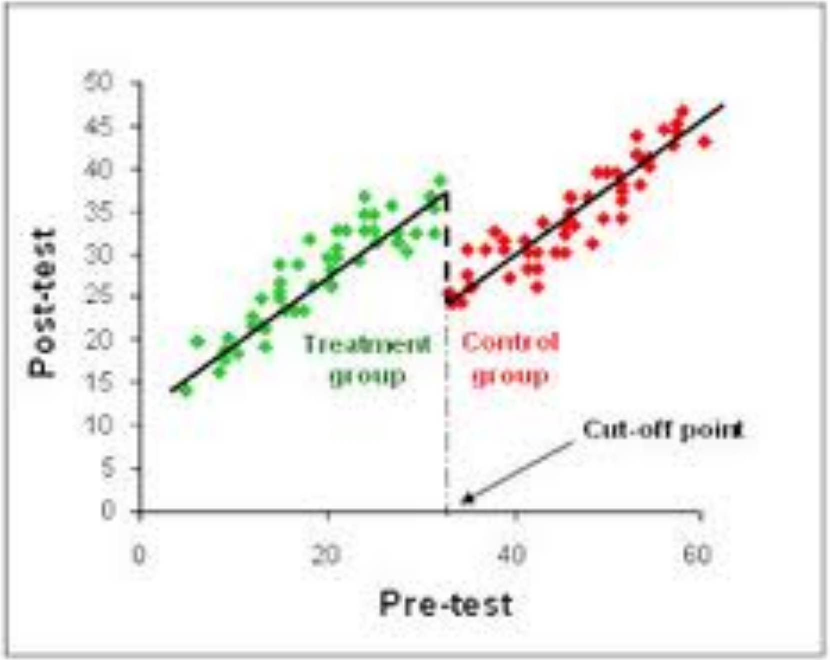
Impact of police on crime
    Terrorism alerts in DC (Klick-Tabarrok)
Impact of veteran status on future income
    Draft lottery (Angrist)
Impact of education on income
    quarter of birth (Angrist-Krueger) [see next slide]
Impact of medical care on health outcomes
    Oregon lottery (Finkelstein, et al.)
Impact of ad impressions on purchases
    Superbowl (Stephens-Davidowitz et al)

Education by quarter of birth, wage by quarter of birth



Regression discontinuity (and kinks)

Treatment applied depending on some score
- Class size in Israel (Angrist-Lavy)
  - selection bias: students
  - max class size = 40
- Position effect of ads
  - selection bias: placement
  - ad rank = bid x ad quality
- Thresholds show up in lots of policies
- ...and also lots of algorithms (trees)



Examine outcomes for subjects near thresholds

Perhaps can apply regression discontinuity approach in tree models?

Difference between subjects slightly above and below thresholds in a tree model could be just due to random noise.

In fact, we could also *explicitly randomize* so we are confident it is random noise. Adding the noise has low expected cost.

Randomization is too important to be left to chance!

1. Write code so that you can experiment

```
if(x > some parameter) do something
```

vs

```
if(x > some parameter + e) do something
```

2. Can fine tune the parameter

3. Once you understand responses you can simulate: “rerun yesterday” with different parameter values

Experiments are not just for high tech

“How big of a change is it that you can now gather so much information about your customers?”

It’s powerful. About 50 percent of our transactions occur on our Panera (PNRA) card. So we have individual information on individual purchase activities. That means that when we’re going to do something, we can actually look at its impact on behavior and build our marketing and our campaigns around individual consumers or small groups of consumers as opposed to the mass market. Everybody is going to get something a little different depending on what their behavior is.

Panera Bread CEO Ron Shaich

Once you understand behavioral responses simulation can be useful

Partially baked idea

Build in experimentation at compile time.
Every language has a statement like this:

```
const price 5.0
```

Consider a statement like this:

```
param price 4.00:6.00 inc .50
outcome sales
```

Compiler generates code to assign random value to “price” at run time and keeps track of outcome variables “sales”

Can run compiler on existing code, making it more powerful
Make experimentation the path of least resistance.

Partially baked idea, further considerations

```
param color (blue, yellow, green)
outcome clicks
```

- Can run many experiments at once if they don’t interfere
- Can specify layers that make it easy to avoid interference
- Not just for computer’s internal environment (ATLAS software optimization tool) , but also on responses in external environment (such as human responses)

Instrumental variables

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

- Problem: confounding variables that affect both e and X
- However, *some* part of the variation in X_i may be independent of error
- Can we find something that changes X, but does not affect error?
  - Estimate coffee demand elasticity: look at supply shifts due to weather
  - Examples: demand for cigarettes, cigarette taxes by state
    - Overall tax rate might be an instrument
- Instrument Z has two properties
  - Z is highly correlated with X [testable]
  - Z is not correlated with the error [not testable]
- Z only affects y via the X variable
- estimate dX/dZ and dY/dZ and then take ratio to get dY/dX

Instrumental variables

Best instrument is randomization: use a coin flip to choose X

If it is random, it isn't correlated with omitted variables (error)

But there may exist some variable that affects X and isn't correlated with the error; that is, the instruments affects y only via its effect on X.

Elaborate econometric theory to deal with instruments.

Difference in differences

Two groups: treatment, control
Two time periods: before and after treatment
(Treatment may or may not be randomly assigned)

after - before = $(T_A - T_B)$

treatment - control = $(T_A - C_A)$

DiD = $(T_A - T_B) - (C_A - C_B)$

DiD: impact of treatment on treated, adjusted by control

	Treatment group	Control group
Before	T_B	C_B
After	T_A	C_A

Regression interpretation

$treat_i$ = 1 if treated, 0 if control
$after_i$ = 1 if after, 0 if before
β_3 = treatment effect

$$y_i = \beta_0 + \beta_1\, treat_i + \beta_2\, after_i + \beta_3\, treat_i{*}after_i + \text{other things that affect } y + e_i$$

	Treatment	Control	Difference
Before	$\beta_0 + \beta_1$	β_0	β_1
After	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_2$	$\beta_0 + \beta_3$
Difference	$\beta_2 + \beta_3$	β_2	β_3

How much should you worry about selection bias?

Economics

- Experiment to determine policy change for population
- Impact of treatment on population
- Worry about selection bias: want random sample

Business

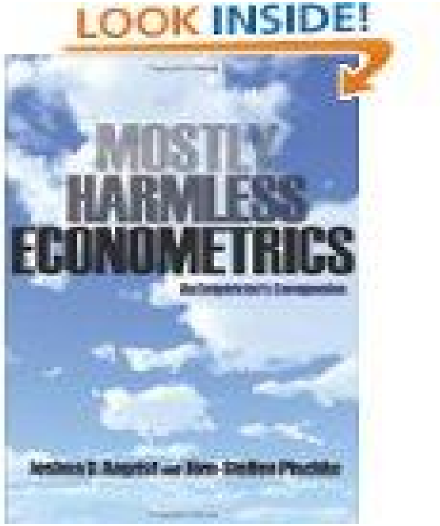
- Impact on advertisers who choose to use new feature or service
- Impact of treatment on *those who choose to be treated*
- Not necessarily worried about selection bias (but may be worried about early adopter bias)

What are lessons?

1. Observational data (usually) can't determine causality, no matter how "big" it is.
2. Causal inference is what you want for decisions
3. Treatment-control with random assignment is gold standard
4. Sometimes you can find natural experiments, discontinuities, instrumental variables, DiD, etc.
5. Prediction is critical to causal inference: predict who is selected, predict counterfactual
6. Interesting research possibilities in systems optimized for testing

Two very good introductory books to follow up

Mostly Harmless Econometrics
Joshua Angrist and Jörn-Steffen Pischke



Introduction to Statistical Learning
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani