Building a Robot Judge: Data Science for the Law

12. Causal Inference with High-Dimensional Data

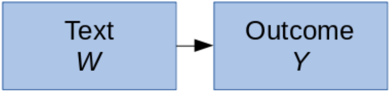
April 14, 2019

Research Design

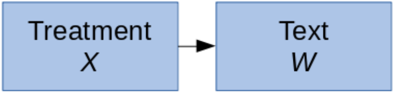
- The goal of social-science research with big data is the same as other social-science research:
  - provide credible tests of social-science hypotheses
  - estimate policy parameters to inform policymakers

Setup

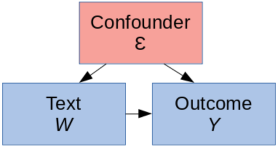
- W, vectorized text
- Y, outcome from the text
  - e.g., the facts of the case W determine the verdict Y



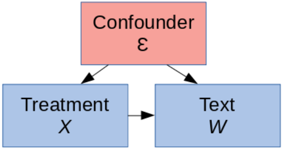
- X, treatment affecting the text
  - e.g., judge political preferences X affect the written opinion W



Empirical Problem: Confounders (ϵ)

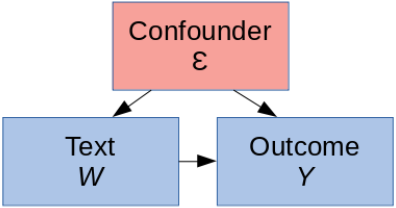


- judge decides Y based on defendant characteristics ϵ as well as case facts W



- judge writes opinion W based on characteristics ϵ as well as her ideology X.
- Key point: **a variable is a confounder only if it affects both sides of a regression** (both W & Y, or both X & W).

Econometrics



- We would like to learn

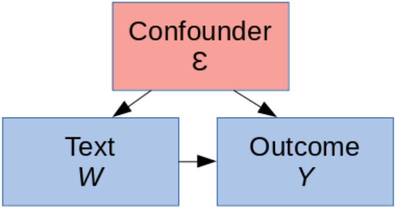
$$f(W;\theta) = \mathbb{E}\{Y|W\}$$

  the conditional expectation function for y, where θ represents the true parameter vector.

  - $f(\cdot)$ and θ describe the arrow from W to Y.

- If we assume linearity and run OLS, the estimates for $\hat{\theta}$ are biased because of the confounder.

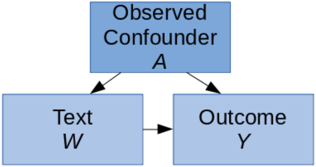
Econometrics + Machine Learning



$$f(W;\theta) = \mathbb{E}\{Y|W\}$$

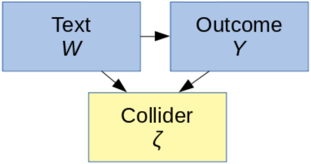
- We could take a machine learning (ML) approach and learn a nonlinear approximation $\hat{f}(W;\theta)$ to predict Y in held-out data.
  - If we obtained more documents W_i for new individual i, we could form a good prediction about the associated Y_i.
- But the ML estimates $\hat{\theta}$ do *not* have a causal interpretation.
  - i.e., if the case facts W were experimentally changed, $\hat{\theta}$ would not provide a counterfactual prediction about how the associated outcome Y would change.

Confounder is Observed



- If confounder A is observed, problem solved:
  - include A in your model, or residualize W and Y on A before estimation.
- This is rarely plausible in practice.

Colliders or Bad Controls

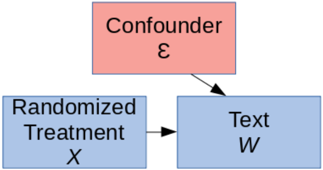


- ζ, colliders (or as most economists would say, "bad controls"), are a third variable that is affected by both your treatment and your outcome.
  - For example, let ζ be the length of the prison sentence, which is affected by the case facts W and the verdict Y.
- **Don't control for colliders**! It introduces bias.
  - put differently, don't condition on a joint outcome.

Empirical Strategies

- In the presence of unobserved confounders, estimating causal parameters presents a significant challenge.
- Modern empirical economics puts an emphasis on obtaining causal estimates using **empirical strategies** or **research designs**.
  - this is why Google/Amazon/etc. hire many PhD economists.
- This involves running a controlled experiment, or approximating one using features of observational data.

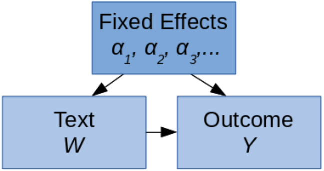
Randomized Experiments



- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - If treatment X is randomly assigned, it is uncorrelated with the confounder by construction $(X \perp \epsilon)$.
- E.g.:
  - randomly assign judges from $X \in \{\text{Party } 1, \text{Party } 2\}$ to cases.
  - The causal effect is the average difference in their written decisions, $\mathbb{E}\{W|X=1\} - \mathbb{E}\{W|X=2\}$.

Fixed Effects

- What if all confounders are at the group level?
  - e.g., (unobserved) defendant characteristics ϵ are the only deconfounder for the verdict, and those are constant over time.
- If same defendant i is observed over multiple cases, can control/adjust for defendant characteristics by including a fixed effect α_i for each i.

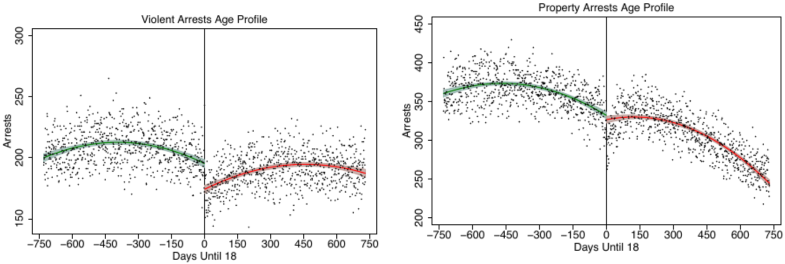


  - in the data, add a dummy variable equaling one for i's cases.
- Equivalently (almost), can center (de-mean) predictors W and outcome Y by defendant.
  - With multiple fixed effects (e.g., defendant, judge, and year), can **residualize**: project predictors W and outcome Y onto matrix of dummy variables, and take residuals $\tilde{W} = W - \hat{W}$ and $\tilde{Y} = Y - \hat{Y}$ for use in model training.

Regression Discontinuity Design (RDD)

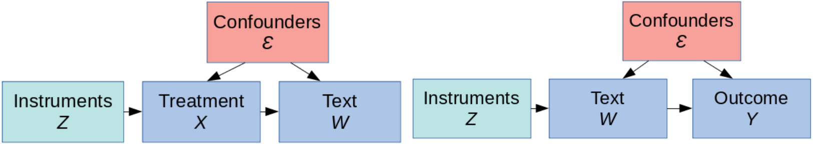
- RDD's exploit threshold rules, where individuals are assigned to treatment if a continuous variable is above some discrete cutoff.
  - The idea is to exploit randomness around this threshold.
- Example "running variables":
  - Score in entry exams, effect of barely making it into college.
  - Income, effect of barely being eligible for poverty subsidy
  - Votes in an election, effect of barely getting a Republican (relative to a Democrat)

Increased Penalties at 18 → Less Crime



Lovett and Zue (2018). California data.

Instrumental Variables



- A valid instrument Z is related to the treatment but not otherwise correlated with the outcome
  - left panel: Z affects X but orthogonal to ϵ.
  - right panel: Z affects W but orthogonal to ϵ.
- First stage:
  - Predict $\hat{X}(Z)$ or $\hat{W}(Z)$.
  - If Z is high-dimensional, use regularized model.
  - Assess relevance with F-statistic or out-of-sample R^2 (Hartford et al 2017).
- Second stage:
  - Predict $W(\hat{X}(Z))$ or $Y(\hat{W}(Z))$

Random Assignment of Judges → Z

- Let Z be a high-dimensional set of characteristics of judges, e.g. political party, cohort, writing style.
- Let W be the text features of the current case.
- Let Y be the outcome, e.g., whether the case is appealed.

- Instrumental variables system:

$$W = g(Z), Y = f(W)$$

  - form ML predictions of $\hat{g}(\cdot)$
  - use those predictions $\hat{W}$ in predicting $\hat{f}(\cdot)$

Fixed Effects and Sparsity

- Recall that standardizing data breaks sparsity structure in high-dimensional sparse data.
  - fixed effects or other residualization steps will also do this.
- Some solutions:
  - Can residualize outcomes but not predictors.
  - Can use use first-differences rather than fixed-effects.
  - Can center on the mode after residualizing.
- Fixed-effects transformations don't have the same interpretation with non-linear models. Not enough research on this yet.

Matching / Synthetic Control

- **matching**: use covariates to find matching individuals
- **synthetic control**: construct a synthetic "match" from a weighted average of other individuals (based on covariates).
- Note:
  - Equivalent to controlling for many observed confounders.
- Can imagine the text documents associated with individual or groups as a set of covariates for matching
  - e.g., text features from the criminal history of each defendant.

Adjusting for confounding with text matching

Roberts, Stewart, and Nielsen (2018)

- Lots of governments try to control online information
- But, censoring the whole internet is hard (# of bloggers $\gg$ # of censors)
- Limited external enforcement $\rightsquigarrow$ self-policing

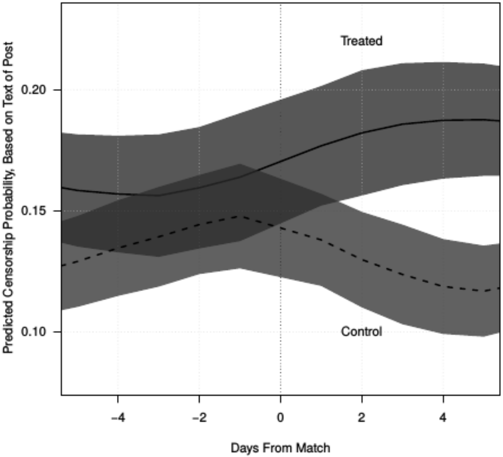
Application to online censorship in China

Roberts, Stewart, and Nielsen (2018)

- Construct a corpus of chinese blog posts, some of which are censored.
  - 593 bloggers, 150,000 posts, 6 months
- They use a variation of propensity score matching to identify almost identical blog posts, some of which were censored, and some of which were not.
- Outcome:
  - Using text of subsequent posts, measure how likely they are to be censored (how censorable)
  - Can see whether censorship has a deterrence or backlash effect.

Censorship has a backlash effect

Roberts, Stewart, and Nielsen (2018)



- Bloggers who are censored respond with more censorable content.

Double/Debiased ML

Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey (2017)

$$Y = \theta T + g(A) + \epsilon$$

- low-dimensional treatment T, high-dimensional set of (observed) confounders A: $T = m(A) + \eta$.
- Because of confounders, forming a prediction $\hat{Y} = \hat{\theta} T + \hat{g}(A)$ will be biased.

Double ML method

Chernozhukov, Chetverikov, Duflo, Hansen, Demirer, and Newey (2017)

1. Predict Y given A: $\hat{Y}(A)$, and T given A: $\hat{T}(A)$, using any ML method
2. Form residuals $\tilde{Y} = Y - \hat{Y}(A)$ and $\tilde{T} = T - \hat{T}(A)$
3. Regress $\tilde{Y}$ on $\tilde{T}$ to learn $\hat{\theta}$.

- Sample split:
  - Run (1) on sample a, then run (2) and (3) on sample b, to estimate $\hat{\theta}_a$
  - and vice versa (run (1) on sample b, and (2/3) on sample a), to learn a second estimate for $\hat{\theta}_b$.
  - average them to get a more efficient estimator:$\hat{\theta} = \frac{1}{2}(\hat{\theta}_a + \hat{\theta}_b)$.

Heterogeneous Treatment Effects

Wager and Athey (2017)

- Estimated effects may be heterogeneous across individuals.
  - These dimensions of heterogeneity may be proxied in text.
  - e.g., Republican judges might be harsher in cases where drug use occurred; Democrats might be harsher in cases where gender discrimination occured.
- I haven't seen any applications like this, but see Wager and Athey (2017) for some tools for data-driven recovery of heterogeneous effects.

Deep Instrumental Variables

Hartford, Lewis, Leyton-Brown, and Taddy (2017)

- *Deep IV: A Flexible Approach for Counterfactual Prediction*
  - use ML algorithms to extend 2SLS to high-dimensional settings
- Causal effect of interest:

$$f(w;\theta) = \mathbb{E}\{y|w\}$$

- Predictors are a function of some instruments:

$$w \sim g(w|z)$$

First stage

Hartford, Lewis, Leyton-Brown, and Taddy (2017)

- Deep IV allows arbitrarily high-dimensional w and z.
- In first stage, approximate $g(w|\gamma(z))$, the distribution of w:
  - assume that $g(\cdot)$ is a mixture density network (a mixture of gaussian distributions) where the parameter vector $\gamma(\cdot)$ includes the weights, means, and variances (Bishop 2006).
  - $\gamma(z)$ is any function of the instruments – can use an MLP, for example.
  - $g(\cdot)$ has to be a parametrized distribution because Deep IV requires that the distribution be integrated in the second stage.

Second Stage

Hartford, Lewis, Leyton-Brown, and Taddy (2017)

- In second stage, want to predict $\hat{y}(w;\theta)$, where $\hat{y}(w;\theta)$ should be a flexibly specified DNN to allow for non-linearities and interactions.
- Hartford et al (2017) show that causal estimates for θ are obtained by minimized the conditional loss function

$$\mathcal{L}(\theta) = \sum_i [y_i - \int \hat{y}(w;\theta) d\hat{g}(w|\gamma(z_i))]^2$$

  - this is true y minus predicted $\hat{y}$, but $\hat{y}$ is conditioned on the instrument-predicted treatment distribution $\hat{g}$.

Second Stage Loss Approximation

Hartford, Lewis, Leyton-Brown, and Taddy (2017)

- The integral in $\mathcal{L}(\theta)$ is approximated by

$$\int \hat{y}(w;\theta)d\hat{g}(w|\gamma(z_i)) \approx \frac{1}{m}\sum_{j}^{m}\hat{y}(\tilde{w}(z_i);\theta)$$

  where you make m draws from the estimated treatment distribution given z_i (the instruments for observation i).

  - Like 2SLS, a prediction for the endogenous regressor with the instruments is used during second-stage estimation.

What about relevance/inference?

Hartford, Lewis, Leyton-Brown, and Taddy (2017)

- Both stages of Deep IV can be validated by out-of-sample prediction in held-out data
  - in the first stage, this guards against weak-instruments bias in the same way that first-stage F-statistics thresholds do for 2SLS

The Blessings of Multiple Causes

Wang and Blei (2018)

- This paper proves an intriguing insight:
  - causal inference with multiple causes (treatments) requires weaker assumptions than classical (single-treatment) causal inference.
- In particular, unbiased causal inference is possible if confounders are shared across multiple treatments.
  - Wang and Blei (2018) provide an ML method to construct a "deconfounder" from the predictors and allow valid inference.

How does the deconfounder work?

Wang and Blei (2018)

- Assume multiple treatments $A_1, ..., A_m$
- Assume there is a latent factor Z that, when taken out from the $\vec{A}$, renders them conditionally independent.
  - If we can learn Z, this will deconfound the treatments.

Argument for Deconfounder Z

Wang and Blei (2018)

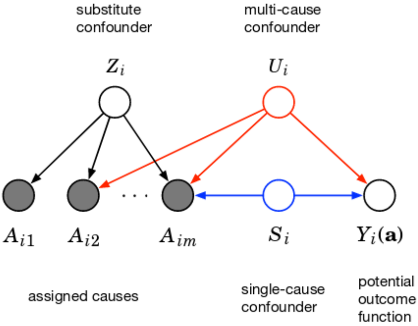


Figure 1: A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij}'s conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1},...,a_{im} \mid z_i) = \prod_j p(a_{ij} \mid z_i)$; if there exists a multi-cause confounder U_i (red) then, by d-separation, conditional independence cannot hold (Pearl, 1988). Note we cannot rule out the single-cause confounder S_i (blue).

Constructing and validating the deconfounder

Wang and Blei (2018)

- Learning the deconfounder is the same as learning any factor model:
    - can use PCA orLDA, for example, or a DNN (e.g. autoencoder)
- To check whether your deconfounder is working, check whether your factor model is capturing distribution of treatment assignment:
    - fit the factor model on training data; it should be able to predict treatment assignment in the test data.
    - the paper provides a formal test statistic.

Best Film Actors: Causal Evidence

Wang and Blei (2018)

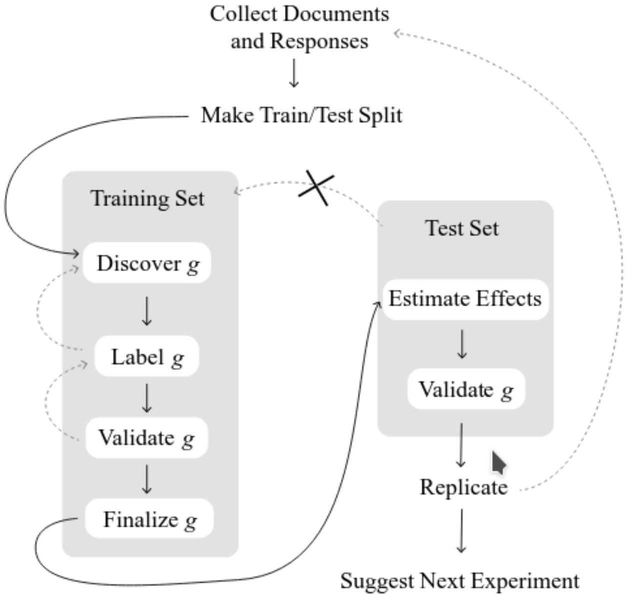
- Top revenue actors, non-causal estimates:
  - Tom Cruise, Tom Hanks, Will Smith, Arnold Schwartzenegger, Robert De Niro, Brad Pitt.
- Top revenue actors, causal estimates:
  - Owen Wilson, Nick Cage, Cate Blanchett, Antonio Banderes.
- Most under-valued actors:
  - Stanley Tucci, Willem Dafoe, Susan Sarandon, Ben Affleck, Christopher Walken.

Setup

Egami, Fong, Grimmer, Roberts, and Stewart (2018)

- There are some latent treatments in the text, represented by W_i
  - Each individual has an outcome Y_i or a non-text treatment X_i
- Text outcome, non-text treatment: $W_i = g(X_i; \theta)$
- Text treatment, non-text outcome: $Y_i = f(W_i; \theta)$

- Learn functional form for $g(\cdot)$ in half the data, and then run causal inference in the other half.



Egami, Fong, Grimmer, Roberts, and Stewart (2018)

Sample Split

Egami, Fong, Grimmer, Roberts, and Stewart (2018)

- The insight/emphasis of Egami et al (2018):
  - the *codebook function* $g(\cdot)$ can take any form (you can use any featurization approach you like)
  - you get valid inference as long as its done in held-out data.
- For example, can assume treatments are represented by frequencies over predictive N-grams, by LDA topics, or document embedding clusters.

How do voters evaluate candidates?

Fong and Grimmer (2016)

- What biographical facts affect voter evaluations?
- Could run a survey experiment:
  - `Document 1: He earned his Juris Doctor in 1997 from Yale Law School, where he operated free legal clinics for low-income residents of New Haven, Connecticut.`
  - `Document 2: He served in South Vietnam from 1970 to 1971 during the Vietnam War in the Army Rangers' 75th Ranger Regiment, attached to the 173rd Airborne Brigade. He participated in 24 helicopter assaults...`
- But hard to generalize what features drive differences.

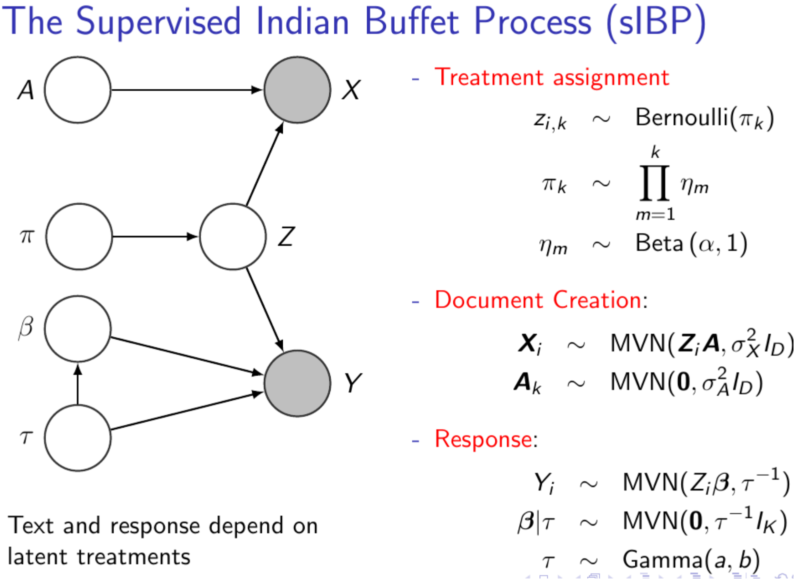
Discovery of Treatments from Text Corpora

Fong and Grimmer (2016)

1. Randomly assign texts, X_i , to respondents
2. Obtain responses Y_i for each respondent
3. Randomly divide text/responses into training and test set
    3.1 Avoid technical issues with using entire sample
    3.2 Ensure we avoid "p-hacking" (false discovery)
4. In training set: Discover mapping from texts to treatments
5. In test set: infer treatments and measure their effects

Supervised Indian Buffet Process

Fong and Grimmer (2016)



The Supervised Indian Buffet Process (sIBP)

- Treatment assignment

$$\begin{aligned} z_{i,k} &\sim \text{Bernoulli}(\pi_k) \\ \pi_k &\sim \prod_{m=1}^{k} \eta_m \\ \eta_m &\sim \text{Beta}\left(\alpha, 1\right) \end{aligned}$$

- Document Creation:

$$\begin{aligned} \boldsymbol{X}_i &\sim \text{MVN}(\boldsymbol{Z}_i \boldsymbol{A}, \sigma_X^2 I_D) \\ \boldsymbol{A}_k &\sim \text{MVN}(\boldsymbol{0}, \sigma_A^2 I_D) \end{aligned}$$

- Response:

$$\begin{aligned} Y_i &\sim \text{MVN}(Z_i \boldsymbol{\beta}, \tau^{-1}) \\ \boldsymbol{\beta}|\tau &\sim \text{MVN}(\boldsymbol{0}, \tau^{-1} I_K) \\ \tau &\sim \text{Gamma}(a, b) \end{aligned}$$

Text and response depend on latent treatments

Candidate Biographies on Wikipedia

Fong and Grimmer (2016)

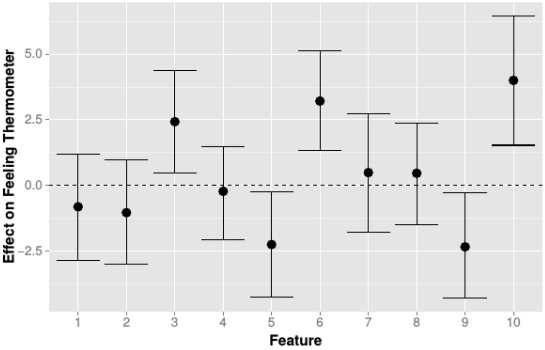
> *Schumacher was born and raised in the Highlandtown neighborhood of East Baltimore, the eldest of the three daughters of Christine Eleanor (nee Kutz) and William Schumacher. Her parents were both of Polish descent; her immigrant great-grandparents had owned a bakery in Baltimore. During her high school years at the Institute of Notre Dame, she worked in her parents' grocery store...*

- Protocol: Each respondent sees up to 3 texts from the corpus of $>$ 2200 biographies
  - Observe text
  - Feeling thermometer rating: 0-100
- 1,886 participants, 5,303 responses
  - 2,651 training, 2,652 test

Results

Fong and Grimmer (2016)

Treatment	Keywords
3	director, university, received, president, phd, policy
5	elected, house, democratic, seat
6	united_states, military, combat, rank
9	law, school_law, law_school, juris_doctor, student
10	war, enlisted, united_states, assigned, army


Effect on Feeling Thermometer
Feature