Sequencing Legal DNA

NLP for Law and Political Economy

10. Causal Inference with Text Data

Research Design

- The goal of social-science research with big data is the same as other social-science research:

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

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

Objectives

1. What is the research question?
2. Corpus and Data
3. **Research design for estimating causal parameters:**
   - What are we trying to estimate?
   - **What identification strategy / research design will get us there?**

Objectives

1. What is the research question?
2. Corpus and Data
3. **Research design for estimating causal parameters:**
   - What are we trying to estimate?
   - **What identification strategy / research design will get us there?**
4. Empirical analysis
   - **Show evidence that identification assumptions hold.**
   - **Produce causal estimates with confidence intervals.**
   - Answer the research question.

Outline

The Empirical Problem

Adjusting for Confounders without Instruments
- Adjustment for Non-Linear Confounding with Double ML
- Matching / Synthetic Control
- Adjusting for Text Confounders with BERT Embeddings
- Decounfounding with Multiple Treatments

Instrumental Variables
- Ash, and Morelli, Vannoni (2020): More Laws, More Growth?
- Galletta-Ash-Chen 2020: Causal Effect of Judicial Sentiment
- Deep IV (Hartford et al 2017)
- Deep IV for Influence of Legal Texts (Ash and Nikolaus 2020)

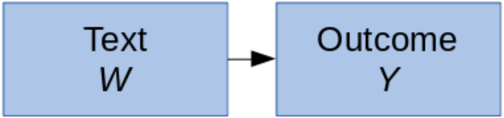
Learning Treatments from Text

Setup

- W, vectorized text

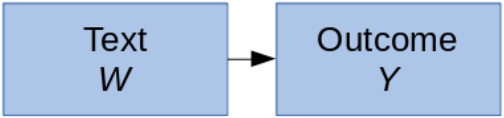
Setup

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

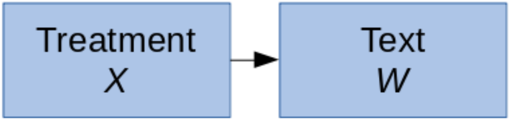


Setup

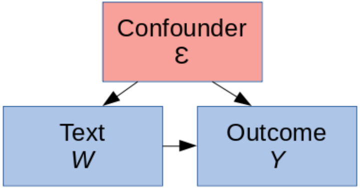
- W, vectorized text
- Y, outcome from the text
  - e.g., the facts of the case W determine the verdict Y



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

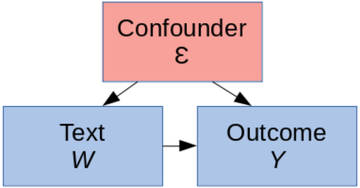


Empirical Problem: Confounders (ϵ)

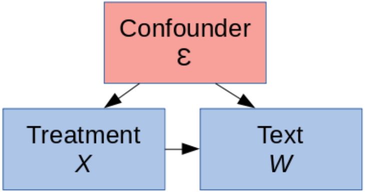


- judge decides Y based on defendant characteristics ϵ as well as case facts W

Empirical Problem: Confounders (ϵ)

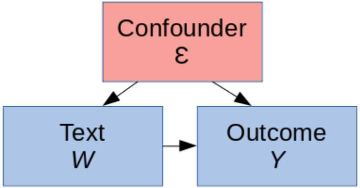


- judge decides Y based on defendant characteristics ϵ as well as case facts W

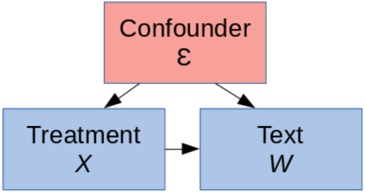


- judge writes opinion W based on characteristics ϵ as well as her ideology X.

Empirical Problem: Confounders (ϵ)

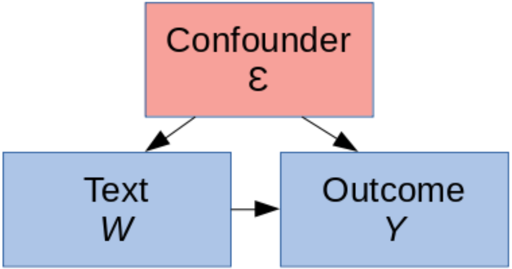


- judge decides Y based on defendant characteristics ϵ as well as case facts W



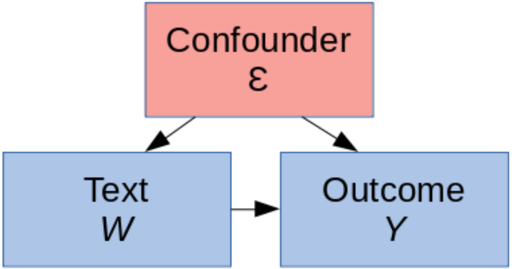
- judge writes opinion W based on characteristics ϵ as well as her ideology X.
- Key point: **a variable is a confounder only if it affects both sides of a regression** (both W & Y, or both X & W).

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.

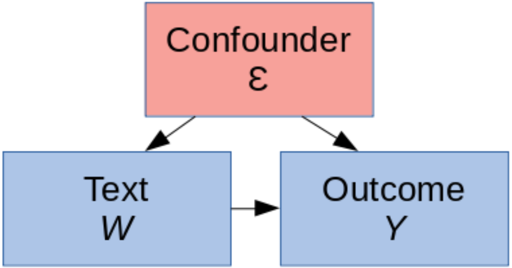
Econometrics



- 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.

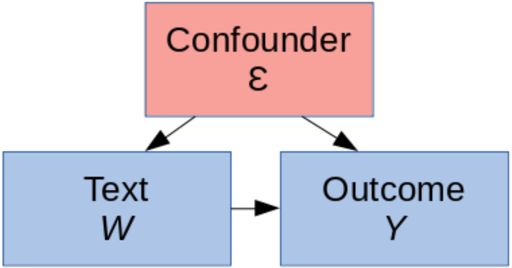
Econometrics + Machine Learning



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

- 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.

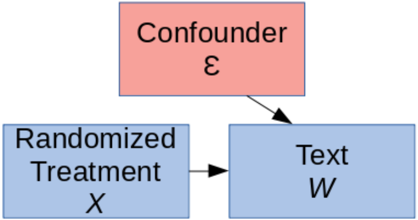
Econometrics + Machine Learning



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

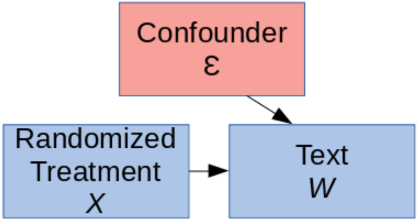
- We could take a machine learning (ML) approach and learn a nonlinear approximation $\hat{f}(W;\theta)$ to predict Y in held-out data.
  - If we obtained more documents W_i for new individual i, we could form a good prediction about the associated Y_i.
- But the ML estimates $\hat{\theta}$ do *not* 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.

Randomized Experiments



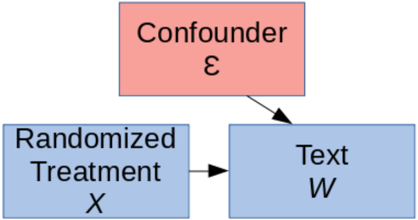
- Lab/field experiments provide a gold standard for obtaining causal estimates.
    - If treatment X is randomly assigned, it is uncorrelated with the confounder by construction ($X \perp \epsilon$).

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.

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, $\mathbb{E}\{W|X=1\} - \mathbb{E}\{W|X=2\}$.

Empirical Economics and Research Design

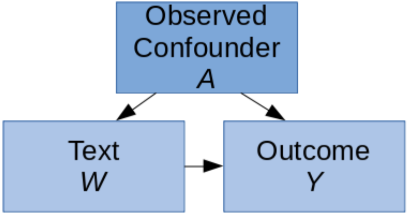
- In the presence of unobserved confounders, estimating causal parameters presents a significant challenge.
  - especially in observational studies where we can't run experiments.

Empirical Economics and Research Design

- In the presence of unobserved confounders, estimating causal parameters presents a significant challenge.
  - especially in observational studies where we can't run experiments.
- 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.

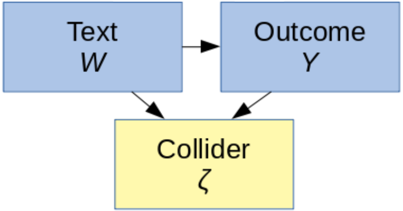
Outline

Confounder is Observed



- If confounder A is observed, problem solved:
  - include A in your model, or residualize W and Y on A before estimation.
- Often a strong assumption; ML can help if A is high-dimensional (more on this later).

Colliders or "Bad Controls"



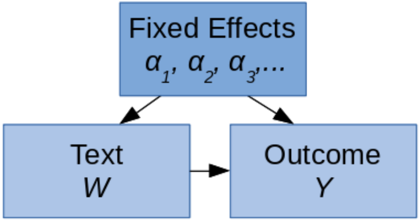
- ζ, colliders (or as most economists would say, "bad controls"), are a third variable that is affected by both your treatment and your outcome.
  - For example, let ζ be the length of the prison sentence, which is affected by the case facts W and the verdict Y.
- **Don't control for colliders**! It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

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.

Fixed Effects

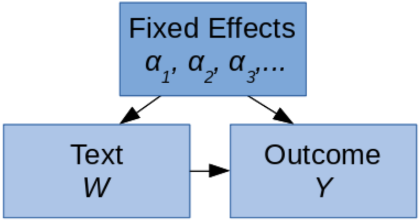
- What if all confounders are at the group level?
  - e.g., (unobserved) defendant characteristics ϵ are the only deconfounder for the verdict, and those are constant over time.
- If same defendant i is observed over multiple cases, can control/adjust for defendant characteristics by including a fixed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



  - in the data, add a dummy variable equaling one for i's cases.

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.

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.

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 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.

Outline

Double/Debiased ML

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

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.
  - this is the "observed confounders" case, but covariates are high-dimensional and non-linearly related to outcome and treatment.

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

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 $\tilde{T} = T - \hat{T}(A)$

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 $\hat{\theta}$.

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 $\hat{\theta}_a$

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:$\hat{\theta} = \frac{1}{2}(\hat{\theta}_a + \hat{\theta}_b)$.

Outline

Matching and 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).

Matching and 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: like double ML, also equivalent to fixed effects or controlling for many observed confounders.
  - but powered up with ML

Matching and 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: like double ML, also equivalent to fixed effects or controlling for many observed confounders.
  - but powered up with ML
- 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)

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

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.

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)

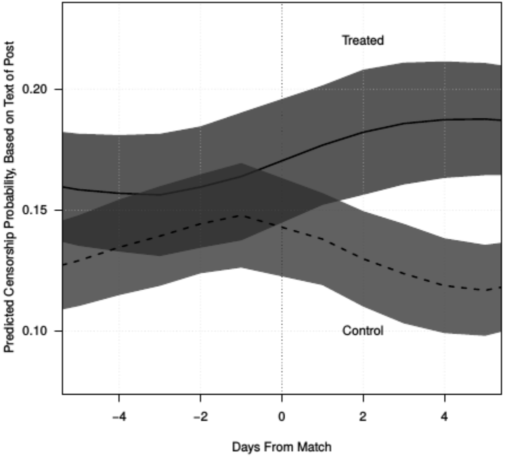
Application to online censorship in China

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

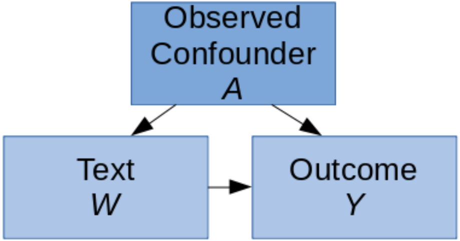
Roberts, Stewart, and Nielsen (2018)



- Bloggers who are censored respond with more censorable content.

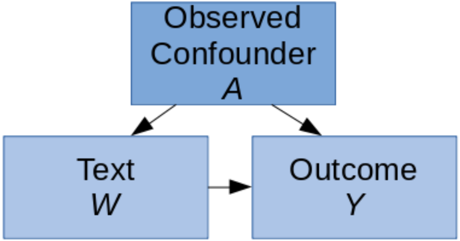
Outline

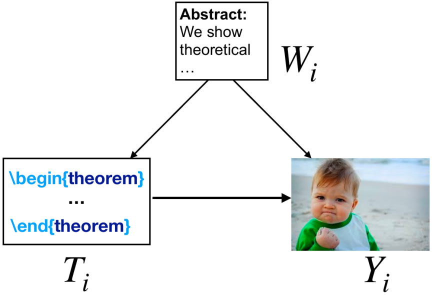
Sridhar, Veitch, and Blei (2019)



- This paper analyzes the problem of the effect of text features on outcomes, where the unobserved confounders are other features of the document.

Sridhar, Veitch, and Blei (2019)





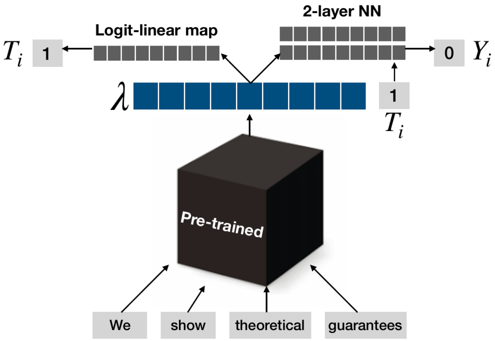
- This paper analyzes the problem of the effect of text features on outcomes, where the unobserved confounders are other features of the document.
- For example, the effect of putting a theorem in your paper on acceptance to a conference/journal.
- This paper is another example of controlling for observed confounds, but in high dimensions.

Approach

- Insight: the confounding part of the text is that which carries information about both treatment and outcome.

Approach

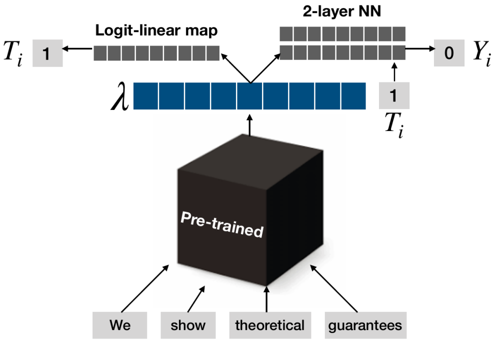
- Insight: the confounding part of the text is that which carries information about both treatment and outcome.



- Start with pre-trained BERT embeddings.
- fine-tune them on an additional multitask objective.

Approach

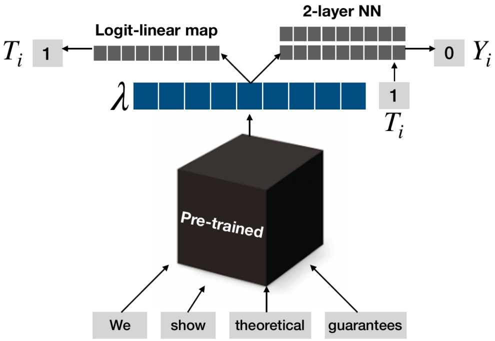
- Insight: the confounding part of the text is that which carries information about both treatment and outcome.



- Start with pre-trained BERT embeddings.
- fine-tune them on an additional multitask objective.
- 1) predict propensity score (probability of treatment given other text features.
- 2) predict outcomes conditional on treatment.

Approach

- Insight: the confounding part of the text is that which carries information about both treatment and outcome.



- Start with pre-trained BERT embeddings.
- fine-tune them on an additional multitask objective.
- 1) predict propensity score (probability of treatment given other text features.
- 2) predict outcomes conditional on treatment.
- the resulting embeddings serve as a sufficient statistic for the unobserved confounders.

Veitch et al

- The applications in the paper are not that interesting/convincing.
- Nice opportunity for an Assignment 3 replication exercise.

Outline

The Blessings of Multiple Causes

Wang and Blei (2018)

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.

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.

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 $A_1, \ldots, A_m$
- Assume there is a latent factor Z that, when taken out from the $\vec{A}$, renders them conditionally independent.

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)

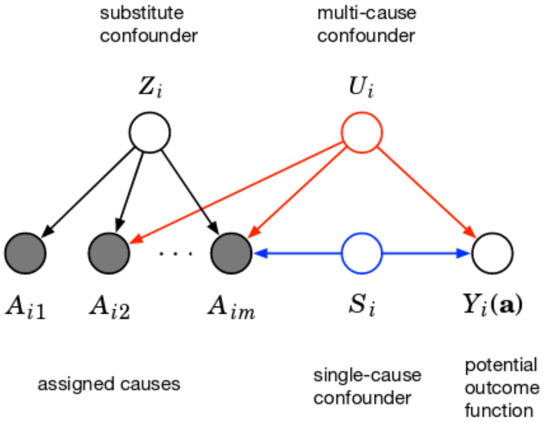


Figure 1: A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij}'s conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1}, ..., a_{im} \mid z_i) = \prod_j p(a_{ij} \mid z_i)$; 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)

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.

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.

Best Film Actors: Causal Evidence

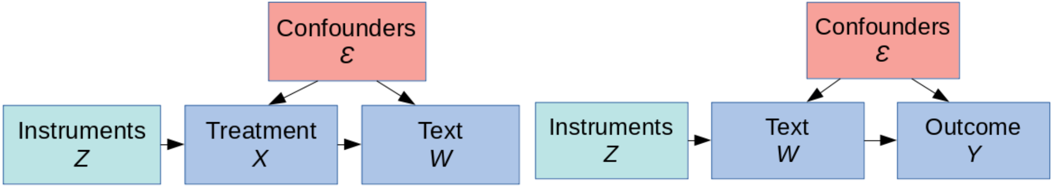
Wang and Blei (2018)

- 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.

- This paper also has a not very interesting application.
- Another good replication exercise!

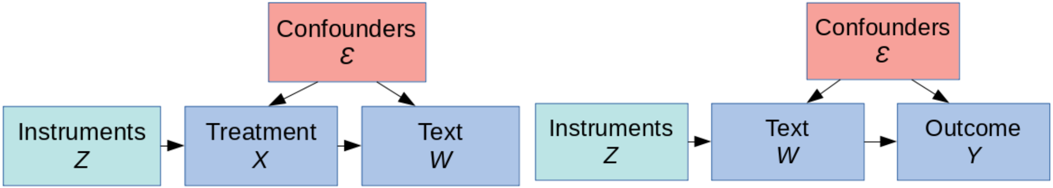
Outline

Instrumental Variables



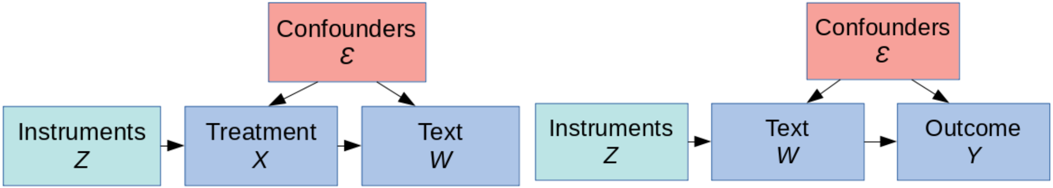
- A valid instrument Z is related to the treatment but not otherwise correlated with the outcome
  - left panel: Z affects X but orthogonal to ϵ.
  - right panel: Z affects W but orthogonal to ϵ.

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 first stage F-statistic

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 first stage F-statistic
- Second stage:
  - Predict $W(\hat{X}(Z))$ or $Y(\hat{W}(Z))$

Random Assignment of Judges $\rightarrow Z$

- 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.

Random Assignment of Judges $\rightarrow Z$

- 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 $\hat{f}(\cdot)$

Regression Discontinuity Design (RDD)

- RDD's are a special type of IV that 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.

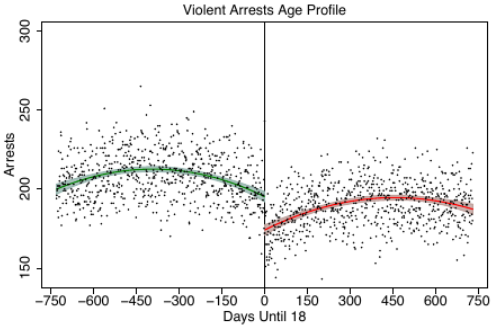
Regression Discontinuity Design (RDD)

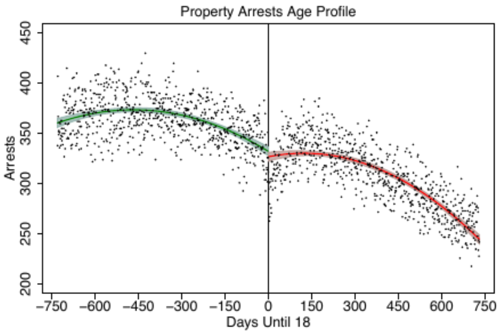
- RDD's are a special type of IV that exploit threshold rules, where individuals are assigned to treatment if a continuous variable is above some discrete cutoff.
  - The idea is to exploit randomness around this threshold.
- Example "running variables":
  - Score in entry exams, effect of barely making it into college.
  - Income, effect of barely being eligible for poverty subsidy

Regression Discontinuity Design (RDD)

- RDD's are a special type of IV that 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



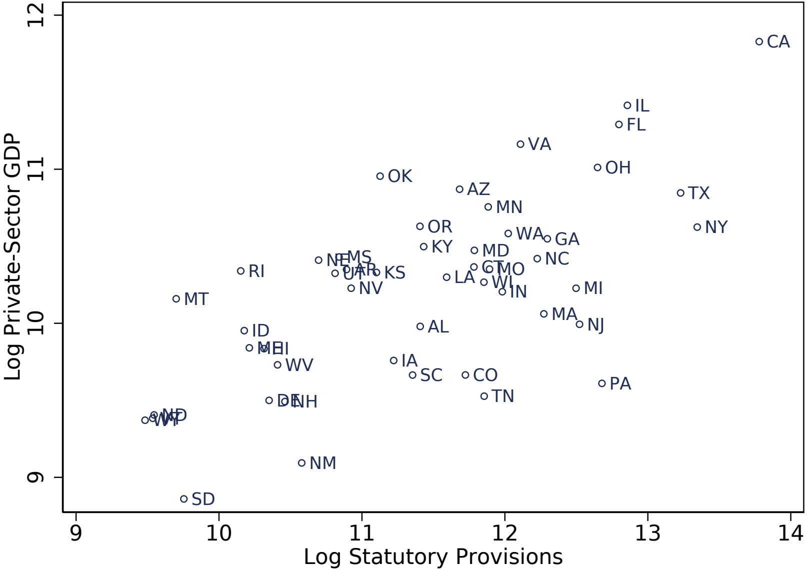


Lovett and Zue (2018). California data.

Outline

Laws $\leftrightarrow$ Growth

Ash, Morelli, and Vannoni (2020)



Second Stage

$$\Delta \log(Y_{st}) = \rho \Delta \log(W_{st}) + \alpha_{st} + X'_{st}\beta + \varepsilon_{st}$$

- Y_{st}, private-sector GDP in state s at biennium t
  - also look at alternative outcomes to understand mechanism

Second Stage

$$\Delta \log(Y_{st}) = \rho \Delta \log(W_{st}) + \alpha_{st} + X'_{st}\beta + \varepsilon_{st}$$

- Y_{st}, private-sector GDP in state s at biennium t
  - also look at alternative outcomes to understand mechanism
- W_{st}, number of legal provisions enacted in state s at biennium t
  - also use log number of words, and log provisions / words

Second Stage

$$\Delta \log(Y_{st}) = \rho \Delta \log(W_{st}) + \alpha_{st} + X_{st}'\beta + \varepsilon_{st}$$

- Y_{st}, private-sector GDP in state s at biennium t
  - also look at alternative outcomes to understand mechanism
- W_{st}, number of legal provisions enacted in state s at biennium t
  - also use log number of words, and log provisions / words
- α_{st}, state and time fixed effects, and state trends
- X_{st}, other observable factors
- ε_{st}, unobservable factors and random noise

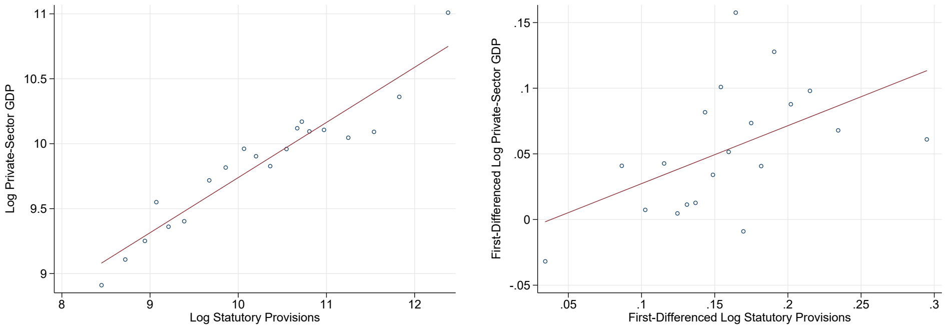
Second Stage

$$\Delta \log(Y_{st}) = \rho \Delta \log(W_{st}) + \alpha_{st} + X'_{st}\beta + \varepsilon_{st}$$

- Y_{st}, private-sector GDP in state s at biennium t
  - also look at alternative outcomes to understand mechanism
- W_{st}, number of legal provisions enacted in state s at biennium t
  - also use log number of words, and log provisions / words
- α_{st}, state and time fixed effects, and state trends
- X_{st}, other observable factors
- ε_{st}, unobservable factors and random noise
- ρ, causal effect of legislation on growth

OLS Relationship between Output Y_{st} and Detail W_{st}

$\log Y_{st} \sim \log W_{st}$ $\qquad$ $\Delta \log Y_{st} \sim \Delta \log W_{st}$



Binned scatterplots for (log) provisions (horizontal axis) and (log) private sector GDP (vertical axis), residualized on year fixed effects. Left panel: cross-sectional relationship. Right panel: first differences.

- We apply latent dirichlet allocation (LDA) to learn topics from text (Blei 2003). Here are 8 most frequent (out of 25):

- We apply latent dirichlet allocation (LDA) to learn topics from text (Blei 2003). Here are 8 most frequent (out of 25):

- Let S_{st} be the set of statutes in state s time t. Each statute $i \in S_{st}$ has a provision count w_i and a distribution over topics v_{il}, $\forall l \in \{1, ..., 25\}$.
- Define legislative flows on topic l in state s during t:

$$W_{slt} = \sum_{i \in S_{st}} v_{il} w_i$$

Shift-Share Instruments

- Classic application (Bartik 1994):
  - Instrument for local employment growth with interaction between pre-treatment local sectoral shares and national growth rates by sector.
  - Isolates changes in employment due to local demand shocks.

Shift-Share Instruments

- Classic application (Bartik 1994):
  - Instrument for local employment growth with interaction between pre-treatment local sectoral shares and national growth rates by sector.
  - Isolates changes in employment due to local demand shocks.
- Other applications:
  - Market size and drug innovation (Acemoglu and Linn, QJE 2004).
  - China shock (Autor, Dorn, and Hanson, AER 2013).
  - Democracy does cause growth (Acemoglu et al, JPE 2019).

Constructing Instrument for Legislative Detail

- Define:
  - W_{slt}, number of provisions on legislative topic $l \in 1, ..., 25\}$ in state $s \in \{1, ..., 50\}$ at biennium t.
  - W_{st}, total number of legislative provisions in state s at t.
  - $\frac{W_{sl0}}{W_{s0}}$, topic shares of legislation for first biennium of data (1963-1964).

Constructing Instrument for Legislative Detail

- Define:
  - W_{slt}, number of provisions on legislative topic $l \in 1, ..., 25\}$ in state $s \in \{1, ..., 50\}$ at biennium t.
  - W_{st}, total number of legislative provisions in state s at t.
  - $\frac{W_{sl0}}{W_{s0}}$, topic shares of legislation for first biennium of data (1963-1964).
- The instrument:

$$\underbrace{Z_{st}}_{\text{instrument}} = \underbrace{\sum_{l=1}^{25} \frac{W_{sl0}}{W_{s0}}}_{\text{shares}} \underbrace{\sum_{r \neq s} \frac{1}{49} \frac{\Delta W_{rlt}}{W_{rlt-1}}}_{\text{shifts}}$$

  - leave-one-out average proportional legislative topic growth in other states, multiplied by this state's pre-treatment topic share.
  - standardized to mean zero and variance one.

First Stage

$$\Delta \log(W_{st}) = \psi Z_{st} + \alpha_{st} + X_{st}'\beta + \eta_{st}$$

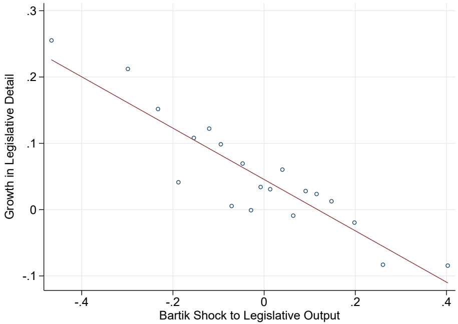
- W_{st}, legal detail in state s at biennium t
- Z_{st}, national topic detail growth weighted by s's pre-treat topic shares
- α_{st}, state/time fixed effects/trends; X_{st}, other covariates

First Stage

$$\Delta \log(W_{st}) = \psi Z_{st} + \alpha_{st} + X'_{st}\beta + \eta_{st}$$

- W_{st}, legal detail in state s at biennium t
- Z_{st}, national topic detail growth weighted by s's pre-treat topic shares
- α_{st}, state/time fixed effects/trends; X_{st}, other covariates

First-Stage Binscatter



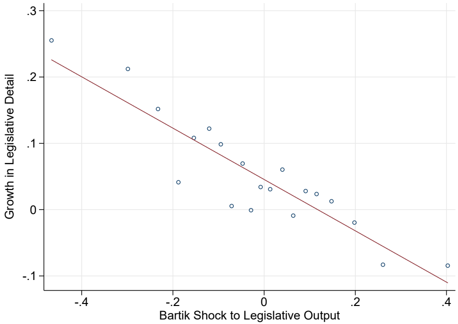
- Relevance:
  - $\hat{\psi}$ is statistically significant ($p = .003$).
  - Robust F-stat in baseline = 46.57.

First Stage

$$\Delta \log(W_{st}) = \psi Z_{st} + \alpha_{st} + X'_{st}\beta + \eta_{st}$$

- W_{st}, legal detail in state s at biennium t
- Z_{st}, national topic detail growth weighted by s's pre-treat topic shares
- α_{st}, state/time fixed effects/trends; X_{st}, other covariates

First-Stage Binscatter



- Relevance:
  - $\hat{\psi}$ is statistically significant ($p = .003$).
  - Robust F-stat in baseline $= 46.57$.
- $\hat{\psi}$ is negative:
  - different from standard Bartik.
  - when state has initially low detail on topic, it is more likely to increase detail in response to national trends.
  - e.g., state can borrow legal language at low cost.

Shift-Share Instruments: Two Approaches to Identification

Shift-Share Instruments: Two Approaches to Identification

1. Assume that **pre-treatment shares are exogenous** (Goldsmith-Pinkham, Sorkin, Swift, 2018; Jaeger, Ruist, Stuhler, 2018).

$$\mathbb{E}\{(\underbrace{\sum_{l=1}^{25} \frac{W_{sl0}}{W_{s0}}}_{\text{shares}}) \cdot \epsilon_{st}\} = 0$$

   - strong assumption: states with different initial topics (e.g. more/less employment regulation) are likely to be on different growth trends.

Shift-Share Instruments: Two Approaches to Identification

1. Assume that **pre-treatment shares are exogenous** (Goldsmith-Pinkham, Sorkin, Swift, 2018; Jaeger, Ruist, Stuhler, 2018).

$$\mathbb{E}\{(\underbrace{\sum_{l=1}^{25} \frac{W_{sl0}}{W_{s0}}}_{\text{shares}}) \cdot \epsilon_{st}\} = 0$$

   - strong assumption: states with different initial topics (e.g. more/less employment regulation) are likely to be on different growth trends.

2. Assume that **current-period shifters are exogenous** (Borusyak & Jaravel, 2017; Adao, Kolesar, Morales, QJE 2019).

$$\mathbb{E}\{(\underbrace{\sum_{r \neq s} \frac{1}{49} \frac{\Delta W_{rlt}}{W_{rlt-1}}}_{\text{shifts}}) \cdot \epsilon_{st}\} = 0$$

   - global shocks are uncorrelated with unobserved determinants of growth in legislative detail; no assumption needed on pre-treatment shares.

Shift-Share Instruments: Two Approaches to Identification

1. Assume that **pre-treatment shares are exogenous** (Goldsmith-Pinkham, Sorkin, Swift, 2018; Jaeger, Ruist, Stuhler, 2018).

$$\mathbb{E}\{(\underbrace{\sum_{l=1}^{25} \frac{W_{sl0}}{W_{s0}}}_{\text{shares}}) \cdot \epsilon_{st}\} = 0$$

   - strong assumption: states with different initial topics (e.g. more/less employment regulation) are likely to be on different growth trends.

2. Assume that **current-period shifters are exogenous** (Borusyak & Jaravel, 2017; Adao, Kolesar, Morales, QJE 2019).

$$\mathbb{E}\{(\underbrace{\sum_{r \neq s} \frac{1}{49} \frac{\Delta W_{rlt}}{W_{rlt-1}}}_{\text{shifts}}) \cdot \epsilon_{st}\} = 0$$

   - global shocks are uncorrelated with unobserved determinants of growth in legislative detail; no assumption needed on pre-treatment shares.
   - with state/time fixed effects and state trends (as in our context), global shocks are allowed to be correlated with exposure-weighted averages of unobservables that linearly vary within state (Borusyak & Jaravel, 2017).

Assessing instrument validity

- Following Borusyak & Jaravel (2017) and Adao, Kolesar, Morales (QJE 2019):
  - Instrument is driven by a majority of topics.
  - Economic growth is uncorrelated with future values of the instrument.
  - Instrument is uncorrelated with pre-treatment state characteristics.
  - Olea & Pflueger (2013) cluster-robust test for weak instruments should have $F > 23.1$; in our data $F > 43.81$.

Assessing instrument validity

- Following Borusyak & Jaravel (2017) and Adao, Kolesar, Morales (QJE 2019):
  - Instrument is driven by a majority of topics.
  - Economic growth is uncorrelated with future values of the instrument.
  - Instrument is uncorrelated with pre-treatment state characteristics.
  - Olea & Pflueger (2013) cluster-robust test for weak instruments should have $F > 23.1$; in our data $F > 43.81$.
- Following Goldsmith-Pinkham, Sorkin, Swift (2018) and Jaeger, Ruist, Stuhler (2018):
  - pre-treatment topic shares uncorrelated with pre-treatment state characteristics.
  - pre-treatment topic shares are uncorrelated with economic growth.

Effect of Legislative Detail on Economic Growth

	(1) OLS	(2) Red. Form	(3) 2SLS	(4) 2SLS	(5) 2SLS	(6) 2SLS	(7) 2SLS	(8) 2SLS	(9) 2SLS
Δ Log W_{st}	0.015*	-0.026**	0.071**	0.075**	0.066**	0.095**	0.063**	0.07**	0.078**
	(0.0062)	(0.0085)	(0.022)	(0.022)	(0.021)	(0.028)	(0.021)	(0.022)	(0.023)
Obs	846	848	846	846	812	846	846	846	796
First-Stage F			46.57	46.47	48.49	47.99	45.25	46.26	46.86
State FE	X	X	X	X	X	X	X	X	X
Time FE	X	X	X	X	X	X	X	X	X
State Trends				X					
Pre $X \times \alpha_t$					X				
Pop/Income						X			
Govt Expevd							X		
Politics								X	
Lagged DV									X

Notes: Column 1 shows the results for the OLS regression model. Column 2 shows the results for the reduced form. Column 3 gives the baseline 2SLS estimate and Column 4 adds state-specific linear trends. Column 5 adds a set of covariates (share child and old population, the fraction urban population, and the share of foreign born population) measured in the pre-treatment period interacted with biennium fixed effects. Column 6,7 and 8 add a series of time-varying covariates, respectively: population and income variables, government expenditure variables and political party control variable. Column 9 adds the lagged dependent variable. All specifications include state and biennium fixed effect; standard errors clustered by state. **p$<$.01; *p$<$.05; +p$<$.1.

- A 10% increase in legislative growth rate increases economic growth rate by ~0.7-0.9%.
  - ~4x larger than OLS, probably due to measurement error and that IV captures LATE.

- A 10% increase in legislative growth rate increases economic growth rate by ~0.7-0.9%.
  - ~4x larger than OLS, probably due to measurement error and that IV captures LATE.
- Additional robustness checks:
  - two-way clustering by state and year.
  - inclusion of topic share controls (Borusyak and Jaravel 2017).
  - use k-means clustering on initial topic vectors to group states by legislative types, then cluster standard errors on these initial-topic groups (Adao, Kolesar, Morales QJE 2019).

- A 10% increase in legislative growth rate increases economic growth rate by ~0.7-0.9%.
  - ~4x larger than OLS, probably due to measurement error and that IV captures LATE.
- Additional robustness checks:
  - two-way clustering by state and year.
  - inclusion of topic share controls (Borusyak and Jaravel 2017).
  - use k-means clustering on initial topic vectors to group states by legislative types, then cluster standard errors on these initial-topic groups (Adao, Kolesar, Morales QJE 2019).
- Other explanatory variables – strong first stage and significant 2SLS treatment effects:
  - log number of pages, log number of words
  - log provisions per page, log provisions per word

- A 10% increase in legislative growth rate increases economic growth rate by ~0.7-0.9%.
  - ~4x larger than OLS, probably due to measurement error and that IV captures LATE.
- Additional robustness checks:
  - two-way clustering by state and year.
  - inclusion of topic share controls (Borusyak and Jaravel 2017).
  - use k-means clustering on initial topic vectors to group states by legislative types, then cluster standard errors on these initial-topic groups (Adao, Kolesar, Morales QJE 2019).
- Other explanatory variables – strong first stage and significant 2SLS treatment effects:
  - log number of pages, log number of words
  - log provisions per page, log provisions per word
- Which industries are growing?
  - effect concentrated in construction, manufacturing, and wholesale

Other Economic Performance Outcomes

	(1) Log Pop	(2) Log Income	(3) Log Firms	(4) Log Profits	(5) Log Emp
Δ Log W_{st}	0.00007	0.0214*	0.00531	0.128**	0.0744*
	(0.00478)	(0.00904)	(0.0110)	(0.0429)	(0.0328)
Observations	846	846	819	548	819
First Stage F-stat	46.57	46.57	44.02	46.09	44.02
State FE	X	X	X	X	X
Time FE	X	X	X	X	X

Notes: All outcomes in logs. Column 1 uses population as dependent variables. Column 2 uses personal income. Column 3 uses the number of companies. Column 4 and Column 5 use respectively firm profits and employment. All specifications include state and biennium fixed effect, as well as standard errors clustered by state. **p$<$.01; *p$<$.05; +p$<$.1.

Outline

Identification

- we want to estimate causal effect of judge sentiments on citizen attitudes.
- but they could be correlated without indicating causation:
  - citizen attitudes could influence judges
  - or there could be a third unobserved factor.

Identification

- we want to estimate causal effect of judge sentiments on citizen attitudes.
- but they could be correlated without indicating causation:
  - citizen attitudes could influence judges
  - or there could be a third unobserved factor.
- Solution: instrumental variables using random assignment of judges, following Belloni et al (2012).
- We have access to 61 variables that refer to judges' biographical characteristics (age, geographic history, education, occupational history, governmental positions, military service, religion, race, gender, political affiliations, etc)
  - Let $\mathbf{J}_i$ = average characteristics for the three judges assigned to case i.
- Let W_i^k be the average similarity of case i to target k. Then, the vector of judge characteristics randomly assigned to target k in circuit c during year t is

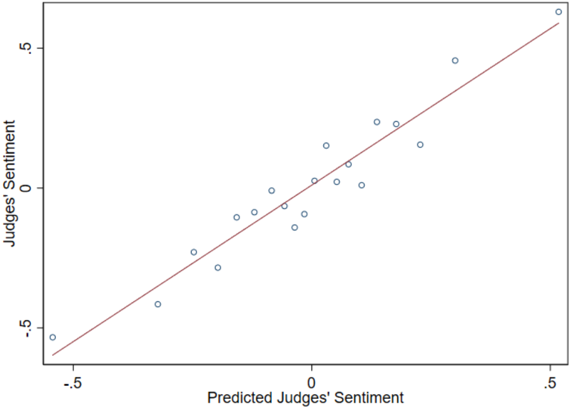
$$\mathbf{J}_{ckt} = \frac{1}{|C_{ct}|} \sum_{i \in C_{ct}} W_i^k \mathbf{J}_i, \tag{1}$$

The many weak instruments problem

- a well-known limitation of IV is that the instruments have to be sufficiently strong (that is, correlated with the treatment), or else the 2SLS estimator is biased.
- Recent papers in econometrics have used lasso to select instruments (Belloni et al 2012) or used ridge regression with cross-validated predictions (Hansen and Kozbur 2014).

The many weak instruments problem

- a well-known limitation of IV is that the instruments have to be sufficiently strong (that is, correlated with the treatment), or else the 2SLS estimator is biased.
- Recent papers in econometrics have used lasso to select instruments (Belloni et al 2012) or used ridge regression with cross-validated predictions (Hansen and Kozbur 2014).
- Z_{ckt} = the cross-validated prediction for S_{ckt} using the randomly assigned judge characteristics:



2SLS System

- The first stage is
$$S_{ckt} = \gamma_k + \gamma_{ct} + \gamma_Z Z_{ckt} + \eta_{ckt}$$
  - γ_{ck} = dummy variables (fixed effects) for each circuit-year
  - γ_k = dummy variables (fixed effects) for each target.
  - Z_{ckt} = the cross-validated prediction for S_{ckt} using the randomly assigned judge characteristics.

2SLS System

- The first stage is

$$S_{ckt} = \gamma_k + \gamma_{ct} + \gamma_Z Z_{ckt} + \eta_{ckt}$$

  - γ_{ck} = dummy variables (fixed effects) for each circuit-year
  - γ_k = dummy variables (fixed effects) for each target.
  - Z_{ckt} = the cross-validated prediction for S_{ckt} using the randomly assigned judge characteristics.
- The second stage is

$$Y_{ckt} = \alpha_k + \alpha_{ct} + \beta S_{ckt} + \epsilon_{ckt} \tag{2}$$

- Y_{ckt} = thermometer response from ANES in circuit c toward taret k at t.

Results

Table 1: Results

	OLS			2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
Judges' sentiment	-0.138***	-0.137***	-0.135***	-0.139***	-0.167***	-0.122**
	(0.017)	(0.017)	(0.017)	(0.052)	(0.051)	(0.058)
Year FE	Y	Y	Y	Y	Y	Y
Circuit FE	Y	Y	Y	Y	Y	Y
Year FE X Circuit FE	N	Y	Y	N	Y	Y
Target FE	N	N	Y	N	N	Y
F-stat				127.286	124.573	101.201
N observations	2678	2678	2678	2678	2678	2678

Notes: The dependent variable is the thermometer score for all respondents in the ANES by circuit-target-year. *Judges' sentiment* is the text-based average sentiment by circuit-target-year. All variables are centered and standardized by target. Standard errors clustered by circuit-year in parenthesis. * $p < 0.1$, ** $p < 0.05$ and *** $p < 0.01$.

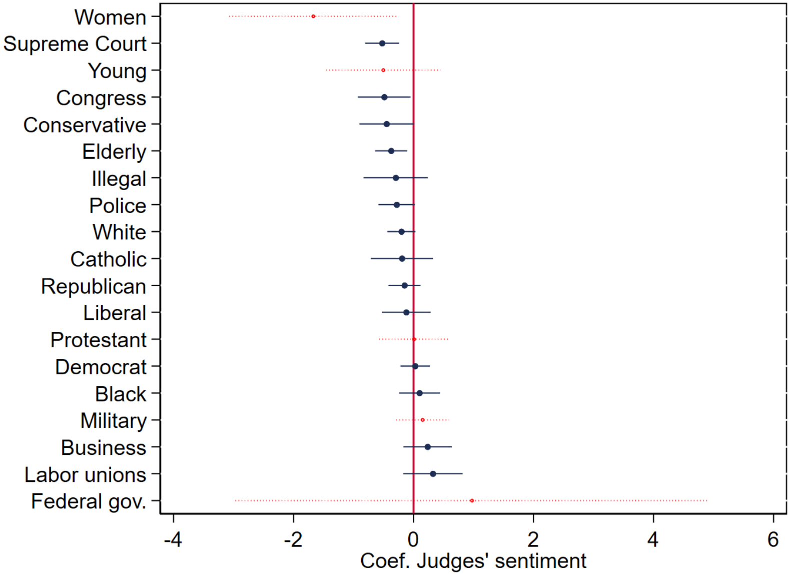
Results

Table 2: Results – Leads/Lags of Dep. variable

	2SLS				
	2 years before	Same year	2 years after	4 years after	6 years after
	(1)	(2)	(3)	(4)	(5)
Judges' sentiment	-0.113	-0.122**	-0.215**	-0.094	-0.051
	(0.083)	(0.058)	(0.085)	(0.092)	(0.189)
Year FE	Y	Y	Y	Y	Y
Circuit FE	Y	Y	Y	Y	Y
Year FE X Circuit FE	Y	Y	Y	Y	Y
Target FE	Y	Y	Y	Y	Y
F-stat	65.546	101.201	77.585	46.285	29.677
N observations	1687	2678	1684	1322	1004

Notes: The dependent variables are the leads and lags of the thermometer score for all respondents in the ANES by circuit-target-year as reported in columns head. *Judges' sentiment* is the text-based average sentiment by circuit-target-year. All variables are centered and standardized by target. Standard errors clustered by circuit-year in parenthesis. * $p < 0.1$, ** $p < 0.05$ and *** $p < 0.01$.

Effect by group



Outline

Deep Instrumental Variables

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

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

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).

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.

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.

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

Outline

Dataset Overview ($n = 71'475$ cases, 382 judges)

Outcome $Y =$ log citations by future judges.
Treatments $X =$ features of written majority opinions:

1. NLP preprocessing of all documents using spaCy
2. Training of a word embedding model using word2vec
3. Clustering of word embeddings into 200 clusters that are interpreted as topics and arguments of the cases
4. Computation of vector of normalized "cluster frequencies" (counts of words from each cluster in a case)

Dataset Overview ($n = 71'475$ cases, 382 judges)

Outcome Y = log citations by future judges.
Treatments X = features of written majority opinions:

1. NLP preprocessing of all documents using spaCy
2. Training of a word embedding model using word2vec
3. Clustering of word embeddings into 200 clusters that are interpreted as topics and arguments of the cases
4. Computation of vector of normalized "cluster frequencies" (counts of words from each cluster in a case)

Instruments Z = Judge writing style

- The average of the cluster frequencies of all other cases that the judges on the case panel were assigned to.

Cluster Frequencies

Table: (Hypothetical) Example of Cluster Frequencies

Cluster ID	**Words**	**Freq. Case 1**	**...**	**Freq. Case 71475**
1	word2, word17, word83, ...	0.00001	...	0.00000
2	word5, word89, word1005, ...	0.00000	...	0.00020
...	...	...	...	...
199	word19, word33, word100, ...	0.00023	...	0.00190
200	word14, word16, word64, ...	0.05010	...	0.00000

- The cluster frequencies of a case sum to 1

Leveraging Random Assignment

- Empirical strategy relies on random assignment of circuit judges, which occurs within the set of judges working on a single court during a year
  - To isolate this endogenous variation, we center Y, X, and Z by court-year
- randomization has been verified in previous papers, but we still need to check that it holds in our context.

Estimation, Prediction, and Evaluation

- We trained the neural nets using Keras/TensorFlow with an 85%/15% train/test split, batch size $= 64$, and early stopping.
  - Tuned hyperparameters: layers, L2 regularization, dropout, learning rate

Estimation, Prediction, and Evaluation

- We trained the neural nets using Keras/TensorFlow with an 85%/15% train/test split, batch size $= 64$, and early stopping.
  - Tuned hyperparameters: layers, L2 regularization, dropout, learning rate
- Take the best model's causal estimate $\hat{\theta}_2$ and form predictions $\hat{Y}_i = f(X_i; \hat{\theta}_2)$ for each case.

Estimation, Prediction, and Evaluation

- We trained the neural nets using Keras/TensorFlow with an 85%/15% train/test split, batch size $= 64$, and early stopping.
  - Tuned hyperparameters: layers, L2 regularization, dropout, learning rate
- Take the best model's causal estimate $\hat{\theta}_2$ and form predictions $\hat{Y}_i = f(X_i; \hat{\theta}_2)$ for each case.
- For comparison we also trained a non-causal "Deep OLS" model:
  - neural network analogous to our second stage
  - predicts Y based on non-instrumented X

Estimation, Prediction, and Evaluation

- We trained the neural nets using Keras/TensorFlow with an 85%/15% train/test split, batch size $= 64$, and early stopping.
  - Tuned hyperparameters: layers, L2 regularization, dropout, learning rate
- Take the best model's causal estimate $\hat{\theta}_2$ and form predictions $\hat{Y}_i = f(X_i; \hat{\theta}_2)$ for each case.
- For comparison we also trained a non-causal "Deep OLS" model:
  - neural network analogous to our second stage
  - predicts Y based on non-instrumented X
- To analyze features, we use a permutation importance approach to identify text features that move $\hat{Y}$ the most when scrambled.

Using Causal Predictions to Analyse the Quality of Judges

- Previous work analyzed the quality of judges using citations directly
- Deep IV provides a means to analyse judge quality based on text-based citation predictions

Court	Judge	Text Quality Rank	Predicted Cites (Deep IV)
5th	Stewart	1	0.3133
5th	Demoss	2	0.2656
5th	Wiener	3	0.2403
5th	Benavides	4	0.1910
5th	Clement	5	0.1817
...	...	...	...
3rd	Hutchinson	27	0.0861
7th	Posner	28	0.0831
4th	Chapman	29	0.0784
2nd	Harlan	30	0.0773
7th	Easterbrook	31	0.0763
...	...	...	...
...	...	382	...

Causal Importance of Text Features

- We analyzed the permutation importance of the treatment features (word clusters / topics):
  - higher impact of procedural language on common-law
  - lower impact of case-specific language (e.g. fraud, corporate accounting, debt)
  - some legal concepts turn out not to matter for legal impact (forensic evidence, fact finder / jury)
  - sanity check: "junk" topics and topics containing typo words rank at the bottom of the impact list

Top 4 Cluster Words	**Topic**	**Deep IV Importance**	**Deep OLS Importance**	**Rank (D-IV)**	**Rank (D-OLS)**
grant, denying, denial, adjudged	grant/deny	0.0882	0.1631	1	1
adjourn, adjournment, reschedule, continuance	schedule	0.0481	0.0001	3	125
certioari, rehearing, rehear, cert	certioari	0.0365	0.0191	6	20
...	...	...	...	...	...
corrupt, defraud, bogus, fraudulent	fraud	0.0002	0.0005	181	145
Cir1991, cir1985, Cir1996, Cir1987	junk	0.0001	0.0022	194	73
finder, trier, factfinder, inference	jury	0.0000	0.0002	199	178

Direction of Treatment Effects

- Use a bootstrap approach to sample from $\hat{X} \sim F(x | \gamma(c, z; \theta_1))$
- Fit OLS using sampled treatment ($Y \sim \hat{X}$):

Top 4 Cluster Words	Deep IV Effect	Deep IV Rank	Deep OLS Rank
complicate, depend, crucial, illustrate	0.091	1	3
implausible, problematic, exaggeration, skeptical	0.059	2	105
reverse, affirm, vacate, reversed	0.043	3	192
...	...	...	...
argument, contention, assertion, suggestion	-0.045	195	199
reconsider, reconsideration, remand, modify	-0.060	199	185
4th, 9th, see, 8th (amendments)	-0.070	200	170

  - nuanced legal reasoning $\rightarrow$ increased citations
  - procedural aspects $\rightarrow$ decreased citations

Outline

Setup

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

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

Setup

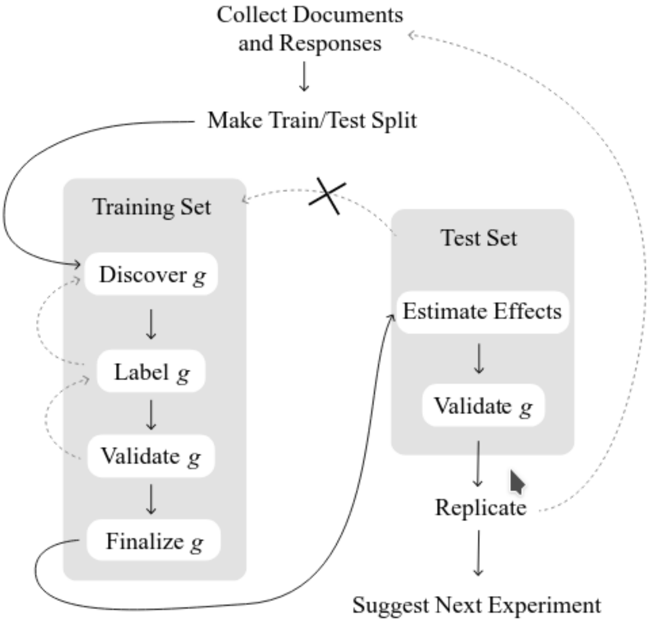
Egami, Fong, Grimmer, Roberts, and Stewart (2018)

- 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: $Y_i = f(W_i; \theta)$

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.

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)

How do voters evaluate candidates?

Fong and Grimmer (2016)

- What biographical facts affect voter evaluations?

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...`

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

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

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)

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

Discovery of Treatments from Text Corpora

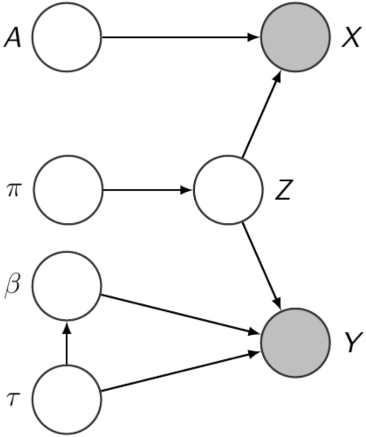
Fong and Grimmer (2016)

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)



Text and response depend on latent treatments

- Treatment assignment

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

- 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:

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

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

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

Candidate Biographies on Wikipedia

Fong and Grimmer (2016)

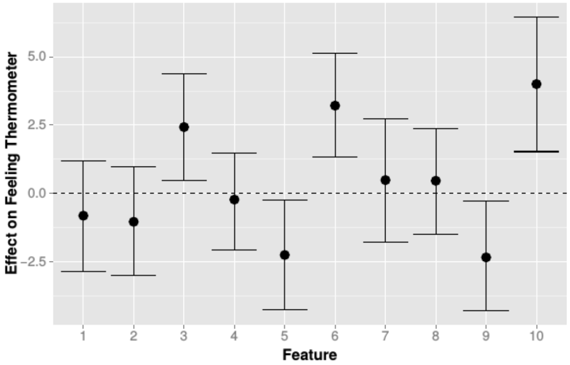
Schumacher was born and raised in the Highlandtown neighborhood of East Baltimore, the eldest of the three daughters of Christine Eleanor (nee Kutz) and William Schumacher. Her parents were both of Polish descent; her immigrant great-grandparents had owned a bakery in Baltimore. During her high school 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



Learning Heterogeneous Treatment Effects

Wager and Athey (2017)

Learning Heterogeneous Treatment Effects

Wager and Athey (2017)

- Estimated effects may be heterogeneous across individuals.
  - These dimensions of heterogeneity may be proxied in text.

Learning 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.

Learning 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.
  - another good replication exercise.