Research Design with Text Data

Elliott Ash

Max Planck Summer School 2017

Research Design

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

Research Design

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

Research Design

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

Causal Inference Meets Text Analysis

- Computer scientists tend to explore exercises based on prediction
- We need less theory for predictive cases as we can simply hold-out data and evaluate accuracy. In my experience prediction turns on questions about the sample and population.
- Very little work in the causal/text space- generally speaking it is pursued by two separate communities.
  - causal inference generally takes the measurement strategy as given
  - text analysis usually does not focus on the downstream inference
- With some areas of text analysis this is relatively straightforward.
- Things get tricky when we start talking about unsupervised methods.

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Causal Inference Basics

- Suppose we have N units, $i = 1, ..., N$.
- Each unit receives a treatment T_i
- We also observe a response $Y_i(T_i)$
- Fundamental problem of causal inference:
  - only observe one potential outcome $Y_i(T_i)$
- With random assignment to treatment $T \in 0, 1$, the treatment effect is just the average difference between the treated and untreated groups.
- Other empirical strategies are designed to simulate this type of random variation

Three Roles for Text in the Causal Pipeline

- Text as outcome: assess a text-based response
  - e.g. text-predicted ideology changes according to senate election schedule (Ash, Morelli, and Van Weeldden 2017)
- Text as treatment: assess the effect of a text
  - discover treatments (Fong and Grimmer 2016)
- Text as confounder: condition on text
  - matching with text (Roberts, Stewart and Nielsen 2017)

Three Roles for Text in the Causal Pipeline

- Text as outcome: assess a text-based response
  - e.g. text-predicted ideology changes according to senate election schedule (Ash, Morelli, and Van Weeldden 2017)
- Text as treatment: assess the effect of a text
  - discover treatments (Fong and Grimmer 2016)
- Text as confounder: condition on text
  - matching with text (Roberts, Stewart and Nielsen 2017)

Three Roles for Text in the Causal Pipeline

- Text as outcome: assess a text-based response
  - e.g. text-predicted ideology changes according to senate election schedule (Ash, Morelli, and Van Weeldden 2017)
- Text as treatment: assess the effect of a text
  - discover treatments (Fong and Grimmer 2016)
- Text as confounder: condition on text
  - matching with text (Roberts, Stewart and Nielsen 2017)

Outline

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Empirical Strategies

- The standards for good empirical work in text projects can and should be the same as those in applied microeconomics:
  - e.g., Angrist and Pischke, *Mostly Harmless Econometrics*
- Classic methods, that still work:
  - lab experiments
  - field experiments
  - differences-in-differences
  - regression discontinuity
  - instrumental variables
  - matching / high-dimensional controls

Lab/Field Experiments

- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - ask your subjects to fill out open-ended survey responses before and after the experiment
  - subjects can talk to each other in a chatroom
  - subjects view randomly assigned text treatments
    - Fong and Grimmer 2015, to be discussed in detail later

Lab/Field Experiments

- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - ask your subjects to fill out open-ended survey responses before and after the experiment
  - subjects can talk to each other in a chatroom
  - subjects view randomly assigned text treatments
    - Fong and Grimmer 2015, to be discussed in detail later

Lab/Field Experiments

- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - ask your subjects to fill out open-ended survey responses before and after the experiment
  - subjects can talk to each other in a chatroom
  - subjects view randomly assigned text treatments
    - Fong and Grimmer 2015, to be discussed in detail later

Lab/Field Experiments

- Lab/field experiments provide a gold standard for obtaining causal estimates.
  - ask your subjects to fill out open-ended survey responses before and after the experiment
  - subjects can talk to each other in a chatroom
  - subjects view randomly assigned text treatments
    - Fong and Grimmer 2015, to be discussed in detail later

Differences in differences

- Measure changes in an outcome due to changes in text using panel data.
  - a lot of strong assumptions to do this, though
- Or measure changes in a text-based metric due to a treatment
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016), effect of political control on tax base
- Include fixed effects and trends for the individuals.
  - Try to illustrate effect with event-study graph

Differences in differences

- Measure changes in an outcome due to changes in text using panel data.
  - a lot of strong assumptions to do this, though
- Or measure changes in a text-based metric due to a treatment
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016), effect of political control on tax base
- Include fixed effects and trends for the individuals.
  - Try to illustrate effect with event-study graph

Differences in differences

- Measure changes in an outcome due to changes in text using panel data.
  - a lot of strong assumptions to do this, though
- Or measure changes in a text-based metric due to a treatment
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016), effect of political control on tax base
- Include fixed effects and trends for the individuals.
  - Try to illustrate effect with event-study graph

Differences in differences

- Measure changes in an outcome due to changes in text using panel data.
  - a lot of strong assumptions to do this, though
- Or measure changes in a text-based metric due to a treatment
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016), effect of political control on tax base
- Include fixed effects and trends for the individuals.
  - Try to illustrate effect with event-study graph

Regression Discontinuity

- Look at local impact of a thresholded treatment on a text-based metric.
  - Electoral RD effect on type of speech used by Congressmen, for example.
- Illustrate result with conditional mean diagrams.

Regression Discontinuity

- Look at local impact of a thresholded treatment on a text-based metric.
  - Electoral RD effect on type of speech used by Congressmen, for example.
- Illustrate result with conditional mean diagrams.

Instrumental Variables

- Use exogenous variation from instruments:
  - Relevance: $F > 10$ in first stage
  - Exclusion restriction: instrument only affects outcome through endogenous regressor channel.
- Text-based outcome:
  - Ash, Morelli, and Van Weelden (2017): news coverage instrument from Snyder and Stromberg (2010)
- See Ash (2016) for review of recent econometric results for high-dimensional two-stage-least-squares.
  - And an application using diffusion of text features through legal communities.

Instrumental Variables

- Use exogenous variation from instruments:
  - Relevance: $F > 10$ in first stage
  - Exclusion restriction: instrument only affects outcome through endogenous regressor channel.
- Text-based outcome:
  - Ash, Morelli, and Van Weelden (2017): news coverage instrument from Snyder and Stromberg (2010)
- See Ash (2016) for review of recent econometric results for high-dimensional two-stage-least-squares.
  - And an application using diffusion of text features through legal communities.

Instrumental Variables

- Use exogenous variation from instruments:
  - Relevance: $F > 10$ in first stage
  - Exclusion restriction: instrument only affects outcome through endogenous regressor channel.
- Text-based outcome:
  - Ash, Morelli, and Van Weelden (2017): news coverage instrument from Snyder and Stromberg (2010)
- See Ash (2016) for review of recent econometric results for high-dimensional two-stage-least-squares.
  - And an application using diffusion of text features through legal communities.

Instrumental Variables

- Use exogenous variation from instruments:
  - Relevance: $F > 10$ in first stage
  - Exclusion restriction: instrument only affects outcome through endogenous regressor channel.
- Text-based outcome:
  - Ash, Morelli, and Van Weelden (2017): news coverage instrument from Snyder and Stromberg (2010)
- See Ash (2016) for review of recent econometric results for high-dimensional two-stage-least-squares.
  - And an application using diffusion of text features through legal communities.

Matching / high-dimensional controls

- "matching" is the use of covariates to weight other observations as better controls.
- You can imagine the text documents associated with individuals as a set of covariates.
  - e.g., compare impacts of state-level policing-policy changes on crime rates, while matching on (controlling for) a high-dimensional represtenation of the rest of a state's laws.
  - Roberts, Stewart, and Nielsen (2016), to be discussed later

Matching / high-dimensional controls

- "matching" is the use of covariates to weight other observations as better controls.
- You can imagine the text documents associated with individuals as a set of covariates.
  - e.g., compare impacts of state-level policing-policy changes on crime rates, while matching on (controlling for) a high-dimensional represtenation of the rest of a state's laws.
  - Roberts, Stewart, and Nielsen (2016), to be discussed later

Matching / high-dimensional controls

- "matching" is the use of covariates to weight other observations as better controls.
- You can imagine the text documents associated with individuals as a set of covariates.
  - e.g., compare impacts of state-level policing-policy changes on crime rates, while matching on (controlling for) a high-dimensional represtenation of the rest of a state's laws.
  - Roberts, Stewart, and Nielsen (2016), to be discussed later

Matching / high-dimensional controls

- "matching" is the use of covariates to weight other observations as better controls.
- You can imagine the text documents associated with individuals as a set of covariates.
  - e.g., compare impacts of state-level policing-policy changes on crime rates, while matching on (controlling for) a high-dimensional represtenation of the rest of a state's laws.
  - Roberts, Stewart, and Nielsen (2016), to be discussed later

Random Assignment of Judges

- In the USA and many other jurisdictions, judges are randomly assigned to the cases they work on.
- This is a great source of experimental variation when studying laws and the legal system.
  - Belloni et al (2012): eminent domain improves housing values but increases inequality
  - Ash and Chen (2017a): religious freedoms strengthen minority religions but weaken mainstream religions, increase religiosity overall
  - Ash and Chen (2017b): pro-environmental-groups cases reduce pollution

Random Assignment of Judges

- In the USA and many other jurisdictions, judges are randomly assigned to the cases they work on.
- This is a great source of experimental variation when studying laws and the legal system.
  - Belloni et al (2012): eminent domain improves housing values but increases inequality
  - Ash and Chen (2017a): religious freedoms strengthen minority religions but weaken mainstream religions, increase religiosity overall
  - Ash and Chen (2017b): pro-environmental-groups cases reduce pollution

Random Assignment of Judges

- In the USA and many other jurisdictions, judges are randomly assigned to the cases they work on.
- This is a great source of experimental variation when studying laws and the legal system.
  - Belloni et al (2012): eminent domain improves housing values but increases inequality
  - Ash and Chen (2017a): religious freedoms strengthen minority religions but weaken mainstream religions, increase religiosity overall
  - Ash and Chen (2017b): pro-environmental-groups cases reduce pollution

Random Assignment of Judges

- In the USA and many other jurisdictions, judges are randomly assigned to the cases they work on.
- This is a great source of experimental variation when studying laws and the legal system.
  - Belloni et al (2012): eminent domain improves housing values but increases inequality
  - Ash and Chen (2017a): religious freedoms strengthen minority religions but weaken mainstream religions, increase religiosity overall
  - Ash and Chen (2017b): pro-environmental-groups cases reduce pollution

Random Assignment of Judges

- In the USA and many other jurisdictions, judges are randomly assigned to the cases they work on.
- This is a great source of experimental variation when studying laws and the legal system.
  - Belloni et al (2012): eminent domain improves housing values but increases inequality
  - Ash and Chen (2017a): religious freedoms strengthen minority religions but weaken mainstream religions, increase religiosity overall
  - Ash and Chen (2017b): pro-environmental-groups cases reduce pollution

Outline

Wager and Athey (2017)

- *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*

In this paper, we develop a **non-parametric causal forest for estimating heterogeneous treatment effects** ... We show that **causal forests are pointwise consistent for the true treatment effect**, and have an asymptotically Gaussian and centered sampling distribution. . . In experiments, **we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially in the presence of irrelevant covariates**.

Chernozhukov et al (2017)

- *Double/Debiased Machine Learning for Treatment and Causal Parameters*

"The resulting method thus could be called a "double ML" method because it relies on estimating primary and auxiliary predictive models. In order to avoid overfitting, our construction also makes use of the K-fold sample splitting, which we call cross-fitting. **This allows us to use a very broad set of ML predictive methods in solving the auxiliary and main prediction problems**, such as random forest, lasso, ridge, deep neural nets, boosted trees, as well as various hybrids and aggregators of these methods."

Hartford et al (2017)

- *Counterfactual Prediction with Deep Instrumental Variables Networks*

This paper provides a **recipe for combining ML algorithms to solve for causal effects in the presence of instrumental variables -- sources of treatment randomization that are conditionally independent from the response**. We show that a flexible IV specification resolves into two prediction tasks that can be solved with deep neural nets: a first-stage network for treatment prediction and a second-stage network whose loss function involves integration over the conditional treatment distribution.

Outline

1. Overview of Relevant Research Designs
2. High-Dimensional Econometrics
3. Applications
   - Text as Treatment: Fong and Grimmer (2015)
   - Roberts, Stewart, and Nielsen (2016)
   - Text IV Using Legislative Diffusion
   - Robot Judges?

Outline

How do voters evaluate candidates?

- 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.

How do voters evaluate candidates?

- 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.

How do voters evaluate candidates?

- 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

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
    1. Avoid technical issues with using entire sample
    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

Discovery of Treatments from Text Corpora

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
    1. Avoid technical issues with using entire sample
    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

Discovery of Treatments from Text Corpora

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
    1. Avoid technical issues with using entire sample
    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

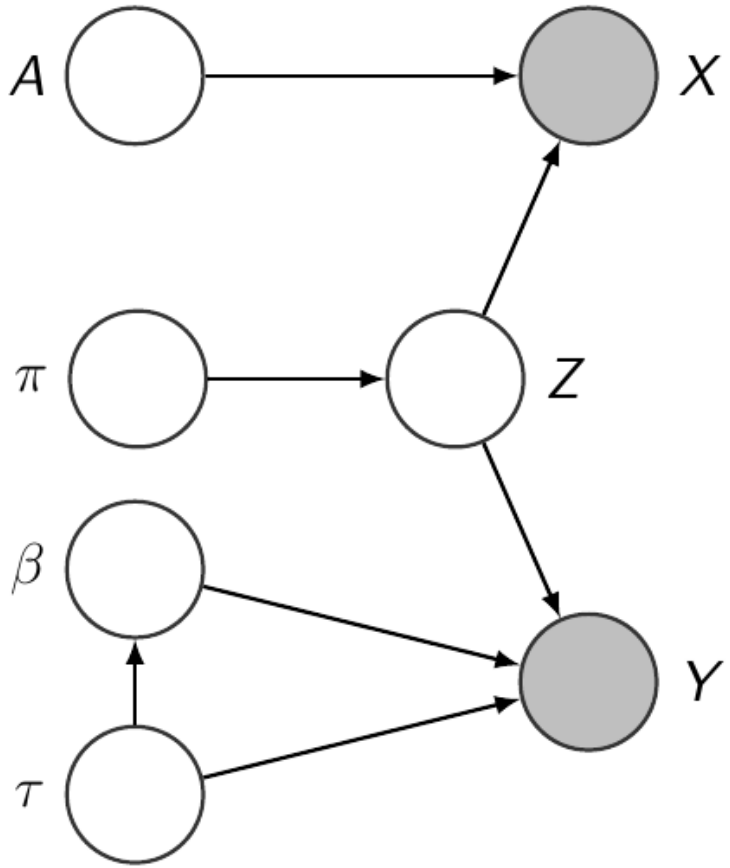
Discovery of Treatments from Text Corpora

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
    1. Avoid technical issues with using entire sample
    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

Discovery of Treatments from Text Corpora

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
    1. Avoid technical issues with using entire sample
    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

The Supervised Indian Buffet Process (sIBP)



Text and response depend on latent treatments

- Treatment assignment

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

- Document Creation:

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

- Response:

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

Candidate Biopgrahies on Wikipedia

`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

Candidate Biopgrahies on Wikipedia

`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

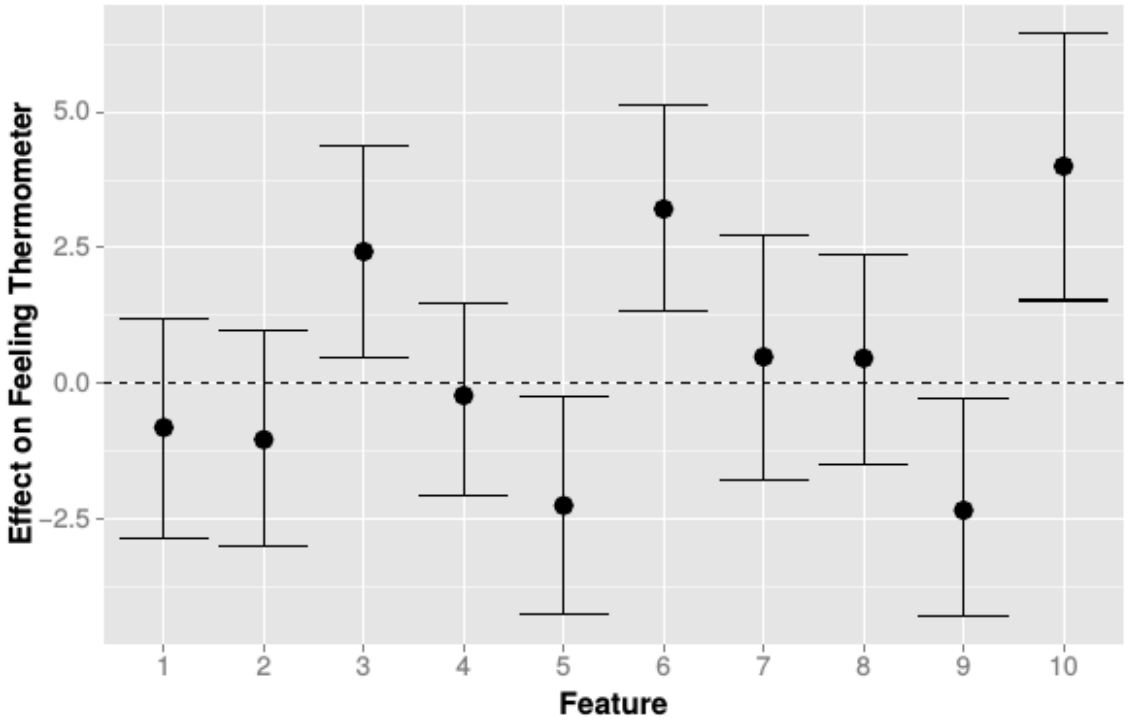
Candidate Biopgrahies on Wikipedia

`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

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



Outline

How do people react to online censorship?

- 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

Text Matching

- Roberts et al (2016) 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.

Text Matching

- Roberts et al (2016) 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.

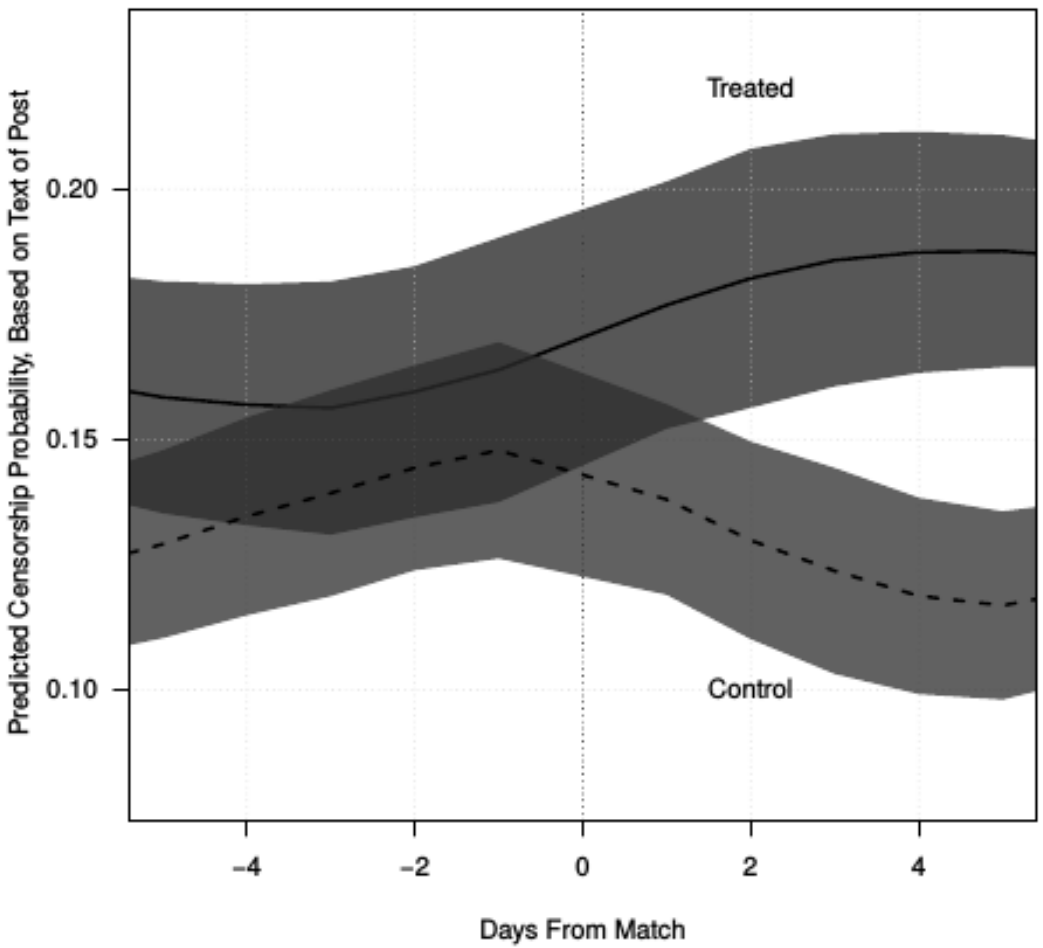
Text Matching

- Roberts et al (2016) 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.

Text Matching

- Roberts et al (2016) 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.

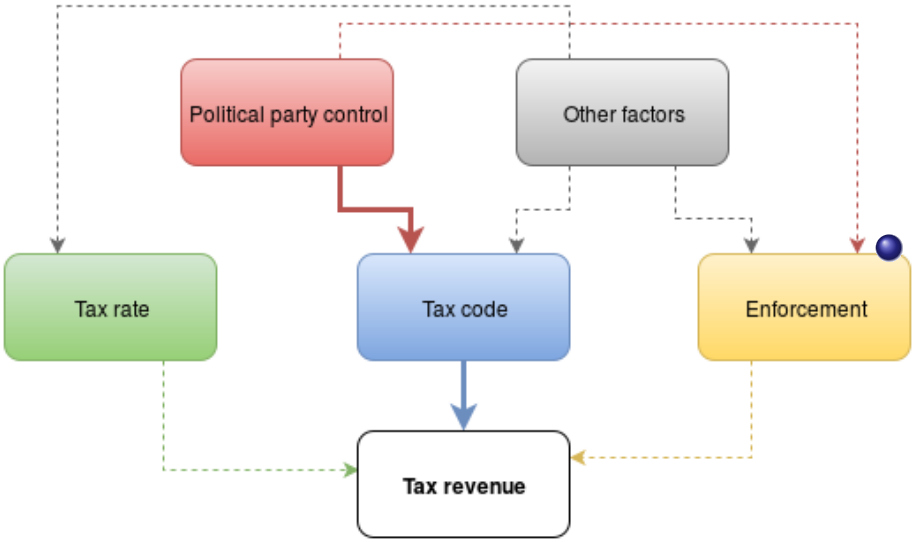
Censorship has a backlash effect



- Bloggers who are censored respond with more censorable content.

Outline

Instrumental Variables Overview



$$g_{st} = \mathbf{x}'_{st}\beta + \varepsilon_{st}$$

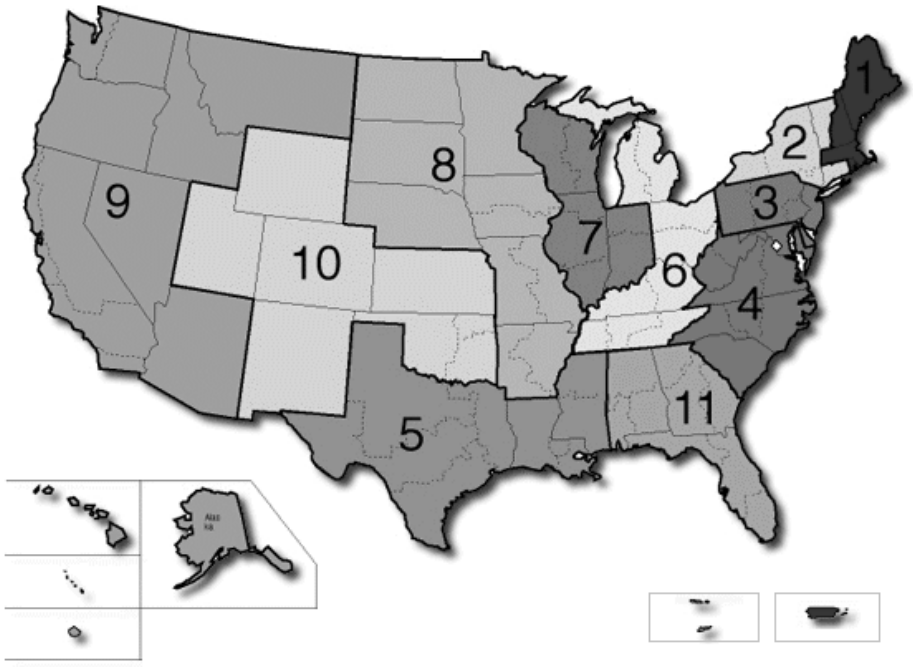
- The goal of instrumental variables is to find a set of instruments $\mathbf{z}_{st}$ that are correlated with the explanatory variables $\mathbf{x}_{st}$ but not with the error term ε_{st}.
- If $\text{Cov}(\mathbf{z}_{st}, \mathbf{x}_{st}) \neq 0$ and $\text{Cov}(\mathbf{z}_{st}, \varepsilon_{st}) = 0$, then

$$\hat{\beta}_{2SLS} = (\mathbf{z}'_{st}\mathbf{x}_{st})^{-1}\mathbf{z}'_{st}g_{st}$$

is a consistent estimator for β.

Instrumental Variables Approach

- Instrument for tax code using language diffusion from other states in same federal appellate judicial circuits:



- Based on Bartik (1991); see also Nakamura and Steinsson (AER 2013), Bertrand et al (QJE 2015).
  - Closest paper methodologically: Acemoglu, Naidu, Restrepo, and Robinson (JPE 2016): Instrument for democratization using diffusion of democratization from nearby countries.

Diffusion of text within circuits

- Federal circuits assigned long ago, primarily for federal law:
  - Circuits form legal communities (Carp, 1972), language diffuses to other states through legal rather than political/economic channels.
- Previous work on this channel:
  - Preferential policy diffusion within circuit rather than just neighboring states (Bird and Smythe, 2008).
  - Actual statute text diffuses to states in same circuit more than neighboring states in other circuits (Hinkle, 2015).

Definition of Instrument

- For each state s, time t, and phrase i, construct *leave-one-out average* phrase frequency for *other* states in same circuit, $J(s)$, for *previous* legislative session, $t-1$:

$$z_{st}^{i} = \frac{1}{|J(s)|} \sum_{j \in J(s)} x_{jt-1}^{i}$$

  - x_{jt}^{i} has been residualized on state fixed effects, party-year fixed effects (in robustness check, also on census-region-year fixed effects).

- First stage for each phrase i:

$$x_{st}^{i} = \mathbf{z}_{st}' \gamma_i + \eta_{st}^{i}, \forall i$$

  - $\gamma_i =$ ith row of $25,000 \times 25,000$ matrix of first-stage coefficients Γ.

- Second stage:

$$g_{st} = \hat{\mathbf{x}}_{st}' \beta + \varepsilon_{st}$$

Definition of Instrument

- For each state s, time t, and phrase i, construct *leave-one-out average* phrase frequency for *other* states in same circuit, $J(s)$, for *previous* legislative session, $t-1$:

$$z_{st}^{i} = \frac{1}{|J(s)|} \sum_{j \in J(s)} x_{jt-1}^{i}$$

  - x_{jt}^{i} has been residualized on state fixed effects, party-year fixed effects (in robustness check, also on census-region-year fixed effects).
- First stage for each phrase i:

$$x_{st}^{i} = \mathbf{z}_{st}' \gamma_i + \eta_{st}^{i}, \forall i$$

  - γ_i= ith row of $25,000 \times 25,000$ matrix of first-stage coefficients Γ.
- Second stage:

$$g_{st} = \hat{\mathbf{x}}_{st}' \beta + \varepsilon_{st}$$

Exclusion Restriction

- Sufficient condition for consistent 2SLS estimates for instruments z_{st}:

$$\text{Cov}(z_{st}^{i}, \varepsilon_{st}) = 0, \forall i$$

  - Lagged tax code changes in other circuit states only affect revenue through changing the tax code.
  - That is, language diffusion is primarily legal rather than political/economic.
- Circuit-specific economic/political shocks?
  - Instruments are residualized on party-control fixed effects.
  - Including covariates for current/lagged own GDP, circuit GDP, expenditures on tax enforcement, etc. doesn't change results.
- Correlated effects of circuit judicial rulings?
  - Own-phrase first-stage effect tends to be strong, consistent with diffusion rather than judicial response.
  - More work needed on this (I'm merging data on text of circuit rulings)

Exclusion Restriction

- Sufficient condition for consistent 2SLS estimates for instruments z_{st}:

$$\text{Cov}(z_{st}^{i}, \varepsilon_{st}) = 0, \forall i$$

  - Lagged tax code changes in other circuit states only affect revenue through changing the tax code.
  - That is, language diffusion is primarily legal rather than political/economic.
- Circuit-specific economic/political shocks?
  - Instruments are residualized on party-control fixed effects.
  - Including covariates for current/lagged own GDP, circuit GDP, expenditures on tax enforcement, etc. doesn't change results.
- Correlated effects of circuit judicial rulings?
  - Own-phrase first-stage effect tends to be strong, consistent with diffusion rather than judicial response.
  - More work needed on this (I'm merging data on text of circuit rulings)

Exclusion Restriction

- Sufficient condition for consistent 2SLS estimates for instruments z_{st}:

$$\text{Cov}(z_{st}^{i}, \varepsilon_{st}) = 0, \forall i$$

  - Lagged tax code changes in other circuit states only affect revenue through changing the tax code.
  - That is, language diffusion is primarily legal rather than political/economic.
- Circuit-specific economic/political shocks?
  - Instruments are residualized on party-control fixed effects.
  - Including covariates for current/lagged own GDP, circuit GDP, expenditures on tax enforcement, etc. doesn't change results.
- Correlated effects of circuit judicial rulings?
  - Own-phrase first-stage effect tends to be strong, consistent with diffusion rather than judicial response.
  - More work needed on this (I'm merging data on text of circuit rulings)

High Dimensional IV Estimation

- Standard 2SLS estimator is consistent only for small numbers of instruments relative to sample size (Chao and Swanson, 2005; Hansen et al, 2008).
  - Here, $P = 25{,}000$ and $N = 1{,}500$
- Active literature in econometrics has applied regularization methods to improve performance of 2SLS with many instruments (Caner, 2009; Gautier and Tsybakov, 2011; Okui, 2011; Carrasco, 2012; Belloni et al., 2012; Lin et al., 2015).
- Here we use elastic net, where key assumption is sparsity:
  - Most instruments have zero effect in a given first stage; a small subset of instruments can approximate effect of all instruments.
  - Adding penalty parameters excludes weak instruments.

High Dimensional IV Estimation

- Standard 2SLS estimator is consistent only for small numbers of instruments relative to sample size (Chao and Swanson, 2005; Hansen et al, 2008).
  - Here, $P = 25{,}000$ and $N = 1{,}500$
- Active literature in econometrics has applied regularization methods to improve performance of 2SLS with many instruments (Caner, 2009; Gautier and Tsybakov, 2011; Okui, 2011; Carrasco, 2012; Belloni et al., 2012; Lin et al., 2015).
- Here we use elastic net, where key assumption is sparsity:
  - Most instruments have zero effect in a given first stage; a small subset of instruments can approximate effect of all instruments.
  - Adding penalty parameters excludes weak instruments.

Outline

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!

Predicting Judicial Outcomes with Text

- Work in progress:
  - Predict circuit court appeals decisions (at 80% accuracy in U.S. Supreme Court)
  - Predict criminal sentences, and sentencing disparities, with text.
  - Look at implicit bias in judges using Word Embedding Association Test
  - Examine whether facts or law drive judicial decisions
  - Plenty more to do here I think...
- Thanks!