Causal Inference Workshop

Week 11 - Intro to text analysis and wrap-up!

Causal Inference Workshop

April 19, 2024

Anna Papp, ap3907@columbia.edu - SDEV 9280

Workshop outline

A. Causal inference fundamentals
   - Modeling assumptions matter too
   - Conceptual framework (potential outcomes framework)

B. Design stage: common identification strategies
   - IV + RDD [coding]
   - DiD, DiDiD, Event Studies, New TWFE Lit [coding]
   - Synthetic Control / Synthetic DiD [coding]

C. Analysis stage: strengthening inferences
   - Limitations of identification strategies, pre-estimation steps
   - Estimation [controls] and post-estimation steps [supporting assumptions]

D. Other topics in causal inference and sustainable development
   - Inference (randomization inference, bootstrapping, etc.)
   - Fixed effects, climate regressions, remote sensing data
   - Intro to text analysis and wrap up!

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

Presenting your paper

Text analysis: A brief introduction[1]

- Why text?
  - Vast amount of text is produced: every minute there are 2.4 million searches on Google, 156 million emails, 350,000+ tweets
  - A lot of otherwise difficult to obtain information is encoded in text

- Text is inherently very **high-dimensional**
  - Oxford English dictionary includes 171,476 words, which means that there are $\sim$15 billion possible pairs of words (ignoring ordering)

- Approaches of text analysis
  - *Close reading*: Reading text to classify or extract info using specific domain knowledge
  - *Distant reading*: Statistical models and ML to process large amounts of text w/ Natural Language Processing (NLP) using little domain knowledge but lots of computation

1. Based on *Text in Politics* lecture from Political Economy.

Text analysis: Basic steps

See Gentzkow et al. 2019 (JEL) for a great intro to text as data in economics research

- Analysis can usually be summarized in three steps:
  1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
  2. Map $\mathbf{C}$ to predicted values of $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$
  3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

- Of course, often transforming actual text to raw text already requires a lot of work

Text analysis: Basic steps

See Gentzkow et al. 2019 (JEL) for a great intro to text as data in economics research

- Analysis can usually be summarized in three steps:
  1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
     → Reduce the dimensionality of the data to a manageable level; usually counts of **tokens**
  2. Map $\mathbf{C}$ to predicted values of $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$
  3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

- Of course, often transforming actual text to raw text already requires a lot of work

Text analysis: Basic steps

See Gentzkow et al. 2019 (JEL) for a great intro to text as data in economics research

- Analysis can usually be summarized in three steps:
  1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
  2. Map $\mathbf{C}$ to predicted values of $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$
     → Apply high-dimensional statistical methods; e.g. predicted spam filter or sentiment
  3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

- Of course, often transforming actual text to raw text already requires a lot of work

Text analysis: Basic steps

See Gentzkow et al. 2019 (JEL) for a great intro to text as data in economics research

- Analysis can usually be summarized in three steps:
  1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
  2. Map $\mathbf{C}$ to predicted values of $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$
  3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis
     → Often prediction is the goal; in social science often want to infer causal relationships

- Of course, often transforming actual text to raw text already requires a lot of work

Text analysis: Basic steps

See Gentzkow et al. 2019 (JEL) for a great intro to text as data in economics research

- Analysis can usually be summarized in three steps:
  1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
  2. Map $\mathbf{C}$ to predicted values of $\hat{\mathbf{V}}$ of unknown outcomes $\mathbf{V}$
  3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

- Of course, often transforming actual text to raw text already requires a lot of work

- Examples of social science studies using $\hat{\mathbf{V}}$ in causal analyses or structural models:
  - Gentzkow and Shapiro 2010: Congressional and news text → political slant → supply and demand forces that determine slant in equilibrium
  - Engelberg and Parsons 2011: Local news coverage of earnings → Relationship between coverage & trading by local investors → Causal effect of news on stock prices
  - Du 2023: Twitter data → Sentiment of tweets → Air pollution and offensive/racist tweets

Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming

1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)

Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming[2]

raw text

> We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.

1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)

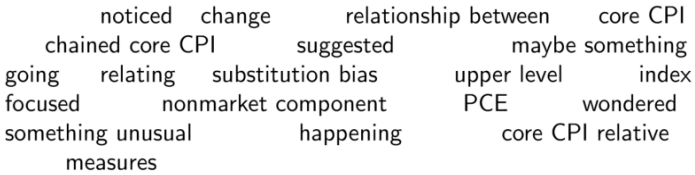
2. Illustration from Stephen Hansen, Imperial College

Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming[2]

raw text → remove stop words



1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)

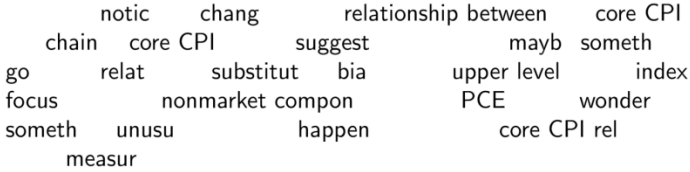
2. Illustration from Stephen Hansen, Imperial College

Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming[2]

raw text → remove stop words → stemming



1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)

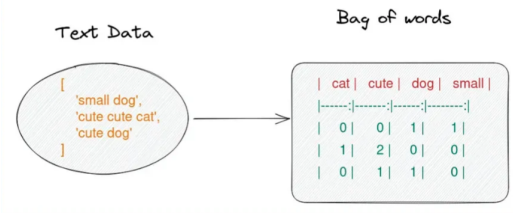
2. Illustration from Stephen Hansen, Imperial College

Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming

1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)



Text analysis: Representing raw text as numerical array

- Text represented as an ordered list of words is too complex to analyze → want to reduce the dimensionality of the data
  - Represent text as a numerical array appropriate for statistical analysis

1a. **Restricting words**, including filtering (very common words, etc.) and stemming

1b. **Mapping of text to C**, tokenizing, e.g., bag-of-words model (but many different word embedding models out there)

→ See Part 1 of `00_nlp_example.ipynb`

Text analysis: Mapping **C** to predicted values of $\hat{\mathbf{V}}$

- Map numerical array representation of text to something we care about → some sort of statistical / machine learning analysis; many different approaches
- **Topic Models** (LDA, STM, DTM, ProdLDA)
- **Text Classification** (e.g., sentiment analysis)
- **Many Other Tasks**

Text analysis: Mapping **C** to predicted values of $\hat{\mathbf{V}}$

- Map numerical array representation of text to something we care about → some sort of statistical / machine learning analysis; many different approaches

- **Topic Models** (LDA, STM, DTM, ProdLDA)
  - Unsupervised statistical method for discovering abstract 'topics' that exist within a collection of documents

  → See part 2 of `00_nlp_example.ipynb`

- **Text Classification** (e.g., sentiment analysis)

- **Many Other Tasks**

Text analysis: Mapping **C** to predicted values of $\hat{\mathbf{V}}$

- Map numerical array representation of text to something we care about → some sort of statistical / machine learning analysis; many different approaches

- **Topic Models** (LDA, STM, DTM, ProdLDA)

- **Text Classification** (e.g., sentiment analysis)
  - Classifying text into various categories or metrics
  - For example, determining the emotional value of a given expression in natural language
  - *Rule-based* sentiment analysis: Basic approach that doesn't use machine learning (lexical features labeled as positive/negative, count # of times positive/negative words appear)

  → See part 3 of `00_nlp_example.ipynb`

- **Many Other Tasks**

Text analysis: Mapping **C** to predicted values of $\hat{\mathbf{V}}$

- **Many Other Tasks**

Text analysis: Mapping **C** to predicted values of $\hat{\mathbf{V}}$

- **Many Other Tasks** , including language modeling

Language modeling: Extremely brief intro[2]

- Task in language modeling is based on predicting the next word given the context
  - From NYT story:
    Stock plunged this...
    Stock plunged this morning, despite a cut in interest rate by the Federal Reserve.

- Commonly used in speech recognition, machine translation, spelling correction, and even at the core of LLMs like ChatGPT

2. Thanks to Junho Choi for the slides this is based on!

Language modeling: Extremely brief intro[2]

- How can we generate the next word, given current context?
  - Rely on notion of the probability of sequence of words → choose highest probability one
  - $P(w_n|w_1, w_2, ..., w_{n-1})$ hard to estimate because the longer the sequence, less likely that it appears in the training data
    - Make Markov Assumption (independence assumption) that seeing a word w_n only depends on the previous $k-1$ words $P(w_n|w_1, w_2, ..., w_{n-1}) = P(w_n|w_{n-k+1}, ..., w_{n-1})$

2. Thanks to Junho Choi for the slides this is based on!

Language modeling: Extremely brief intro[2]

- How can we generate the next word, given current context?
  - Rely on notion of the probability of sequence of words → choose highest probability one
  - $P(w_n|w_1, w_2, ..., w_{n-1})$ hard to estimate because the longer the sequence, less likely that it appears in the training data
    - Make Markov Assumption (independence assumption) that seeing a word w_n only depends on the previous $k-1$ words $P(w_n|w_1, w_2, ..., w_{n-1}) = P(w_n|w_{n-k+1}, ..., w_{n-1})$

- Special cases of Markov Assumption
  - **Bi-gram** language model: current word only depends on the previous word
  - **Tri-gram** language model: current word depends only on last two words

2. Thanks to Junho Choi for the slides this is based on!

Language modeling: Extremely brief intro[2]

- How can we generate the next word, given current context?
  - Rely on notion of the probability of sequence of words → choose highest probability one
  - $P(w_n|w_1, w_2, ..., w_{n-1})$ hard to estimate because the longer the sequence, less likely that it appears in the training data
    - Make Markov Assumption (independence assumption) that seeing a word w_n only depends on the previous $k-1$ words $P(w_n|w_1, w_2, ..., w_{n-1}) = P(w_n|w_{n-k+1}, ..., w_{n-1})$

- Special cases of Markov Assumption
  - **Bi-gram** language model: current word only depends on the previous word
  - **Tri-gram** language model: current word depends only on last two words

- Where do these probabilities come from?
  - Some **corpus**: large, structured collection of texts used for linguistic research or NLP

2. Thanks to Junho Choi for the slides this is based on!

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

Presenting your paper

Causal inference using observational data

1. **Framework:** A counterfactual approach to causality
   - Potential outcomes framework *week 2*
   - Causal graph (DAG) framework *week 2*

2. **Design stage**
   - IV *week 3* [coding, week 4]
   - RD *week 3* [coding, week 4]
   - DiD, DiDiD, TWFE *week 5* [coding, week 6]
   - SCM *week 7* [coding, week 7]

3. **Analysis stage**
   - Pre-estimation: restructuring data *week 8*
   - Estimation: regression controls, inference *week 8 & 9* [coding, week 9]
   - Post-estimation: Checking/supporting assumptions *week 8*

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

1. Framework
2. Design stage
3. Analysis stage

Presenting your paper

What am I trying to estimate?

- We want to estimate a TE, defined by potential outcomes
- We can only measure observed outcomes (fundamental problem of causal inference)

→ We estimate a TE using relationship between observed outcomes and potential outcomes

individual treatment effects (TEs)	$Y_i^1 - Y_i^0 \; \forall i$	*ideally estimate;* unknowable
average treatment effect (ATE)	$\mathbb{E}[Y_i^1 - Y_i^0]$	*reasonably estimate;* unknowable, but can be estimated
average treatment effect on the treated (ATT)	$\mathbb{E}[Y_i^1 - Y_i^0 \mid D_i = 1]$	*reasonably estimate;* unknowable, but can be estimated
difference in average observed outcomes	$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$	*what we can estimate*

Independence assumptions identify TEs and give unbiased estimators

- We can only compute the difference in average *obesrved* outcomes
- We can recover an **unbiased** estimator of a causal effect iff an **identifying/independence assumption** holds:
  - if IA holds $((Y_i^0, Y_i^1) \perp\!\!\!\perp D_i) \rightarrow$ estimate ATT
  - if ~~IA~~, but CIA $((Y_i^0, Y_i^1) \perp\!\!\!\perp D_i | X_i) \rightarrow$ can estimate ATT in each stratum (and then combine)
  - if ~~CIA~~, need relevant exogenous source of variation in D_i (e.g., $(Y_i^0, Y_i^1) \perp\!\!\!\perp Z_i$; $Z_i \perp\!\!\!\perp D_i) \rightarrow$ estimate a LATE
- Estimate of $\hat{\beta}$ is never the true effect, only one realization of an unbiased *distribution*

Expressing TE as linear regression

$$Y_i = Y_i^0 + (Y_i^1 - Y_1^0)D_i = \ldots = \alpha + \beta D_i + u_i$$

Which means that:

- $\hat{\beta}_{OLS}$ is unbiased for the ATT iff:
  - there is no selection bias (identification problem; independence)
  - e is uncorrelated with D (regression problem, endogeneity)

In observational studies:

- Excluding a confounder creates bias, so we must adjust for all confounders
- With all confounders adjusted for, we have an unbiased estimator of an average TE
- As we can rarely be certain to have measured all confounders, turn to **identification strategies** that rely on other assumptions

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

1. Framework
2. Design stage
3. Analysis stage

Presenting your paper

Common Identification Methods

1. **RCT** or 'natural' randomization of treatment
2. **IV, RD**: *If there may be selection based on unobservables, we use an instrument or discontinuity that induces exogenous variation in treatment status.*
3. **DiD, DiDiD, Event Studies, SCM**: *If we have repeated observations and want to estimate the effects of an event, we use research designs that assume or construct parallel trends and only time-invariant confounds.*
   - Event-studies: All units get treated; we identify effects from *within* variation only
   - DiD: Some units never get treated; we identify effects from *within* and *between* variation
   - SCM: Combine control units to construct synthetic unit (counterfactual)

→ We reviewed the assumed DGP, identifying assumptions, the estimated, the estimator used, and best practices for each!

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

1. Framework
2. Design stage
3. Analysis stage

Presenting your paper

All Other Things (Besides Identification!)

1. **Pre-estimation**: Restructuring data, so we rely on functional form of $f(X)$ less
2. **Estimation**: What controls do we include?
   - *Required/forbidden controls*: adjust for all confounding paths, but do not adjust for post-treatment variables that may be affected by treatment
   - *Optional good/bad controls*: can adjust for pre-treatment covariates that are strong determinants of Y (may increase efficiency of $\hat{\beta}$); do not adjust for determinants of D (may decrease precision)
3. **Post-estimation**: "*Falsification tests set you up for failure: do them.*" - Doug
   - *Falsification tests:* Show that specification doesn't find an effect when one should not exist
   - *Diagnosis tests of modeling assumptions*: Easy to test basic assumptions of CLRM (non-linearity, homoskedasticity of errors, leverage of observations)

Outline

Workshop outline

Introduction to text analysis

Recap of the Causal Inference Workshop

Presenting your paper

Characterizing the empirical strategy

Your presentation/paper should contain – to some degree, explicitly:

1. **Research question**

   What causal effect of interest are we trying to estimate?
2. **Ideal experiment**

   What ideal experiment would capture the causal effect?
3. **Identification strategy**

   How are the observational data used to make comparisons that approximate such an experiment?
4. **Estimation method** *(including assumptions made when constructing standard errors)*
5. **Falsification tests** *that support the identifying assumptions.*

Putting the paper in perspective

In addition to the paper's empirical strategy, one may want to discuss:

- **Contributions to the literature**
- **Methodological contributions**
- **Internal validity** of the statistical analysis
  - *Are the identifying assumptions plausible? Could there be measurement error? Are there unexplained results?*
- **External validity** of the statistical analysis
  - *w.r.t. to policy:* Is there a gap between policy questions and the analyses performed?
  - *w.r.t. to the literature:* How does the paper account for its results compared to other results in the literature?
  - *w.r.t. to other settings:* Are the results generalizable to other populations and settings?

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
  - Topic and research question (should get to it relatively quickly; general vs. specific research question)
  - Answer to question (findings)
  - Relationship/contribution to literature
  - Importance - why should we listen to you?
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
  - Topic and research question (should get to it relatively quickly; general vs. specific research question)
  - Answer to question (findings)
  - Relationship/contribution to literature
  - Importance - why should we listen to you?
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
  - Everything you do
  - Nice to have agenda slides to keep audience on track (but don't need to describe it)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
  - Topic and research question (should get to it relatively quickly; general vs. specific research question)
  - Answer to question (findings)
  - Relationship/contribution to literature
  - Importance - why should we listen to you?
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
  - Everything you do
  - Nice to have agenda slides to keep audience on track (but don't need to describe it)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)
  - Takeaways you want audience to remember

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)

- Dealing with questions
  - Practice polite but firm things to say for questions and to control your own timing

Structuring presentation

- Introduction [forecast]: $\sim$10 minutes (5-7 for 30-min talk)
- Middle [do]: $\sim$30 minutes (12-15 min for 30-min talk)
- End [takeaways]: $\sim$5-10 minutes ($<$5 min for 30-min talk)

- Dealing with questions
  - Practice polite but firm things to say for questions and to control your own timing

- Delivery
  - Practice talking about research
  - Practice out loud

Questions? Comments?

Thank you!

References

Heavily based on Claire Palandri's 2022 version of the Causal Inference Workshop.

Du, Xinming. 2023. *Symptom or Culprit? Social Media, Air Pollution, and Violence.* Working Paper.

Engelberg, Joseph E., and Christopher A. Parsons. 2011. "The Causal Impact of Media in Financial Markets." *The Journal of Finance* 66 (1): 67–97. https://doi.org/https://doi.org/10.1111/j.1540-6261.2010.01626.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01626.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01626.x.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74. https://doi.org/10.1257/jel.20181020. https://www.aeaweb.org/articles?id=10.1257/jel.20181020.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78 (1): 35–71. ISSN: 00129682, 14680262, accessed April 18, 2024. http://www.jstor.org/stable/25621396.