Perspectives on Computational Analysis: Observational Data

Dr. Richard W. Evans

October 8, 10, 15, 2018

Data and Methods in MACSS

- MACSS 30000 focuses on data and research strategies
- MACSS 30100 focuses on methods and tools
- Data and methods overlap in both courses

Data and Methods in MACSS

- MACSS 30000 focuses on data and research strategies
- MACSS 30100 focuses on methods and tools
- Data and methods overlap in both courses

This Course: Perspectives on Computational Analysis

Interspersed in this course will be some methods as examples, but focus is on data and strategies (approaches).

Large digital data

Def’n: Digital traces

Digital data byproduct of people’s everyday behavior

Large digital data

Def’n: Digital traces

Digital data byproduct of people’s everyday behavior

- What digital traces have you left already today?
- What social scientific questions could we answer?

Observational Data

Def'n: Observational data

Any data that results from observing a social system without intervening in some way.

Observational Data

Def'n: Observational data

Any data that results from observing a social system without intervening in some way.

- Does NOT involve:
- DOES involve:

Observational Data

Def'n: Observational data

Any data that results from observing a social system without intervening in some way.

- Does NOT involve:
  - Talking with people (surveys); also does not include large scale government surveys
  - Deliberately changing people's environments (experiments)
- DOES involve:

Observational Data

Def’n: Observational data

Any data that results from observing a social system without intervening in some way.

- Does NOT involve:
  - Talking with people (surveys); also does not include large scale government surveys
  - Deliberately changing people’s environments (experiments)
- DOES involve:
  - Many traditional digital datasources: Twitter, Facebook, IRS, outside sensors, cell phone, newspapers, images, satellite images.
  - Natural experiments

Two Approaches

Two general approaches to research (obs. data or other)

Two Approaches

Two general approaches to research (obs. data or other)

1. Purist (high risk, high return)

2. Incrementalist (low risk, low return)

Two Approaches

Two general approaches to research (obs. data or other)

1. Purist (high risk, high return)
   - Benefits
   - Drawbacks
2. Incrementalist (low risk, low return)
   - Benefits
   - Drawbacks

Two Approaches

Two general approaches to research (obs. data or other)

1. Purist (high risk, high return)
   - Benefits
     - Might be first person to use, contribution likely impactful
   - Drawbacks
     - You might not be able to get the data
     - If first to data, might require significant cleaning
2. Incrementalist (low risk, low return)
   - Benefits
     - Already have the dataset
     - Can build off of work someone else has done
     - Innovation often new method or combo of data
   - Drawbacks
     - Most “low-hanging fruit” has been plucked
     - More difficult to do something impactful

One Data Characteristics Decomposition

- **Complexity**
- **Cleanliness**
- **Representativeness**
- **Source**

One Data Characteristics Decomposition

- **Complexity**
    - dimensionality, uniformity, frequency, format
- **Cleanliness**
- **Representativeness**
- **Source**

One Data Characteristics Decomposition

- **Complexity**
  - dimensionality, uniformity, frequency, format
- **Cleanliness**
  - How well organized and codified are the data?
  - How and why the data were collected?
  - Often, the data you have are not the data you want
- **Representativeness**
- **Source**

One Data Characteristics Decomposition

- **Complexity**
  - dimensionality, uniformity, frequency, format
- **Cleanliness**
  - How well organized and codified are the data?
  - How and why the data were collected?
  - Often, the data you have are not the data you want
- **Representativeness**
  - Who is in the data and who isn't?
  - How and why was the selection made?
  - How does that selection influence your ability to answer social scientific questions?
- **Source**

One Data Characteristics Decomposition

- **Complexity**
  - dimensionality, uniformity, frequency, format
- **Cleanliness**
  - How well organized and codified are the data?
  - How and why the data were collected?
  - Often, the data you have are not the data you want
- **Representativeness**
  - Who is in the data and who isn't?
  - How and why was the selection made?
  - How does that selection influence your ability to answer social scientific questions?
- **Source**
  - Corporate, government (public), academic
  - Source is important for motivation
  - Corporate: Twitter, Google, Banks, Healthcare
  - Government (public): administrative records, IRS, schools, vital statistics, social security, Medicare

Main Observational Data Decomposition

Generally helpful for research

- Big
- Always on
- Nonreactive

Generally problematic for research

- Incomplete
- Inaccessible
- Nonrepresentative
- Drift
- Algorithmically confounded
- Dirty
- Sensitive

Large digital data

Common Definition: Big Data

- So large in required memory that it must be stored across multiple hard drives (requires specific access and analysis tools)
- So complex that it requires multiple cores of processors to manipulate.

Large digital data

Common Definition: Big Data

- So large in required memory that it must be stored across multiple hard drives (requires specific access and analysis tools)
- So complex that it requires multiple cores of processors to manipulate.

WARNING

Use of the term “big data” is out of fashion

- Use the term “**large data**” or just “**data**”
- “large data” is basically a protest of “big data”
- Most data is now large data
- Most software has implemented tools for working with large data

Large digital data

- Good for study of rare events
- Good for heterogeneity
- Large N gives significance bias
- Complexity can be a problem

Always on

Data that are being constantly collected

- High frequency panel data
- Time trends
- Unexpected events
- Natural experiments

Nonreactive

Measurement can change behavior

Large observational data are usually nonreactive to observation because the individuals creating the data often don’t know their data are being recorded, or they are so used to it that it doesn’t change their behavior.

- Experiments often change people’s behavior

Incomplete

- Finish this slide

Inaccessible

- Finish this slide

Nonrepresentative

- Finish this slide

Drift

- Finish this slide

Algorithmically Confounded

- Finish this slide

Dirty

- Finish this slide

Sensitive

- medical, financial, sexual, other personal
- emotional harm, embarrassment
- economic harm, loss of employment
- legal harm
- Hard to decide what is sensitive (Netflix prize lawsuit)
- Privacy is a high standard

Research Strategies

1. Counting things

2. Forecasting and nowcasting

3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
   - Synonyms for count: estimate, fit, describe, relate
   - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning

2. Forecasting and nowcasting

3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
   - Synonyms for count: estimate, fit, describe, relate
   - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
   - Farber (2015) taxi cab data
   - Kossinets and Watts (2009) social network
2. Forecasting and nowcasting
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
   - Synonyms for count: estimate, fit, describe, relate
   - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
   - Farber (2015) taxi cab data
   - Kossinets and Watts (2009) social network
2. Forecasting and nowcasting
   - nowcasting = present or short horizon forecasting
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
   - Synonyms for count: estimate, fit, describe, relate
   - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
   - Farber (2015) taxi cab data
   - Kossinets and Watts (2009) social network
2. Forecasting and nowcasting
   - nowcasting = present or short horizon forecasting
   - Statistical learning models ("machine learning") seem to have the best forecasting performance
     - We will study these in depth next term
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
    - Synonyms for count: estimate, fit, describe, relate
    - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
    - Farber (2015) taxi cab data
    - Kossinets and Watts (2009) social network
2. Forecasting and nowcasting
    - nowcasting = present or short horizon forecasting
    - Statistical learning models ("machine learning") seem to have the best forecasting performance
        - We will study these in depth next term
    - Horse races are important. Compare one forecast against some baseline.
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
   - Synonyms for count: estimate, fit, describe, relate
   - Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
   - Farber (2015) taxi cab data
   - Kossinets and Watts (2009) social network
2. Forecasting and nowcasting
   - nowcasting = present or short horizon forecasting
   - Statistical learning models ("machine learning") seem to have the best forecasting performance
     - We will study these in depth next term
   - Horse races are important. Compare one forecast against some baseline.
   - Gopalan (2018) MACSS thesis
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things
2. Forecasting and nowcasting
3. Approximating experiments (natural experiments)

Research Strategies

1. Counting things

2. Forecasting and nowcasting

3. Approximating experiments (natural experiments)
   - Causal inference is hard (endogeneity).
     - Read a book [see Imbens and Rubin (2015), Pearl (2009), Morgan and Winship (2014)]
     - Take a class in causal inference or econometrics

Research Strategies

1. Counting things

2. Forecasting and nowcasting

3. Approximating experiments (natural experiments)
    - Causal inference is hard (endogeneity).
        - Read a book [see Imbens and Rubin (2015), Pearl (2009), Morgan and Winship (2014)]
        - Take a class in causal inference or econometrics
    - Rosenzweig and Wolpin (2000) on natural experiments.

Research Strategies: Counting things

- Synonyms for count: estimate, fit, describe, relate
- Descriptive papers, estimation papers, network analysis, causal inference, natural language processing, labeling with supervised learning
- Farber (2015) taxi cab data
- Kossinets, et al (2009) social network

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

- Example from Zunda’s MACSS thesis

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

- Example from Zunda's MACSS thesis
  - Worked on model with interesting features.

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

- Example from Zunda's MACSS thesis
  - Worked on model with interesting features.
  - What are the most interesting features?

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

- Example from Zunda's MACSS thesis
  - Worked on model with interesting features.
  - What are the most interesting features?
  - What is interesting question that uses those features?

Research Strategies: Counting things

LESS EFFECTIVE strategy according to Salganik (2018)

I am going to count something that no one has ever counted before.

MORE EFFECTIVE strategy according to Salganik (2018)

Look for questions that are either important or interesting (policy relevant).

- Sometimes means just counting things in different way

- Example from Zunda's MACSS thesis
  - Worked on model with interesting features.
  - What are the most interesting features?
  - What is interesting question that uses those features?
  - Precautionary labor supply

Farber (2015) taxi cab data

1. Better (larger) data than Camerer, et al (1997)

Farber (2015) taxi cab data

1. Better (larger) data than Camerer, et al (1997)
2. Great question: test two theories
   - Behavioral reference dependence: because of target earnings, higher wage results in reduced hours worked: Camerer, et al (1997)'
   - Neoclassical: higher wage results in increased hours
   - Random shock in Farber (2015) is weather.

Farber (2015) taxi cab data

1. Better (larger) data than Camerer, et al (1997)
2. Great question: test two theories
    - Behavioral reference dependence: because of target earnings, higher wage results in reduced hours worked: Camerer, et al (1997)'
    - Neoclassical: higher wage results in increased hours
    - Random shock in Farber (2015) is weather.
3. Taxi data is public (see notebooks)
4. Opened up more counting of similar data (Uber, Lyft, other taxi data)

Farber (2015) taxi cab data

1. Better (larger) data than Camerer, et al (1997)
2. Great question: test two theories
   - Behavioral reference dependence: because of target earnings, higher wage results in reduced hours worked: Camerer, et al (1997)'
   - Neoclassical: higher wage results in increased hours
   - Random shock in Farber (2015) is weather.
3. Taxi data is public (see notebooks)
4. Opened up more counting of similar data (Uber, Lyft, other taxi data)

Can we come up with another question?

Question that leverages key features

Kossinets and Watts (2009) social network data

- Assignment 2, exercise 3
- Great example of bringing a lot of new tools to a dataset in order to count things differently
- Network analysis
- Language processing

Kossinets and Watts (2009) social network data

- Assignment 2, exercise 3
- Great example of bringing a lot of new tools to a dataset in order to count things differently
- Network analysis
- Language processing

Major missed opportunity

No great network visualization in paper

Research Strategies: forecasting and nowcasting

- nowcasting is just present or short horizon forecasting
- Statistical learning models ("machine learning") seem to have the best forecasting performance
  - We will study these in depth next term
  - OLS is dominated for prediction
- Horse races are important. Compare one forecast against some baseline.
- Gopalan (2018, MACSS thesis), Predicting infant mortality

Gopalan (2018), Infant mortality

Gopalan, Sushmita, "Predicting Infant Mortality: Minimizing False Negatives," unpublished MACSS thesis (2018).

- Old data: National Family Health Survey India
- New methods: Logistic regression, Random forest, AdaBoost, XGBoost, class imbalance adjustment, minimize certain error (False negatives)

Gopalan (2018), Infant mortality

Gopalan, Sushmita, "Predicting Infant Mortality: Minimizing False Negatives," unpublished MACSS thesis (2018).

- Old data: National Family Health Survey India
- New methods: Logistic regression, Random forest, AdaBoost, XGBoost, class imbalance adjustment, minimize certain error (False negatives)

Results

- Prediction error rate for false negatives drops from 74% to 7% (a 90% decrease in error rate)
- False positive rate stays constant at 6%.

Research Strategies: Approximating experiments

- Causal inference is hard (endogeneity)
    - Read a book [see Imbens and Rubin (2015), Pearl (2009), Morgan and Winship (2014)]
    - Take a class on it (causal inference, econometrics)
- Two main approaches: natural experiments and matching
- Rosenzweig and Wolpin (2000) on natural experiments.
- Einav, et al (2015), eBay sales price data

Natural experiments

Def'n: Natural experiment

Something happens in the world that randomly assigns a treatment to some subjects while other subjects do not receive treatment (control)

Always-on Data $+$ Random Treatment $=$ Natural Experiment

Natural experiments

Def'n: Natural experiment

Something happens in the world that randomly assigns a treatment to some subjects while other subjects do not receive treatment (control)

Always-on Data $+$ Random Treatment $=$ Natural Experiment

Drawbacks

Natural experiments

Def'n: Natural experiment

Something happens in the world that randomly assigns a treatment to some subjects while other subjects do not receive treatment (control)

Always-on Data $+$ Random Treatment $=$ Natural Experiment

Drawbacks

- Is the treatment randomly assigned?

Natural experiments

Def'n: Natural experiment

Something happens in the world that randomly assigns a treatment to some subjects while other subjects do not receive treatment (control)

Always-on Data $+$ Random Treatment $=$ Natural Experiment

Drawbacks

- Is the treatment randomly assigned?
- Is everything else held constant (could be other treatments)

Angrist (1990), draft lottery, military service, earnings

Angrist, Joshua D., “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records,” *American Economic Review*, 80:3 (1990), pp. 313-336.

- Draft was random-ish: draw birthday’s at random
- U.S. Social Security Administrative data: always on

Angrist (1990), draft lottery, military service, earnings

Angrist, Joshua D., “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records,” *American Economic Review*, 80:3 (1990), pp. 313-336.

- Draft was random-ish: draw birthday’s at random
- U.S. Social Security Administrative data: always on
- Conclusion: Veterans’ earnings were, on average -15%

Mas and Moretti (2009), peer effect productivity

Mas, Alexandre and Enrico Moretti, “Peers at Work,” *American Economic Review*, 99:1 (2009), pp. 112-145.

- Does worker productivity increase when they have more productive coworkers?
- Always on grocery store cash registers
- Random assignment of who works with whom

Mas and Moretti (2009), peer effect productivity

Mas, Alexandre and Enrico Moretti, “Peers at Work,” *American Economic Review*, 99:1 (2009), pp. 112-145.

- Does worker productivity increase when they have more productive coworkers?
- Always on grocery store cash registers
- Random assignment of who works with whom
- Conclusion: Being with a co-worker with 10% higher productivity increases my productivity by 1.5%

Rosenzweig and Wolpin (2000), natural experiments

Rosenzweig, Mark R. and Kennith I. Wolpin, "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, 38:4 (Dec. 2000), pp. 827-874.

- Instrument (or natural experiment treatment variable) is something that is correlated with the variable of interest and not with the other variables variables.
- Major drawback is that assumed random treatment is often not random
- Absence of explicit models in natural experiment literature
- Conclusion:

Rosenzweig and Wolpin (2000), natural experiments

Rosenzweig, Mark R. and Kennith I. Wolpin, "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, 38:4 (Dec. 2000), pp. 827-874.

- Instrument (or natural experiment treatment variable) is something that is correlated with the variable of interest and not with the other variables variables.
- Major drawback is that assumed random treatment is often not random
- Absence of explicit models in natural experiment literature
- Conclusion:
  - Be careful about selection in random experiment.
  - Be humble about conclusions and assumptions.
  - Identification is always hard

Approximating Experiments: Matching

Def'n: Matching

Statistically adjusting non-experimental data to account for preexisting differences among those who did and did not receive the treatment.

- Involves pruning: getting rid of observations that have no obvious match.

Einav, et al (2015), auction start price vs. end price

Einav, Liran, Theresa Kuchler, Jonathan Levin, Neel Sundaresan, “Assessing Sale Strategies in Online Markets Using Matched Listings,” *American Economic Journal: Microeconomics*, 7:2 (2015), pp. 215-247.

How might you test this question?

Einav, et al (2015), auction start price vs. end price

Einav, Liran, Theresa Kuchler, Jonathan Levin, Neel Sundaresan, “Assessing Sale Strategies in Online Markets Using Matched Listings,” *American Economic Journal: Microeconomics*, 7:2 (2015), pp. 215-247.

How might you test this question?

- How do you deal with heterogeneity: sellers, products, participants?

Einav, et al (2015), auction start price vs. end price

Einav, Liran, Theresa Kuchler, Jonathan Levin, Neel Sundaresan, “Assessing Sale Strategies in Online Markets Using Matched Listings,” *American Economic Journal: Microeconomics*, 7:2 (2015), pp. 215-247.

How might you test this question?

- How do you deal with heterogeneity: sellers, products, participants?
- Could experiment: A/B test, individual sellers do

Einav, et al (2015), auction start price vs. end price

Einav, Liran, Theresa Kuchler, Jonathan Levin, Neel Sundaresan, “Assessing Sale Strategies in Online Markets Using Matched Listings,” *American Economic Journal: Microeconomics*, 7:2 (2015), pp. 215-247.

How might you test this question?

- How do you deal with heterogeneity: sellers, products, participants?
- Could experiment: A/B test, individual sellers do
- Einav, et al (2015) match many items start, end price to common reference (average price)
- Observation is a matched set: ex, TaylorMade Burner 09 Driver by `budgetgolfer`