Analysis with Simulated Data

Dr. Richard W. Evans

November 2, 2016

Machine Learning on *The Walking Dead*

- Matt Yancey, First Analytics (Indiana)
- Machine Learning to predict who dies
- http://livingwithdata.com/blog/wd-post.html

Syllabus

- Simulated data
- Data description and visualization
- Collaboration (Chap. 5, *Bit by Bit*)
- GitHub repo link

Assignment	Quantity	Points	Total Points	Percent
Short papers	4	15	60	50%
Problem sets	4	10	40	33%
Final exam	1	20	20	17%

Scientific Method

1. Formulate a question
2. State a hypothesis
3. Make predictions/implications from hypothesis
4. Testing/experiments
5. Analysis of results, determine what results show

Where we've been

Goal of the course

Social science applications of computational approaches to complex data

- Soltoff's section
  - Observational data
  - Collecting your own data
  - Experiments
  - Very good survey of readings

Ethics

- Collection of data can violate privacy
- Experiments can hurt individuals

What is simulation?

- the imitation of the operation of a real-world process or system over time, usually with a computational or mathematical model

- involves creation of synthetic data

When do we use simulation?

1. when experiments on human subjects are unethical
    - Do economies recover faster from recessions if the government provides no extra help to the poor?
    - University of Iowa “Monster Study”

When do we use simulation?

University of Iowa "Monster Study"

The Monster Study was a stuttering experiment on 22 orphan children in Davenport, Iowa, in 1939 conducted by Wendell Johnson at the University of Iowa. Johnson chose one of his graduate students, Mary Tudor, to conduct the experiment and he supervised her research. After placing the children in control and experimental groups, Tudor gave positive speech therapy to half of the children, praising the fluency of their speech, and negative speech therapy to the other half, belittling the children for every speech imperfection and telling them they were stutterers. Many of the normal speaking orphan children who received negative therapy in the experiment suffered negative psychological effects and some retained speech problems during the course of their life. Dubbed "The Monster Study" by some of Johnson's peers who were horrified that he would experiment on orphan children to prove a theory, the experiment was kept hidden for fear Johnson's reputation would be tarnished in the wake of human experiments conducted by the Nazis during World War II. The University of Iowa publicly apologized for the Monster Study in 2001.

When do we use simulation?

1. when experiments on human subjects are unethical
2. when no past data is available
   - How effective is massive monetary policy injection (TARP) in the face of a global equities crisis stemming from the U.S. housing market?
   - Does the use of iPads in K-12 education increase mathematics performance and global competitiveness?
3. when no past data is observable or has a lot of measurement error
   - Does happiness increase when a population becomes more equal (in terms of income)?
4. when real-world experiments are prohibitively expensive
   - black hole dynamics
   - estimating effects of all presidential candidate's tax policies

Simulation requires data generating process

True data generating process

The complete set of actors, forces, exogenous shocks, etc. that combine to make some observable phenomenon

Model data generating process

The incomplete and stylized set of actors, forces, exogenous shocks, etc. that combine to make some observable phenomenon

- The goal in modeling is to:
  - make the model DGP match the true data
  - incorporate the most important factors from the true process into the model process
- Simulation allows us to perform experiments

Example: How much will you earn?

- How much will you earn after you graduate?
- Talk about paper estimating the process from IRS data
- Look at Problem Set 1

Which comes first: Data or Theory?

Sherlock Holmes [quoted in Mankiw (2012), p. 17]

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to fit facts."

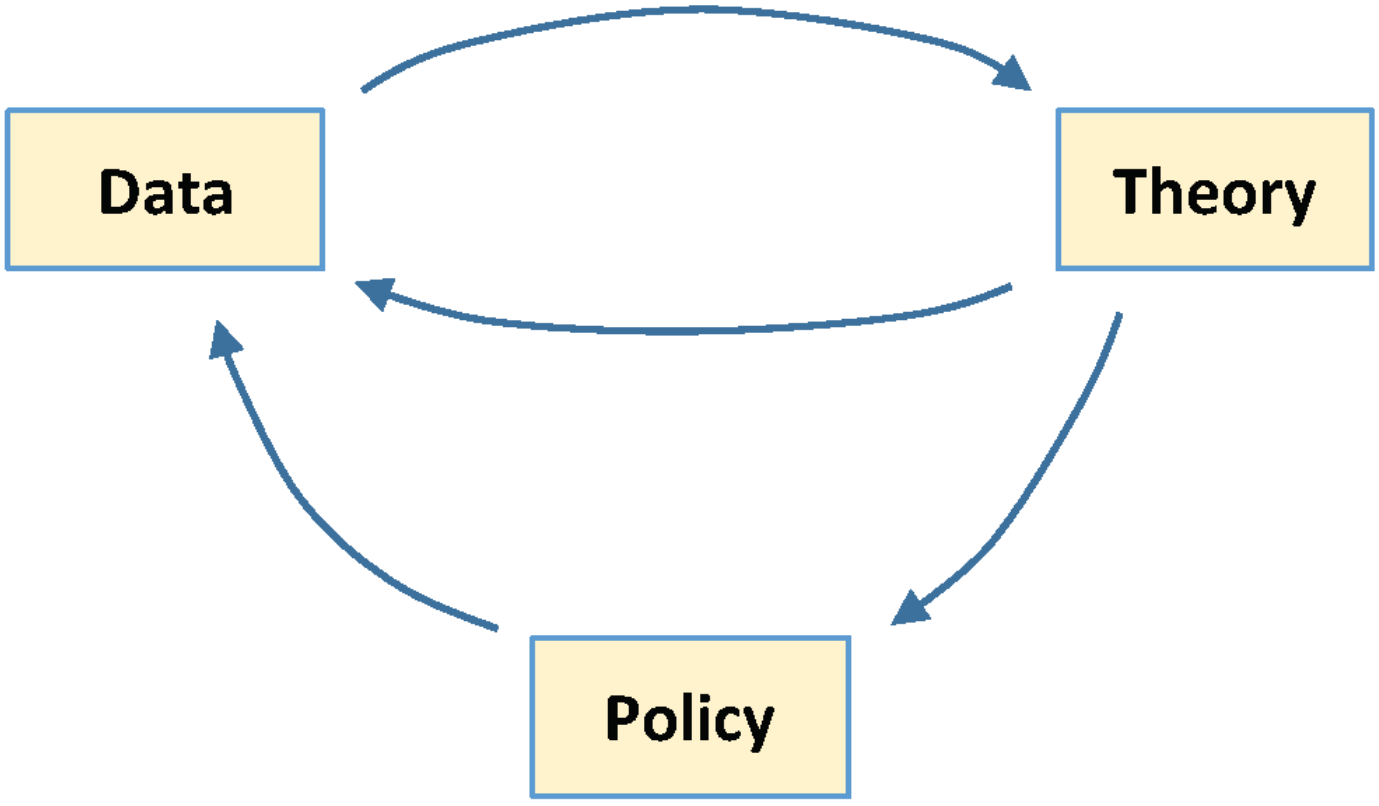
Michael Keane (2005), pp. 29-30

"...[W]e cannot even begin the systematic assembly of facts and empirical regularites without a preexisting theoretical framework that gives the facts meaning, and tells us which facts we should establish."

Multiple approaches to research

- Workshop speaker (Oct. 20), Scott Duke Kominers:
  - “top down” vs. “bottom up”
- Start with data and build to theory
- Start with theory and test the data
- Interplay of policy between data and theory (see figure)

Data and Theory and Policy



Limits of inference without theory

Definition: statisical inference

The process of deducing parameters θ of a model $\boldsymbol{f}(\boldsymbol{x}|\theta)$ from data.

Definition: simulation

The process of generating synthetic data $\boldsymbol{x}$ from a model $\boldsymbol{f}(\boldsymbol{x}|\theta)$ given parameter values θ.

Limits of inference without theory

- Koopmans (1947) essay, “Measurement without Theory”
  - “But, the decision not to use theories of man’s economic behavior, even hypothetically, limits the value to economic science and to the maker of policies, or the results obtained by the methods developed.”
  - Theory helps you know what data you should look for
  - Theory places a hierarchy on the importance of certain relationships between variables

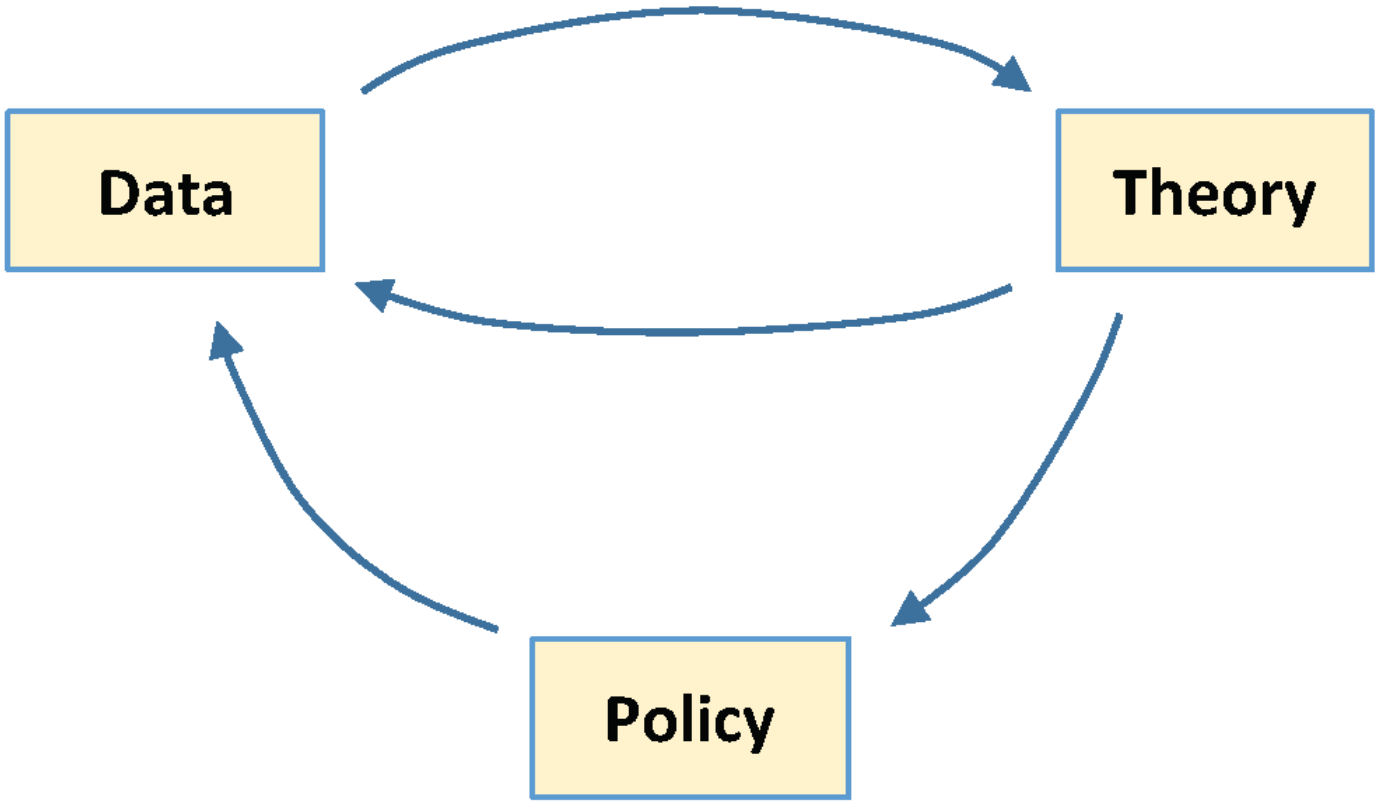
Limits of inference without theory

- "The absence of theory in inferential empirical work is pervasive. For example, of all the papers in the January 2009 maiden issue of the new *American Economic Journal: Applied Economics*, all of which were inferential, non contained an explicit model of 'man's economic behavior'."

- "...sharper inferences do indeed result from drawing explicit connections to coherent and relevant theory. The empirical approach, be it structural, quasi-structural, quasi-experimental, or experimental is of secondary importance. Theory provides the only way to fully appreciate the behavioral assumptions that underlie inference from data."

Data and Theory and Policy

- How could one neglect theory in this process?
- How should one include theory in this process?



Indirect inference: Take model directly to data

- Direct inference:
  1. Get data $\boldsymbol{x}$,
  2. Plug data $\boldsymbol{x}$ into model $\boldsymbol{f}(\boldsymbol{x}|\theta)$
  3. Estimate parameters $\hat{\theta}$ to best fit model output.

$$\min_{\boldsymbol{\theta}} \ \|\boldsymbol{f}(\boldsymbol{x}|\theta)\|$$

- Indirect inference:
  1. Simulate data $\boldsymbol{x}_{sim}$ from model $\boldsymbol{f}(\boldsymbol{x}_{sim}|\theta)$ given parameters θ
  2. Calculate a set of statistics from simulated data $\boldsymbol{m}(\boldsymbol{x}_{sim}|\theta)$ that is observable in real data $\boldsymbol{m}(\boldsymbol{x}_{real})$
     - We often do this if we can't observe the real world data $\boldsymbol{x}_{real}$ but we can observe some results of that data $\boldsymbol{m}(\boldsymbol{x}_{real})$
  3. Estimate parameters $\hat{\theta}$ to make model moments best fit data moments.

$$\min_{\boldsymbol{\theta}} \ \|\boldsymbol{m}(\boldsymbol{x}_{sim}|\theta) - \boldsymbol{m}(\boldsymbol{x}_{real})\|$$

Examples

- Income after graduation model, Problem Set 1

$$\ln(inc_{2018}) = \ln(inc_0) + \ln(\varepsilon_{2018}) \quad \text{and}$$

$$\ln(inc_t) = (1-\rho)\Big[\ln(inc_0) + g(t-2018)\Big] + \rho\ln(inc_{t-1}) + \ln(\varepsilon_t)$$

$$\text{for} \quad 2019 \leq t \leq 2057$$

  - Direct inference
  - Indirect inference

- FiveThirtyEight.com Methodology

Simulation and tax policy

- Baker, Bejarano, Evans, Judd, Phillips (2016), "A Big Data Approach to Optimal Sales Taxation"

- DeBacker, Evans, Phillips (2016), "Integrating Microsimulation Models of Tax Policy into a DGE Macroeconomic Model: A Canonical Example"