

# Issues with Experimental Data

January 23<sup>th</sup>, 2007

# Outline

1

## Data Types

- About Data
- Experimental Data
- NonExperimental Data
- Natural Experiment Data

2

## Pros & Cons of Each

- The Case against NonExperimental Data
  - How Experiments Circumvent Sample Selection Bias
- Other Reasons to Use Experimental Data
- The Case Against Experimental Data
- The Case against Natural Experiments
- The Case For Natural Experiments

3

## Where to Get Data

# Outline

1

## Data Types

- About Data
- Experimental Data
- NonExperimental Data
- Natural Experiment Data

2

## Pros & Cons of Each

- The Case against NonExperimental Data
  - How Experiments Circumvent Sample Selection Bias
- Other Reasons to Use Experimental Data
- The Case Against Experimental Data
- The Case against Natural Experiments
- The Case For Natural Experiments

3

## Where to Get Data

# Validity

All datasets face “Threats to Validity”

- Internal Validity
- External Validity
- Construct Validity

# Purpose of your Data

- Should provide exogenous variation in key explanatory variables
- Should provide comparison groups
- Should be able to help you probe the hypothesis in question

# What is Experimental Data?

## Definition

Data gathered from a **Controlled** experiment.

## Facets

- Typically have a **Control** group and a **Treatment** group.
- Assignment is **Random**.

# Prevalence in Economics

- Widely used in:
  - Medical Studies
  - Agricultural Experiments
  - Biological Research
- Moderately Used in Economics:
  - JTPA
  - NIT
  - HIE—Considered the best economic experiment
- Why??

# Prevalence in Economics

- Widely used in:
  - Medical Studies
  - Agricultural Experiments
  - Biological Research
- Moderately Used in Economics:
  - JTPA
  - NIT
  - HIE—Considered the best economic experiment
- Why??

# Prevalence in Economics

- Widely used in:
  - Medical Studies
  - Agricultural Experiments
  - Biological Research
- Moderately Used in Economics:
  - JTPA
  - NIT
  - HIE—Considered the best economic experiment
- Why??

# Prevalence in Economics

- Widely used in:
  - Medical Studies
  - Agricultural Experiments
  - Biological Research
- Moderately Used in Economics:
  - JTPA
  - NIT
  - HIE—Considered the best economic experiment
- Why??

# Prevalence in Economics

- Widely used in:
  - Medical Studies
  - Agricultural Experiments
  - Biological Research
- Moderately Used in Economics:
  - JTPA
  - NIT
  - HIE—Considered the best economic experiment
- Why??

# What is NonExperimental Data?

## Definition

Data gathered from microdata sources.

## Facets

- Usually survey or census data.
- Commonly referred to as **Observational** or **Econometric** data.

# What is a Natural Experiment?

## Definition

Data on outcome measures for observations in Treatment and Comparison Groups.

## Facets

- Treatment is **NOT** randomly assigned.
- Exogenous reason for treatment assignment.

# Outline

1

## Data Types

- About Data
- Experimental Data
- NonExperimental Data
- Natural Experiment Data

2

## Pros & Cons of Each

- The Case against NonExperimental Data
  - How Experiments Circumvent Sample Selection Bias
- Other Reasons to Use Experimental Data
- The Case Against Experimental Data
- The Case against Natural Experiments
- The Case For Natural Experiments

3

## Where to Get Data

# Sample Selection Bias

## A Simple Example

- Let  $Y_0$  and  $Y_1$  represent the outcomes of interest.
- Let  $D = 1$  indicate assignment to the treatment group.

# Sample Selection Bias-A Simple Example

## Goal:

- Estimate the **Treatment Effect (TE)**.
- $TE = Y_1 - Y_0$ .

## Counterfactual:

- Depending on assignment ( $D$ ) either  $Y_1$  or  $Y_0$  is unobserved.
- Estimation of TE becomes impossible.

# Sample Selection Bias-A Simple Example

## Goal:

- Estimate the **Treatment Effect (TE)**.
- $TE = Y_1 - Y_0$ .

## Counterfactual:

- Depending on assignment ( $D$ ) either  $Y_1$  or  $Y_0$  is unobserved.
- Estimation of TE becomes impossible.

# Sample Selection Bias-A Simple Example

## Change of Course

- Due to the counterfactual, interest lies in the distribution of TE,  $F(TE)$ .
- One places interest in particular moments of  $F$ , mean, median, variance, etc.

## Also of Interest

- **Treatment on the Treated (TTE).**
- $TTE = (Y_1 - Y_0 | D = 1)$ .

# Sample Selection Bias-A Simple Example

## Key Observation

- In nonexperimental settings people **SELECT** whether or not  $D = 1$ .
- In controlled experiments  $D = 1$  is **RANDOMIZED** across participants.

# Sample Selection Bias-A Simple Example

## Selection Bias

$$E(Y_0 | D = 1) - E(Y_0 | D = 0) \neq 0.$$

# The Benefit of Randomization

## Selection Bias No Longer Exists

- Let  $R = 1$  denote randomization **into** the Treatment group.
- Then  $E(Y_0 | D = 1) = E(Y_0 | R = 1 \text{ and } D = 0)$ ,  
**which is Observed.**

## Bias Balance

- Randomization acts as an instrumental variable.
- The bias is not removed, it is **BALANCED**.

# Why Use Experimental Data?

## Prime Benefits

- **Direction** of Causality is known.
- Nonexperimental Methods **CANNOT** solve the selection problem.
- Experiments are grounded on **More Plausible** assumptions.
- Results are **Easier** to Interpret and Explain to Noneconometricians.
- Experimental results lead to a **Consensus**.
- New policies, that have never before existed can be analyzed.

# Selection in nonexperimental data

## What to do?

Selection is an issue of missing data.

# Selection in nonexperimental data

What to do?

**COLLECT BETTER DATA**

Or

Use specification tests

# Selection in nonexperimental data

## What to do?

**COLLECT BETTER DATA**

Or

**Use specification tests**

# Plausible Assumptions

## Implicit to Experiments

- No Randomization Bias.
- No Substitution Bias.
- No Hawthorne Effect.

# Plausible Assumptions

## Implicit to Experiments

- No Randomization Bias.
- No Substitution Bias.
- No Hawthorne Effect.

## Definition

Randomization Bias:

Randomization alters the process of selection into the program.

# Plausible Assumptions

## Implicit to Experiments

- No Randomization Bias.
- No Substitution Bias.
- No Hawthorne Effect.

## Definition

Substitution Bias:

Those in the control group have access to a close substitute for the treatment.

# Plausible Assumptions

## Implicit to Experiments

- No Randomization Bias.
- No Substitution Bias.
- No Hawthorne Effect.

## Definition

Hawthorne Effect:

Program participants change their behavior within the experiment.

# Are Results Easier to Explain?

## Apparent vs. Real Simplicity

- Results are quite difficult to explain in the presence of Randomization or Substitution Bias.
- Before-After Comparisons are better; No need to explain control/treatment groups and no substitution bias is present.
- Evaluation centers on one or two parameters, not all the ones of interest though.

# Is there a Consensus?

## Better to have 'one' number

- Monopoly vs. Scholarship Consensus.
- Many experimental organizations are reluctant to share data.
- Example: SIME-DIME data (Hannan and Tuma 1990).

# Why not use Experimental Data?

## Prime Disadvantages

- **Cannot** answer many interesting questions
- **Intrinsic Variability** in Evidence.
- Randomization and Substitution Bias.
- Institutional Limitations.
- **Cannot** capture general equilibrium effects
- Controlled experiments are **Costly**
- Reference Frame
- Program Entry
- Sample Contamination
- Treatment Contamination

# Questions of Interest

## What if I want to know about

- Median Treatment Effect
- Variance of Individual Treatment Effects
- Fraction w/ a Positive Treatment Effect
- Overall Shape of Distribution of Treatment Effects

# Intrinsic Variability

## No Joint Distribution

Due to the counterfactual nature of experiments we never observe the joint distribution of  $Y_1$  and  $Y_0$ .

## Wide Range of Prospective Answers

Methods to assess key features of the joint distribution lead to a wide array of potential, plausible outcomes. This makes judging joint estimates tricky.

# Experimental Biases

## Randomization Bias

- Ethical and Public Relations Issues raise concerns over randomization, making bias more prevalent
- Selective Failure to Participate

## Substitution Bias

- What if, aside from the experiment, other agencies are also offering services?
- Some experiments actually offered a list of alternative service providers!

# Institutional Limitations

## Variety of Shortcomings

- Optimal Placement of Random Assignment
  - At which point do I randomize subjects into treatment and control?
- Could Make it Impossible to Evaluate Experimental Estimates of Individual Treatments
  - There could be more than one 'type' of treatment
  - Job Training vs. Job Search Assistance
- Voluntary Participation of Sites May Discredit External Validity
  - Treatment centers that volunteer may not be representative of nationwide treatment centers

# What Can Go Wrong?

## Prime Shortcomings

- Omitted Variables
- Trends in Outcomes
- Misspecified Variances
- Mismeasurement
- Political Economy
- Simultaneity
- Selection
- Attrition
- Omitted Interactions

# Why Look for Them?

## Advantages

- Low Cost
- Less Time Intensive

# Outline

1

## Data Types

- About Data
- Experimental Data
- NonExperimental Data
- Natural Experiment Data

2

## Pros & Cons of Each

- The Case against NonExperimental Data
  - How Experiments Circumvent Sample Selection Bias
- Other Reasons to Use Experimental Data
- The Case Against Experimental Data
- The Case against Natural Experiments
- The Case For Natural Experiments

3

## Where to Get Data

# Sources of Data

- Online Journal Data Archives—JAE, JBES, JHR
- Statlib
- BEA, BLS, NBER, Census, PSID, NLS, NLSY, etc.
- Centers at Universities (Harvard's Development Center for example)
- Researcher's Webpages,

# What to do with the Data?

- If you did not collect the data make sure you know how it was gathered and constructed
- Make sure you know the problems inherent in your data