# Causal Inference and Selection Bias

**TAF-CEGA Impact evaluation workshop**

Prashant Bharadwaj
University of California, San Diego

March 23rd, 2010

Material in this presentation developed from CEGA and World Bank materials.

# Outline

1. Why causal inference?
2. Sample selection
3. Omitted variables bias
4. From correlation to causation
    1. Experimental methods
    2. Quasi-experimental methods

# Why causal inference?

1. Simple correlations can lead to misguided policy
2. Among many different options, important to choose the *most effective* intervention
3. Accurate cost-benefit analysis possible

# Some facts

Singapore is one of the fastest growing economies of South East Asia

Singapore has one of the highest per capita execution rates in the world

Singapore has some of the best food anywhere in the world

# Misguided policy

If we only knew those facts, we would be tempted to make the following claims:

*Singapore’s high economic growth is due to its high execution rate*

*Singapore’s high execution rate is due to its amazing food*

*Singapore’s high growth is due to its amazing food*

**Cannot make informed policy based on correlations**

# Choosing the most effective option

Many ways to get at corruption and absenteeism in developing countries

1. Monitoring
    1. Top-down vs. grassroots
2. Smart cards in India
3. Cameras for students

Causal inference enables us to decide objectively across many options

# Helps with cost-benefit analysis

1. Accurate measures of effect size
2. Spillover benefits often missed
    1. ARV therapy for adults in a household
3. Sometimes even costs are difficult to compute

Causal inference helps infer program effectiveness

# Goal of causal inference

What is the impact of an intervention (X) on an outcome (Y)

1. Hard to evaluate
2. Need to compute *counterfactuals*
3. Challenge: same person cannot both get treatment and not get treatment

# Computing counterfactuals

The treated group and the counterfactual group should have identical characteristics on average, except for benefiting from the intervention

→ **only reason** for different outcomes between treatment and counterfactual is the **intervention**

# Wrong Counterfactual 1

**Before and After**

Same group of villagers before and after treatment/information program – community policing

Compare conflict before and after

Findings: Conflict is lower after the program

Did the program succeed?

# Wrong Counterfactual 1

<u>Before and After</u>

What else is happening over time?

- Poor and irregular rainfall?
- Other interventions
- **Effect of treatment and time-varying variables on outcome cannot be separated**

# Wrong Counterfactual 2

**<u>Compare participants to non-participants at the same time</u>**

**Non-participants:**

Communities who <u>choose not to enroll</u> in program: Communities who do not need community policing or support

or

Those who were <u>not offered</u> the program, ineligible: Richer communities, safer communities

# Wrong counterfactual 2

Villagers who are more enterprising, and have more to gain from community policing are the ones participating – hence, they would have participated regardless

Participants have differing (pre-existing) characteristics relative to non-participating communities and individuals that also affect outcomes of interest

**Non-participants → a poor counterfactual for treatment group**

# Barriers to causal inference

1. Sample selection
2. Omitted variables

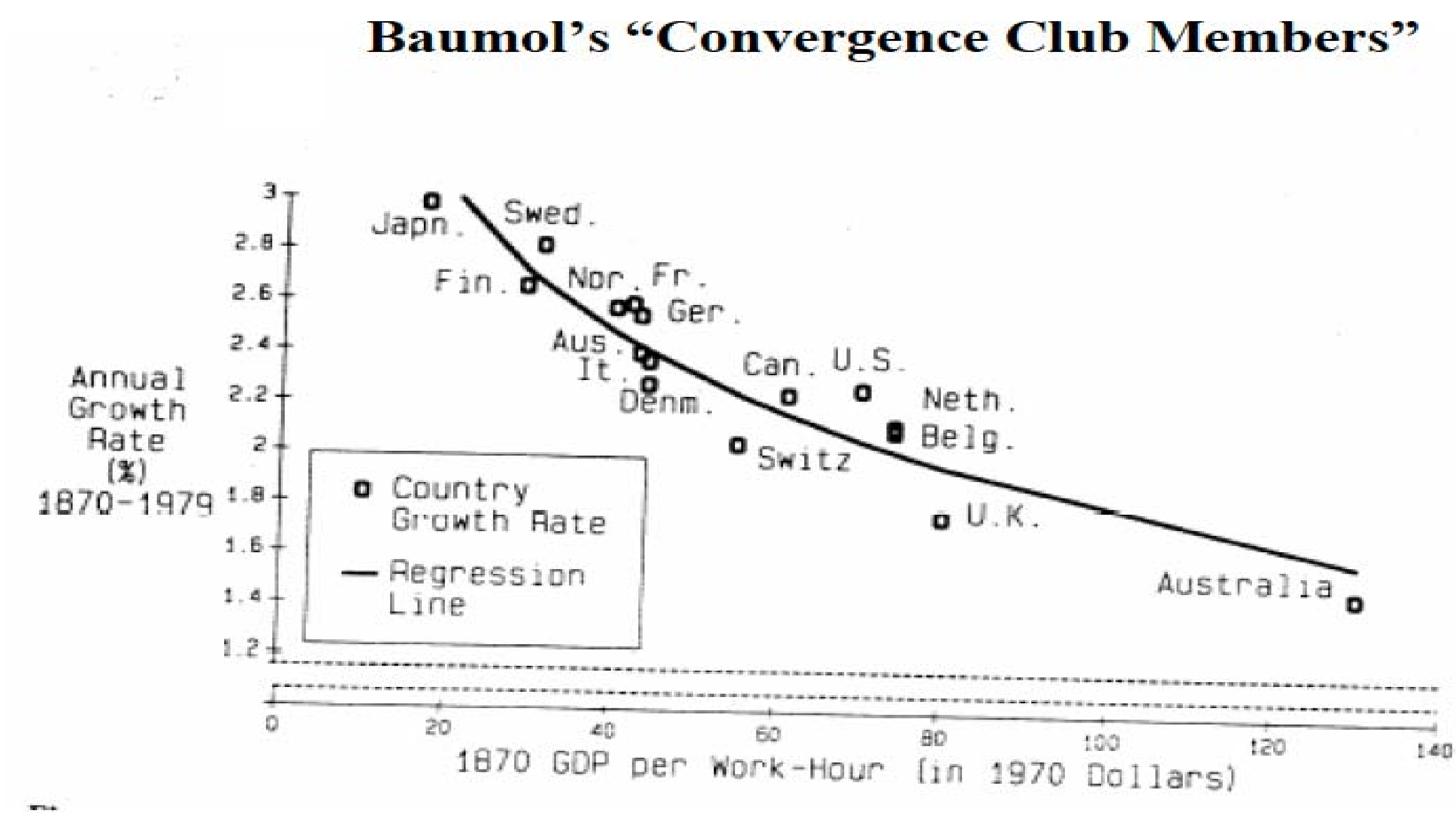
# Sample selection

1. People non-randomly select into various programs which we would like to evaluate
2. Type of data we use

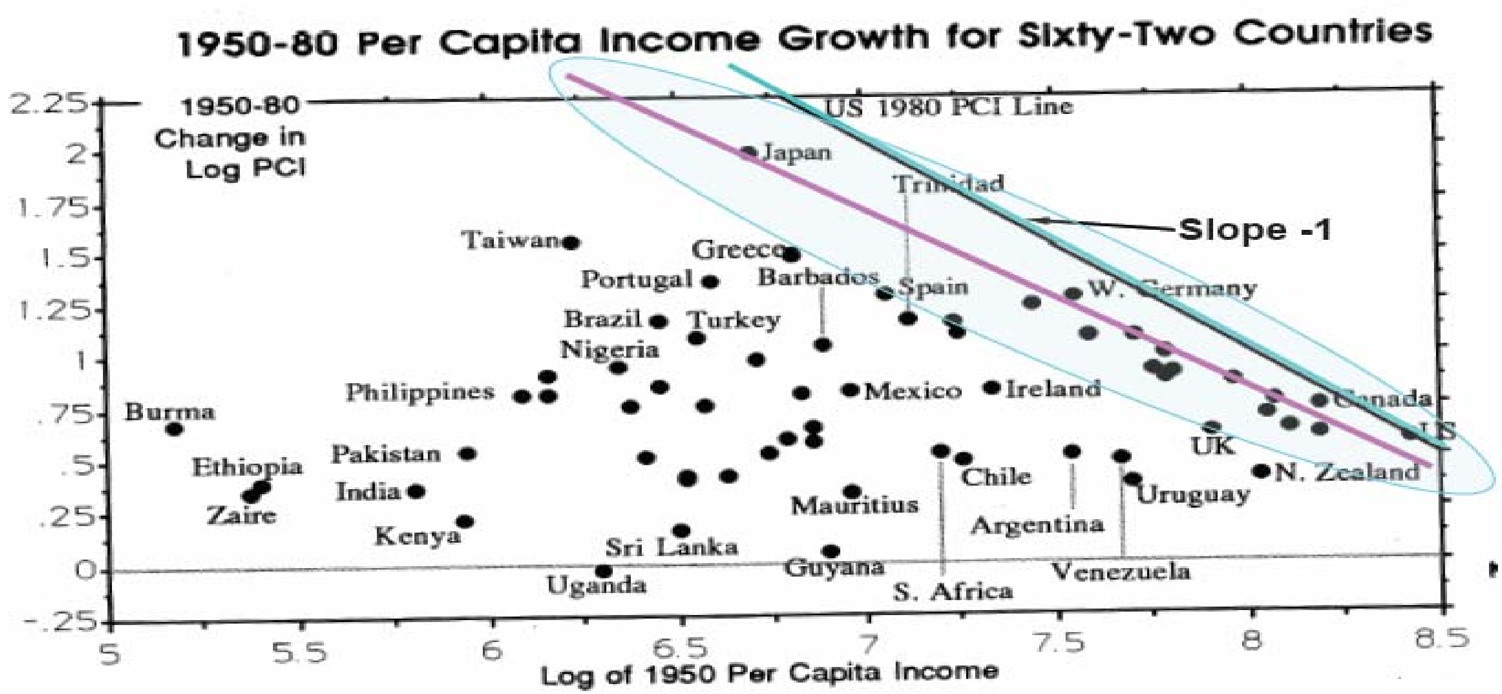
Examples:

1. Who shows up in court?
2. Migration

# Eg: Sample selection



# Eg: Sample selection




Where is the "convergence club" now? (DeLong)

# Omitted variables bias

Do radio and television destroy social capital?

- Negative *correlation* between television watching and social capital formation
- Perhaps societies that watch more TV are wealthier, and wealth correlated with participation
- *Reverse causation*: incompetent village head, less incentives to participate, hence watch TV more

# Instrumental variables

1. If randomization is not possible or too expensive, *natural experiments* can get at causal relationships
2. *Instrumental variables* is a popular way to address causality with non experimental data

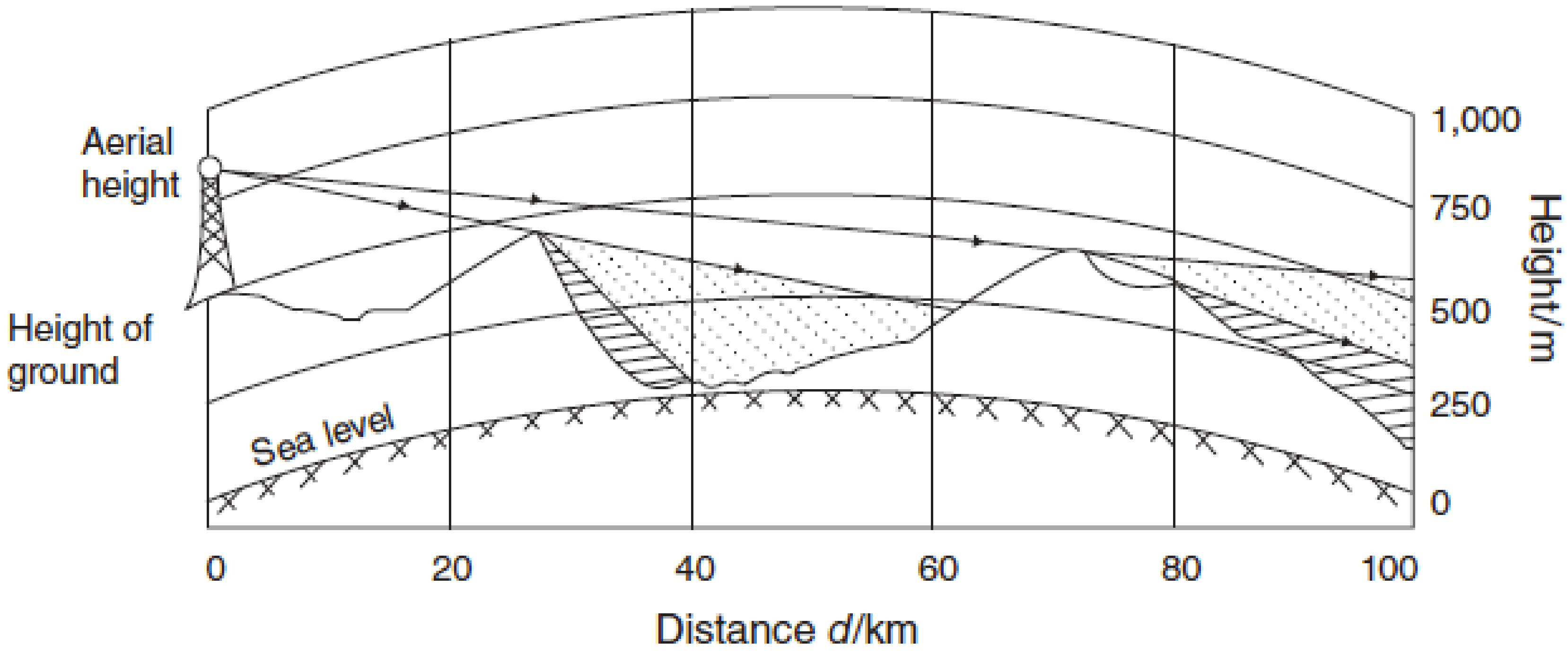
# Instrumental variables in math

$Y = b X + u$

1. Interested in the causal impact of X on Y
2. *u* contains a variable, M that affects both X and Y
3. To get to causality, need something else, Z that does the following:
   1. Affects X
   2. Is not related to anything that *u* might contain
   3. Why not just examine $Y = r Z + u$?
   4. *r* does not have the same economic interpretation as *b*

# IV in action

Topography plays a role in the strength of TV reception




*from: Olken (2009)*

# Getting around sample selection and OVB

Experimental methods

Quasi experimental methods

1. Instrumental variables
2. Difference in differences
3. Regression discontinuity
4. Propensity score matching

# Take away

When thinking about the causal effect of a program on some outcome ask yourself:

1. Are there other things correlated with the program and the outcome that could be driving the results?
2. Is the sample you have data for different for some reason that is correlated with the program/outcome

# Conclusions

- To identify effective interventions and compare alternatives, we need to be able to attribute causality
- Need a valid counterfactual: a group that would behave the same as the treated group in the absence of the intervention
- Invalid counterfactuals:
  - Before and after: time-varying variables
  - Participants vs. non-participants: characteristics
- Options: Choice of method depends on program design, operational considerations, and the question