INSTITUTE FOR DEFENSE ANALYSES

Introduction to Observational Studies

Dean Thomas, *Project Leader*

Yevgeniya K. Pinelis

March 2018

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-9020

Log: H 2018-000122

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-09-2299(90), "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For more information:
Dean Thomas, Project Leader
dthomas@ida.org • (703) 845-6986

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-9020

Introduction to Observational Studies

Dean Thomas, *Project Leader*

Yevgeniya K. Pinelis

Introduction to Observational Studies

Jane Pinelis

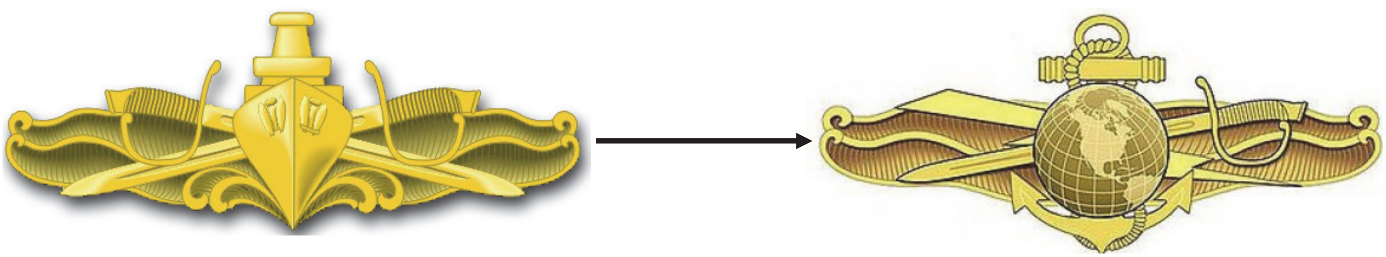
22 March 2018

Outline

- Motivating example
- Observational studies vs. randomized experiments
- Observational studies: basics
- Some adjustment strategies
- Matching / stratification
- Difference-in-difference estimators
- Instrumental variables

Should Navy officers be denied lateral transfer by their supplying communities?

- A Navy officer can apply for a lateral transfer to another community if openings exist.
- The lateral transfer board ensures the receiving community gets the best and fully qualified officers.
- Transfer also needs approval from the supplying community.
- Officers who get denied may leave the Navy.
  - Reason for denial is not recorded in the data.



Navy officer lateral transfers and retention

What's the **causal effect** of being denied on retention?

Should supplying community quotas be reconsidered?

Navy officer lateral transfers and retention

- Problem: Officers who get denied could be:
  - Not best and fully qualified for the job
  - Not great at their job
  - Needed in their current job
- Denied officers could be ‘worse’ than those who get approved
  - Failure to promote correlated with denial and loss rate
  - Are they likely to leave the Navy anyway?

Navy officer lateral transfers and retention

How do we **compare retention** among officers who got approved to that of officers who got denied?

Navy officer lateral transfers and retention

How do we **compare retention** among officers who got approved to that of officers who got denied?

Wouldn't it be nice if officers were approved / denied **at random**?

- Maybe for statisticians, but probably not for the Navy
- Approved officers are probably different from denied officers

Navy officer lateral transfers and retention

How do we **compare retention** among officers who got approved to that of officers who got denied?

Wouldn't it be nice if officers were approved / denied **at random**?

- Maybe for statisticians, but probably not for the Navy
- Approved officers are probably different from denied officers

How do we adjust for **observed and unobserved** qualities?

Navy officer lateral transfers and retention

How do we **compare** retention among officers who got approved to that of officers who got denied?

Wouldn't it be nice if officers were approved / denied **at random**?

- Maybe for statisticians, but probably not for the Navy
- Approved officers are probably different from denied officers

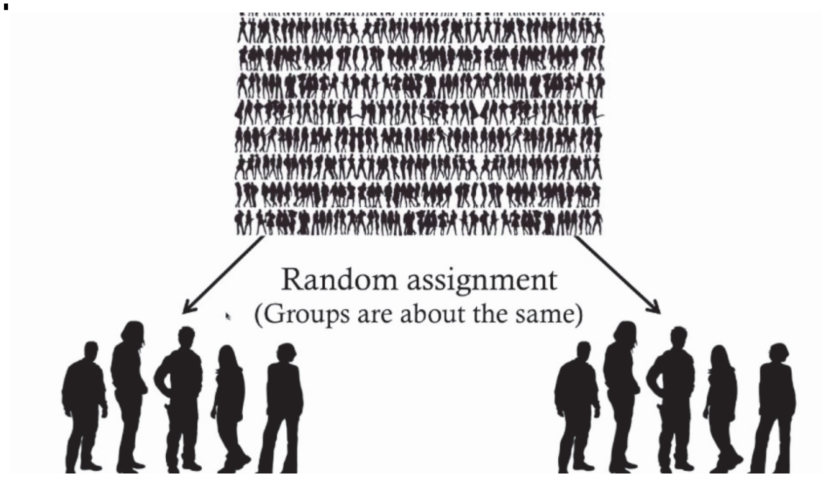
How do we adjust for **observed and unobserved** qualities?

Can we ever get to **causal** effects?

Randomized Experiments vs. Observational Studies

In a randomized experiment, ‘treatment’ is **randomly** assigned.

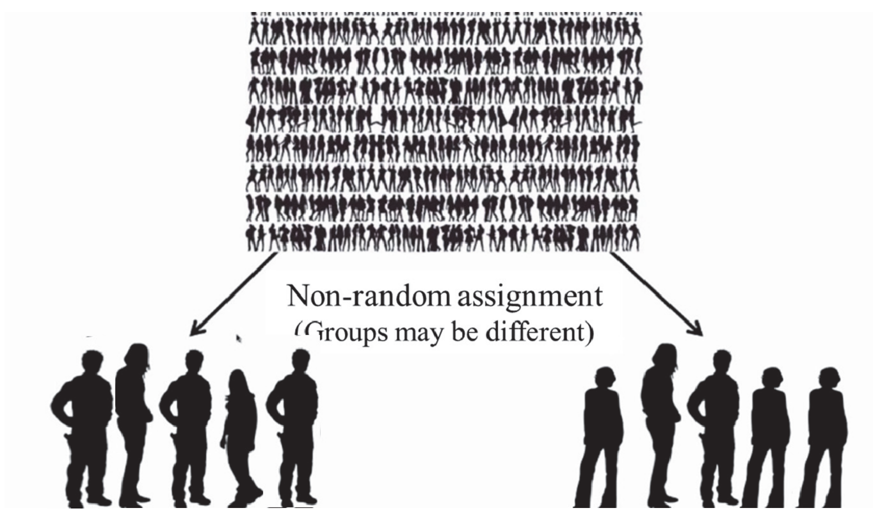
- Probability of being assigned to ‘treatment’ is the same for everyone (or everyone within a group).
- As n gets larger, observed and unobserved characteristics of the treated and control groups start approaching **balance.**
- Difference in outcomes can be attributed to treatment (**causation**).



Randomized Experiments vs. Observational Studies

In an observational study, ‘treatment’ assignment **may be** applied non-randomly.

- Different subjects may have **different probabilities of treatment assignments.**
- Observed and unobserved characteristics of treatment groups may not be balanced.
- Difference in outcomes between groups is much harder to attribute to treatment alone.



A bit of history

1950s and 1960s: interest in causal relationship between smoking and lung cancer

- Establishment of the field of observational studies

Cochran (1965) clarified the benefits of learning from reliably planned, measured, and analyzed observational studies.

- He provided an infrastructure for planning and analysis.

Cochran focuses on two main study characteristics:

- The objective is to elucidate cause-and-effect relationships.
- It is not feasible to use controlled experimentation.

Observational Studies: the basics

Cross-Sectional

- Individual-level data collected at a specific point in time

Case-Control

- Individual-level data collected for cases (subjects with the outcome of interest) vs controls

Cohort

- Following a cohort of subjects over time
- Can be prospective or retrospective

Longitudinal

- Repeated observations of variables over time

Ecological

- At least one variable is measured on the population level

Potential outcomes

Let r_{Ti} be the response of applicant officer i to being denied lateral transfer ('treatment') and r_{Ci} be the response of applicant officer i to being approved ('control'). Then the potential outcomes are:

- r_i=1 if officer leaves the Navy
- r_i=0 if officer stays in the Navy

For each officer i, potential outcomes and treatment effects are:

r_{Ti}	r_{Ci}	δ_i	Explanation
0	0	0	Officer stays in the Navy no matter what
0	1	-1	Approval causes the officer to leave the Navy
1	0	1	Denial causes the officer to leave the Navy
1	1	1	Officer leaves the Navy no matter what

The fundamental problem of causal inference

- For officer i, the treatment effect is δ_i = $r_{Ti} - r_{Ci}$
- Average treatment effect (ATE) for the sample is $\frac{1}{n}\sum_{i=1}^{n} \delta_i$
- You could also estimate
  - Attributable effect
    - Number of events among treated subjects that were caused by the treatment (the number of officer losses that were caused by denials)
  - Average effect of treatment on the treated (ATT)
- **For each officer, we observe only r_{Ti} or r_{Ci} but never both**
- Sample treatment effect estimation is an issue of inference and not arithmetic
- Defined in Holland (1986)

Causal inference – statistical questions

In randomized experiments

- Does denial ***cause*** officers to leave the Navy? (tests of no effect)
  - Fisher 1935 – randomization inference
- How much more likely is a denied officer to leave the Navy? (estimates of magnitude of the effect)

Causal inference – statistical questions

In observational studies

- What could the officers have done if approved or denied? (Potential outcomes framework)
  - Neyman 1923, Rubin 1974
- Adjustment for officer demographics and quality (overt biases)
  - Tests of no effect
  - Estimates of magnitude of the effect
- What if we missed something important? (sensitivity to hidden bias)

Some adjustment strategies

- Matching / Stratification
  - Propensity Scores
  - Prognostic Scores
- Difference-in-difference estimators
- Instrumental Variables

- Multiple other schools of thought
  - Recommended reading: *Causality* by Judea Pearl

Some standard assumptions

The Stable Unit Treatment Value Assumption (SUTVA)

- Each officer decides to stay or leave the Navy regardless of other officers' approval / denial or the approval process
- Potential outcomes for a subject are independent of treatment assignment for all other units and of the assignment mechanism
- SUTVA is usually assumed, but is rarely tested
- Interference between units can result in violations
  - Rubin (1990)

Some standard assumptions

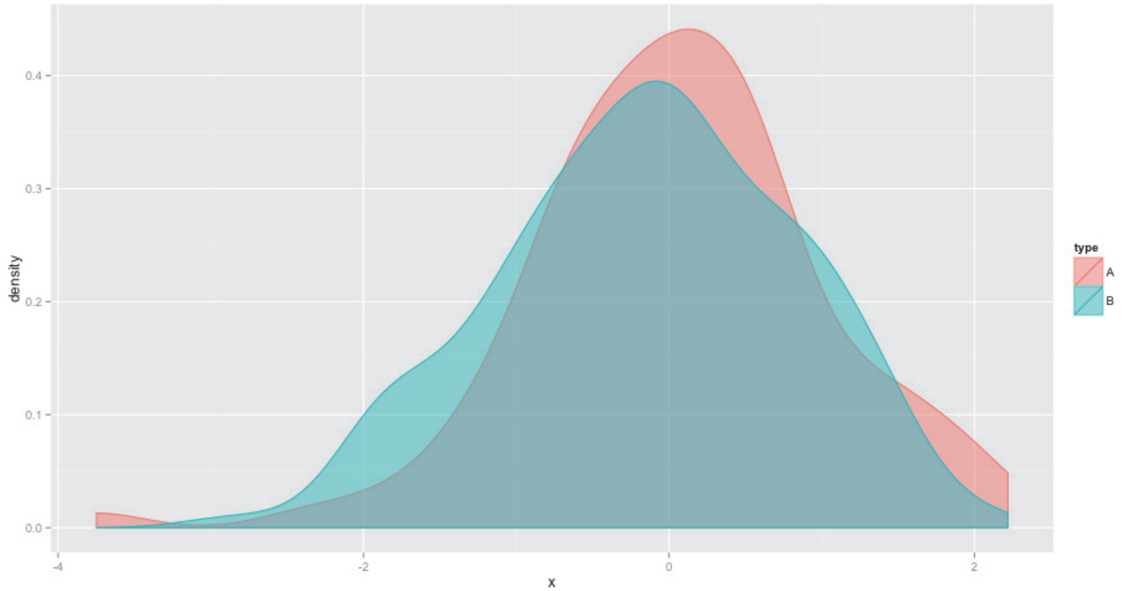
Strong Ignorability of Treatment Assignment

- Application approval or denial depends only on variables we measured and recorded
- A.k.a. Conditional Independence Assumption (CIA)
  - Rosenbaum and Rubin (1983)
- Assumes selection into treatment based on observed covariates
- Critical in matching, stratification, and covariance adjustment

Some standard assumptions

Common Support Condition

- Each officer can be approved and denied
- Probability of assignment to treatment is bounded away from zero and one
- Rosenbaum and Rubin (1983)



Curse of dimensionality

- There are a lot of variables that matter to approval (demographics, officer quality, accession source, etc.)
  - Concern about having to adjust for many potentially causal or "important disturbing variables" (Cochran 1965)
- 20 covariates each with just 2 levels results in over a million categories
  - Exact matches are hard to find
  - Approximate matches are hard to characterize
  - Rosenbaum and Rubin (1985)
- Hence the focus on dimension-reduction techniques
  - Propensity score (Rosenbaum and Rubin, 1983)
  - Prognostic score (Hansen, 2007)

Propensity scores

- In an experiment, we would compare officers who are similar in all important respects except for getting denied (the ‘treatment’).
- We can create such data configurations using *propensity score matching.*
  - Propensity score for each officer is the estimated probability of getting denied lateral transfer given their demographics and quality.
  - The propensity score the probability of “treatment” given observed covariates.
  - It reduces a multivariate **X** to a one-dimensional score.
  - Matching on it should balance variables between the two groups.
  - Matching can result in unbiased estimates of treatment effects.
  - Importantly, we can check whether matching ‘worked’ before we proceed with analysis of the impact of approval / denial.

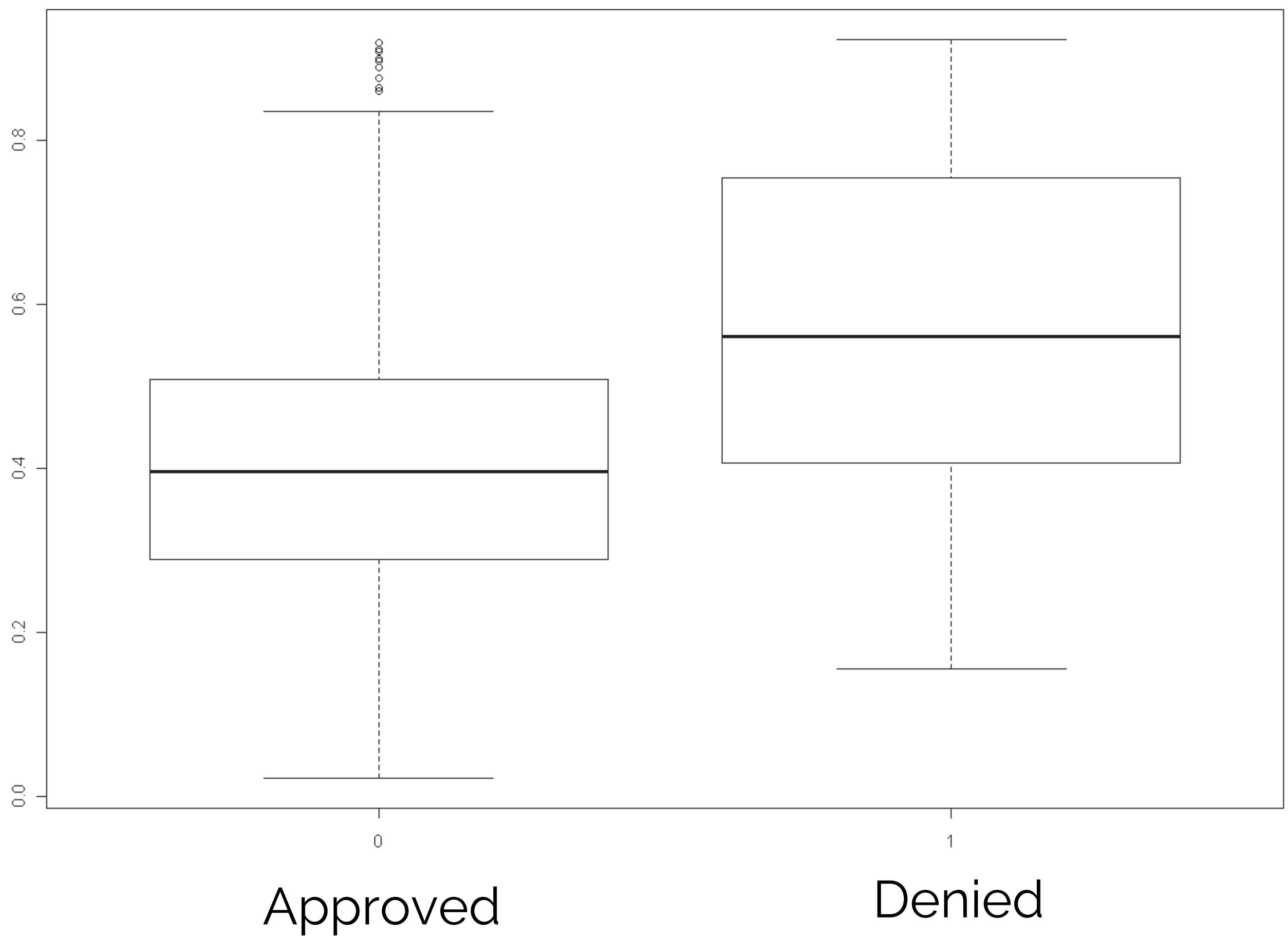
Propensity scores

- Vary with covariates for each officer
- Can be higher for denied officers (treated subjects)
- Overlap is important!
- Are an estimated quantity and that's OK
  - Subclassification on the propensity score should balance the observed covariates that went into its estimation.
- Within subclasses, the joint distribution of observed covariates should be similar between treated and control subjects.

Propensity score shortcomings (Rubin 1997)

- They only help adjust for observed covariates, and unobserved to the extent that they are correlated with observed.
- They work better in large samples.
- Covariates related to treatment assignment and not the outcome are treated the same as the ones strongly related to the outcome and not treatment.
- Misspecification is difficult to diagnose, and the consequences of it are elusive.

Officer Propensity Scores



Prognostic Scores

- Basic idea: Not all covariates are created equal.
  - Balancing covariates strongly related to the outcome may be more important (Hansen 2008).
- The prognostic score measures the relationship between observed variables and potential outcomes.
  - First, retention (the outcome) is modeled just for officers who got approved (in the control group).
  - Then, the obtained model is used to predict retention (the response) for officers who got denied (in the treated group).
  - The fitted values from the model are the prognostic score.
- Allows the comparison of officers who would have responded similarly to being approved.
- Controversial practice of using outcomes at this stage of the analysis.

Stratification and Matching

- An attempt to 'recover' the hidden block-randomized experiment from observational data (Hansen, 2009)

- Stratification first addressed by Cochran (1968)

- Matching options
  - On covariates
  - On propensity or other scores
  - Within calipers

- Matching algorithms
  - Greedy / nearest-neighbor
  - Optimal

Balance assessments

How do you know if matching / stratification worked?

- Are observed covariates any more balanced than they were before?
- Unobserved covariates balance to the extent that they are correlated with observed covariates that got balanced.

Balance assessments

To test balance or not?

- Unresolved debate in statistical literature
- Population hypothesis tests
- Randomization inference

Inference

- Standard inference approaches apply
- Randomization inference (tests of no effect)
  - Fisher exact test
  - Wilcoxon's signed rank and rank sum tests,
  - Mantel-Haenszel-Birch test
  - Logrank test
- Parametric techniques
  - ANOVA (comparing groups)
  - Regression (estimating treatment effects)
- Nonparametric covariance adjustments (Rosenbaum 2002)

Slide 31

AMR10 What are all of these things testing? I find people respond better when you talk about the inference/outcome you're testing for ("Are the characteristics of the two groups the same?" for example) rather than just give the names of the tests. The names of the tests are useful if people want to go back later and find more info/perform the test, but they often just serve to confuse people when context for their purpose isn't understood

Avery, Matthew R, 3/14/2018

Sensitivity Analysis

Basic questions:

- How big of an effect does my missing variable have to have in order to break down my result?
- How likely is a variable like that to exist?

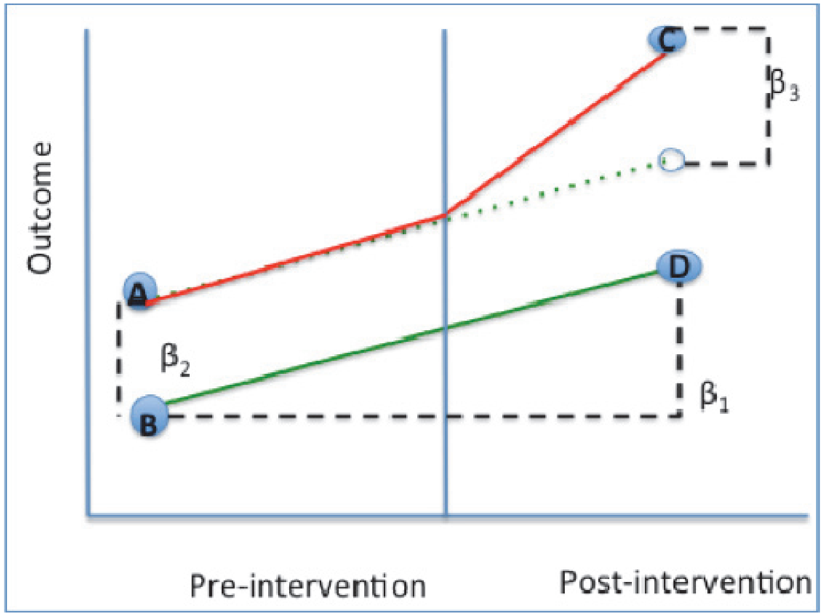
Difference-in-Difference Estimators

- Compares the average change in outcome for the treatment group to the average change in outcome for the control group
- Uses panel data to compare differences using longitudinal data
- Assumptions:
  - Standard OLS assumptions
  - SUTVA
  - Parallel trends assumption (in the absence of treatment, the difference between the 'treatment' and 'control' group is constant over time)

Difference-in-Difference Estimators

Usually implemented as an interaction term between time and treatment group dummy variables in a regression model

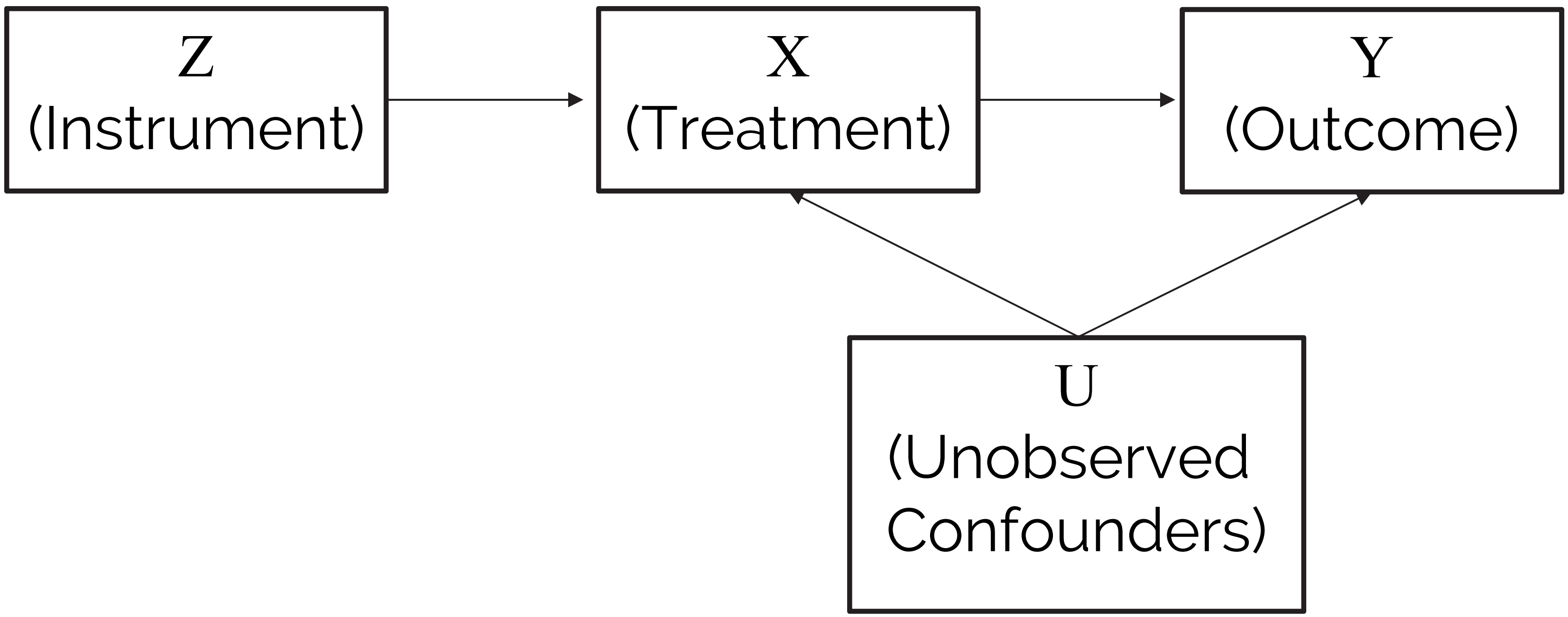
Coefficient	Calculation	Interpretation
β_0	B	Baseline average
β_1	D-B	Time trend in control group
β_2	A-B	Difference between two groups pre-intervention
β_3	(C-A)-(D-B)	Difference in changes over time



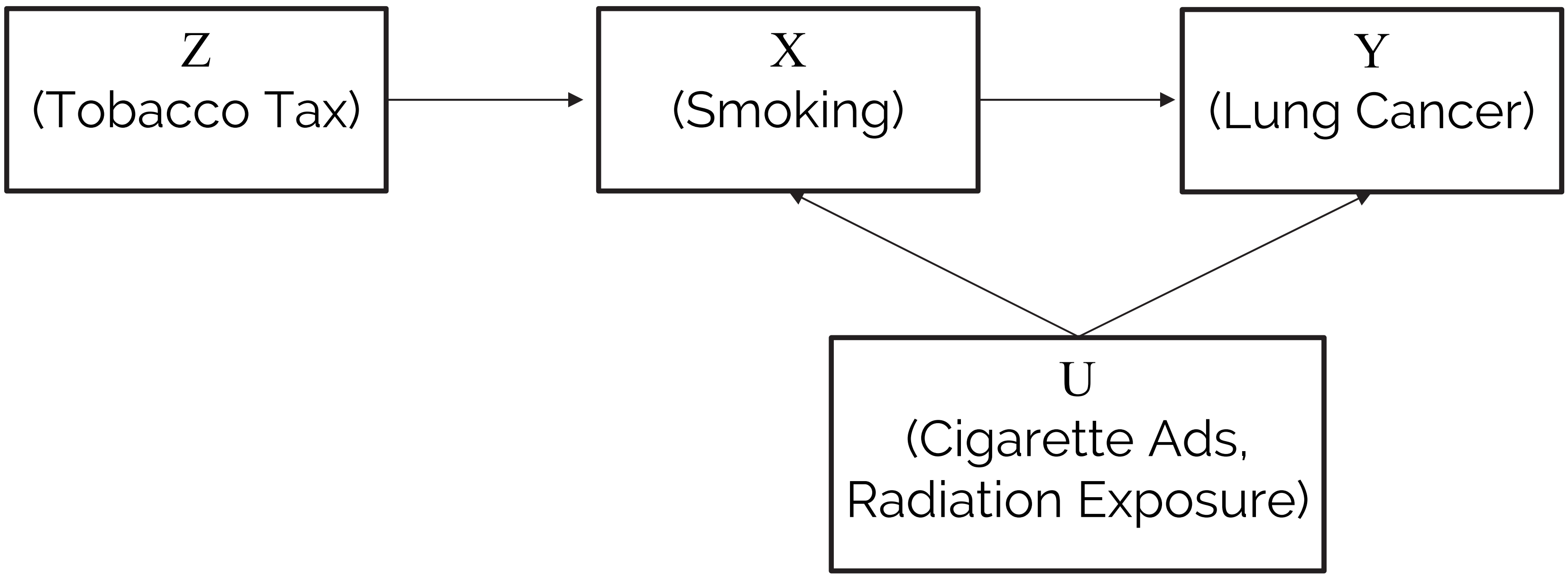
Source: https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation

Instrumental variables

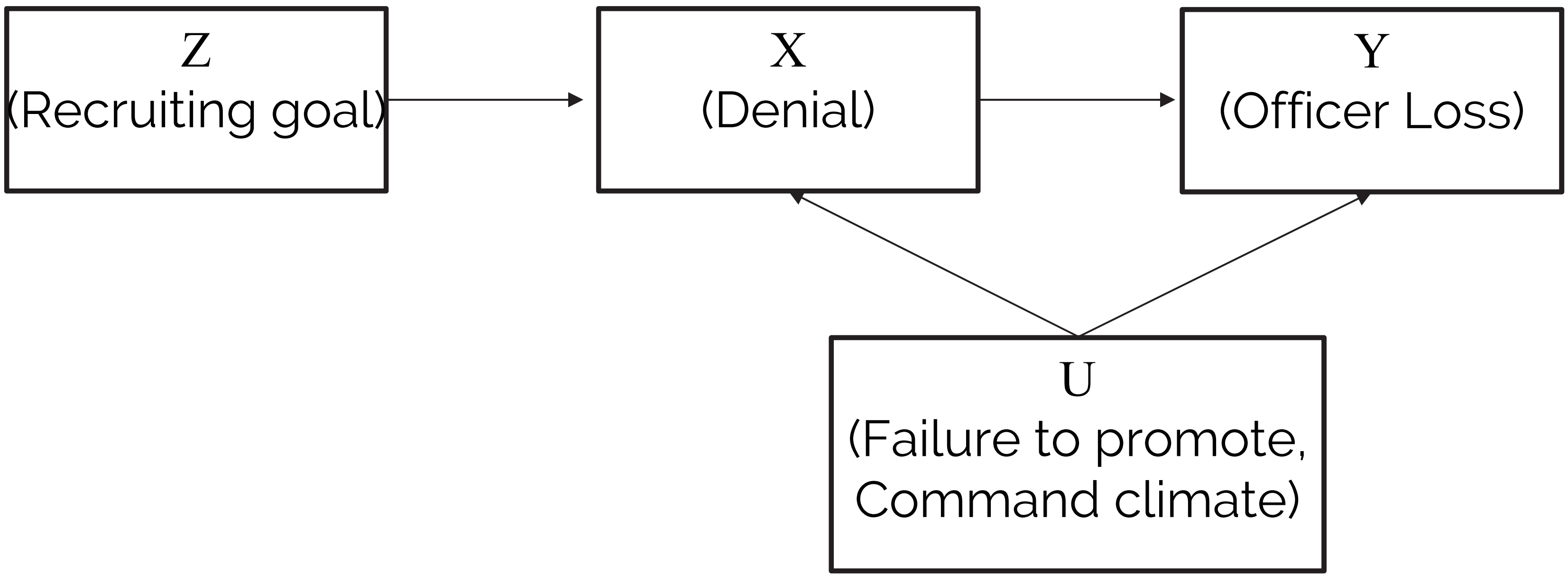
Concept introduced in 1928 by Philip G. Wright in a book called *The Tariff on Animal and Vegetable Oils*



Instrumental variables



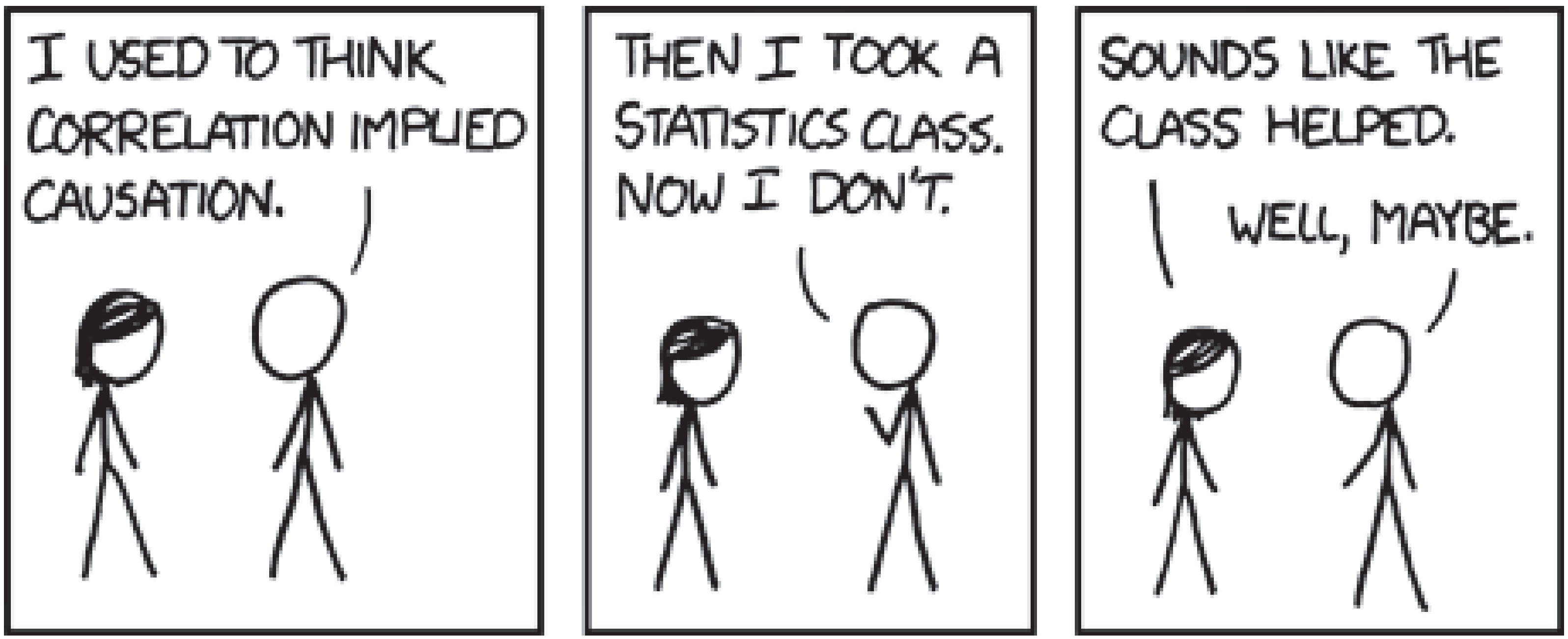
Instrumental variables



Instrumental variables

- Basic idea: in $y_i = \beta x_i + u_i$, x_i are correlated with u_i
- To estimate β, we can use IV z_i and two stage least squares regression to replace x_i with $\hat{x_i}$ that are correlated with x_i but not with u_i
  - First, regress X on Z
    - Predicted values from this regression are $\hat{x_i}$
  - Then regress Y on $\hat{X}$
    - Resulting estimates of β are consistent
    - Can also be interpreted as a Generalized Method of Moments estimator

Conclusion


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.