Introduction to Statistical Methods
SOC-GA 2332

Lecture 8: Causality in Social Sciences

Siwei Cheng

NEW YORK UNIVERSITY

Independence and Association

- Consider two random variables, X and Y. How shall we describe the relationship between the two variables?
- We say that X and Y are **independent** if knowing the values on X does not add information about the distribution of Y. We write independence as: $X \perp\!\!\!\perp Y$.
- For example, consider a sample of families with two children. Let X denote the gender of the first child, and let Y denote the gender of the second child. Does knowing the gender of the first child tell us anything more about the gender of the second child?

Independence and Association

- Relationship between joint distribution, marginal distribution, and independence:
- Statistically speaking, two distributions X and Y are independent if and only if the **joint distribution**, $f(X, Y)$, is the same as the product of the two **marginal distributions** (i.e. $g(x) \times h(y)$, where $g(x)$ and $h(y)$ are marginal distributions of X and Y respectively).

- Relationship between conditional distribution and independence:
- Random variables X, Y are independent if and only if the *conditional distribution* of Y given X is, for all possible realizations of X, equal to the **unconditional distribution** of Y.

Independence and Association

- For example, consider a sample of journal articles. Let X=author's last name initial and Y=number of citations. What does $X \perp\!\!\!\perp Y$ imply?

- If author's last name initial is *not* associated with number of citations, then we expect to see identical distributions of number of citations among authors with different last name initials.

Independence and Association

- If two variables are not independent, then we say that they are associated.

- When two variables are associated, knowing the value of one variable can give us some information on the distribution of the other variable.

- e.g. If we “hypothesize” that last names starting with ABC’s get more citations than last names with XYZ’s, what is the implication for the distribution of citations by last name initials?

Two Variables May be Conditionally Independent

- Two variables can be **associated** but also **conditionally independent** - that is, there is no association between X and Y within the same Z value.
- For example, father's occupational prestige is usually associated with son's occupational prestige. But maybe such father-son association will fade away if the son has obtained a college degree?
- In fact, this is a finding shown in Hout (1988). The study suggests that the log-odds (an indicator of "likelihood") on a higher-status destination versus a lower-status one rise sharply with origin status among workers who lack a college degree. Yet, the log-odds on high status are uniformly high for college graduates of all origin statuses.

Two Variables May be Conditionally Independent

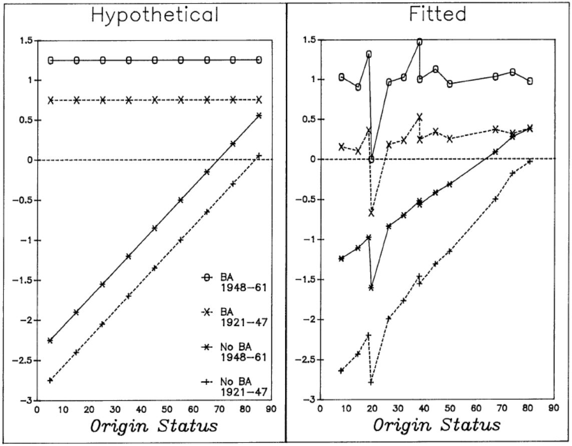
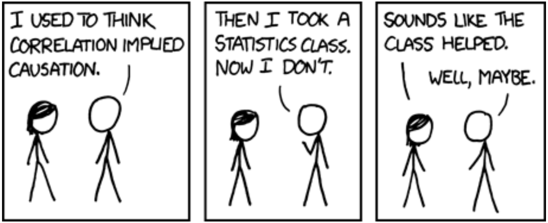


FIG. 2.—Log-odds on salaried professional destination relative to operative (nonmanufacturing) or service work expected under the SAT model by college degree and cohort for persons in the labor force, United States.

Causality under the Potential Outcome Framework

- Now, let's move from association to **causality**.
- Perhaps the most important take-home message is that: Association does not necessarily imply causality.


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

Causality under the Potential Outcome Framework

- Defining causality can be a very complicated philosophical problem.
- Generally, we think of “causality” in a pretty natural way: we say that X causes Y if a change in X will result in a change in Y.
- Some statisticians tend to think of “causality” or “causation” in an experimental sense. This is reflected in a common motto: “ No Causation without manipulation.” (Holland 1986).

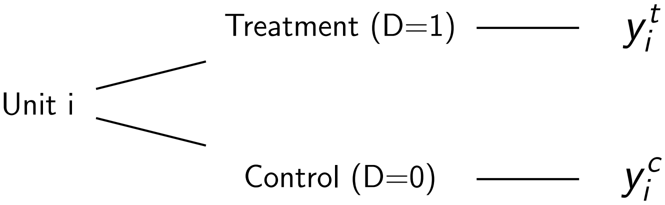
Causality under the Potential Outcome Framework

- In keeping with the current and long-standing literature in causal inference in social sciences, we are going to adopt a **counterfactual** (or: **potential outcome**) framework to study causality.
- In today's lecture, we are going to focus on the example of a **binary cause**. e.g. college attendance, party membership, state-level Medicaid expansion...
- The binary cause is defined by a treatment indicator D.
- We define D=1 if the unit is in the treatment group, and D=0 if the unit is in the control group.

Potential Outcomes on the Individual Level

Potential Outcomes on the Individual Level

- Let Y denote the outcome variable, and let y_i denotes the outcome for unit i.
- y_i^t and y_i^c are **"potential outcomes" for unit i**. y_i^t is the potential outcome for unit i in the treatment state, y_i^c is the potential outcome for unit i in the control state. See below:



Treatment (D=1) —— y_i^t

Unit i

Control (D=0) —— y_i^c

- If the unit receives the treatment, then we observe $y_i = y_i^t$. If the unit does not receive the treatment (i.e. in the control group), then we observe $y_i = y_i^c$.

Potential Outcomes on the Individual Level

- The **treatment effect** for this individual is defined as:

$$\delta_i = y_i^t - y_i^c \tag{1}$$

- But we also know that there might be some randomness in the realized value of Y for each individual unit. In other words, for each individual, there is a *distribution* of potential outcomes under treatment and control states.
- Hence, we can extend the definition in Equation (1) to:

$$\delta_i = E(y_i^t) - E(y_i^c) \tag{2}$$

Potential Outcomes on the Individual Level

- But do we observe the potential outcome for both treatment and control states?
- The answer is **NO**. This is because we only observe one status (D=1 or D=0) for each unit.
- If unit i received the treatment, then we observe y_i^t, but not y_i^c, and so y_i^c is the **counterfactual**.
- If unit i did not receive the treatment, then we observe y_i^c, but not y_i^t, and so y_i^t is the **counterfactual**.
- The **fundamental problem of causal inference** - We can never observe both of the potential outcomes for the same unit.

Potential Outcomes on the Population Level

- Next, let's move to the population level.
- We define two additional variables:
- Y^T denotes the population outcome variable **if we set** D=1, and Y^C denotes the population outcome variable **if we set** D=0.
- We define the *causal effect* in the population as:

$$E[\Delta] = E[Y^T] - E[Y^C] \tag{3}$$

- More precisely, Equation (3) defines the **Average Treatment Effect (ATE)**.

Potential Outcomes on the Population Level

- In the **observational** data, units are grouped into the "treatment group" (D=1) and "control group" (D=0).

- For those in the treatment group, we only observe Y^T, so Y^C is the counterfactual for this group; for those in the control group, we only observe Y^C, so Y^T is the counterfactual for this group:

Group	Y^T	Y^C
Treatment group	Observed	Counterfactual
Control group	Counterfactual	Observed

Potential Outcomes on the Population Level

- **Quick exercise**: what are the “treatment”, “outcome” and “counterfactuals” in the following example?
  1. Does Internet usage cause increase in earnings?
  2. Does access to railroads cause increase in county-level GDP?
  3. Does living in high-poverty neighborhood cause reduction in test scores?

The Problem of Confounding

- We have just established the counterfactual framework (i.e. the potential outcome perspective) for understanding the causal effect.
- Now, let's focus on the causal effect as defined earlier in Equation (3)

$$E[\Delta] = E[Y^T] - E[Y^C] \tag{4}$$

- If we would like to estimate $E[\Delta]$ from our sample, what would you do?

The Problem of Confounding

- The **naive approach** is to take the difference in the outcome variable Y between the sample units in the treatment group and sample units in the control group...
- Formally, this estimator can be written as:

$$E[Y^T|D=1] - E[Y^C|D=0] \tag{5}$$

The Problem of Confounding

- What's the potential problem with this naive approach?
- In fact, if the treatment and controls are *randomly assigned* (e.g. in an experiment or a natural experiment), then we can say that the treatment and control **do not differ systematically** from each other, except for the fact that the treatment group received the treatment (D=1) but the control group did not (D=0).
- This means that:

$$E[Y^T|D=1] = E[Y^T|D=0] = E[Y^T], \tag{6}$$

$$E[Y^C|D=0] = E[Y^C|D=1] = E[Y^C], \tag{7}$$

The Problem of Confounding

- And thus we have:

$$\underbrace{E[Y^T|D=1]-E[Y^C|D=0]}_{\text{What we use as estimator}} \quad = \quad \underbrace{E[Y^T]-E[Y^C]}_{\text{What we want to estimate}} \tag{8}$$

- In short, if treatment is randomly assigned, taking the difference in the data can give us an **unbiased** estimate of the average treatment effect.
- The real problem for causal inference occurs when the treatment and control groups are **not** randomly assigned.
- In this case, there will be **selection bias**: the extent to which the treatment and control groups differ in terms of their potential outcome under the control condition.

Selection Bias

- A little mathematical detour:
- It can be shown that:

$$
\begin{aligned}
& E[Y^T|D=1] - E[Y^C|D=0] && (9)\\
&= E[Y^T|D=1] - E[Y^C|D=1] + E[Y^C|D=1] - E[Y^C|D=0] && (10)\\
&= \underbrace{E[Y^T|D=1] - E[Y^C|D=1]}_{\text{treatment effect on the treated}} + \underbrace{E[Y^C|D=1] - E[Y^C|D=0]}_{\text{selection bias}} && (11)
\end{aligned}
$$

Selection Bias

- A little mathematical detour:
- It can be shown that:

$$
\begin{aligned}
& E[Y^T|D=1] - E[Y^C|D=0] && (9)\\
&= E[Y^T|D=1] - E[Y^C|D=1] + E[Y^C|D=1] - E[Y^C|D=0] && (10)\\
&= \underbrace{E[Y^T|D=1] - E[Y^C|D=1]}_{\text{treatment effect on the treated}} + \underbrace{E[Y^C|D=1] - E[Y^C|D=0]}_{\text{selection bias}} && (11)
\end{aligned}
$$

- In many settings, we can assume that the treatment effect on the treated is the same as the average treatment effect. (but of course this may be not true, e.g. college attendance).
- This basically tells us that, if we take the "naive" approach, our estimator of the causal effect will be "biased" because it is the treatment effect that we want to estimate plus a "selection bias."

Selection Bias

- Why will there be selection bias? i.e. Why would we expect the treatment and control group to be systematically different?
    1. Units may "self-select" into or out of the treatment group: e.g. attending college, participation in job training.
    2. Units may be "selected" into or out of the treatment group by others: e.g. selection into marriage, college admission, county's railroad access.
    3. In policy analysis, the eligibility criteria of policies may require the treated and untreated individuals to be different on some observed characteristics. e.g. dependent coverage term in Affordable Care Act requires that individuals who are newly eligible for parental insurance coverage to be under 26-years-old. Hence, those who are not covered by their parents are older than those in the treatment group.

© picture-alliance/dpa/Zoo Dortmund/Marcel Stawinoga

Exogeneous Shocks

- One way to “randomize” the assignment into treatment and control groups is to have some “exogeneous shocks” that do the randomization for us.
- For example, “natural experiments” are commonly used in empirical causal inference research.
- “Natural experiments” are scenarios in which the assignment into treatment/control group is determined by nature or by other factors outside the control of the investigators, but the process governing the exposures arguably resembles random assignment.

Example of Natural Experiments

- For example, a recent article by Torche and Shwed (2015) examines the effect of prenatal exposure to environmental stressors on birth outcomes.
- They compare births in Northern Israel (dark gray) which were in the range of Hezbollah's missiles in 2006, to the out-of-reach regions in the Center, Sharon, and Jerusalem regions (pale blue), excluding the South and the West Bank (pale gray) which were subjected to a host of armed conflict events in the study years. Areas in bright blue identify bodies of water. (Their notes to Figure 1 in the article.)

Example of Natural Experiments

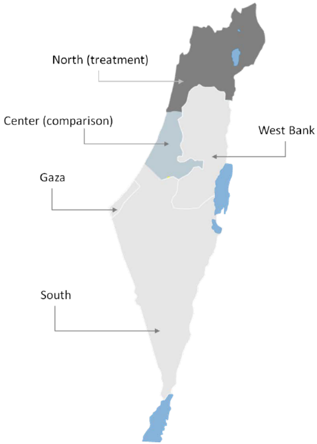
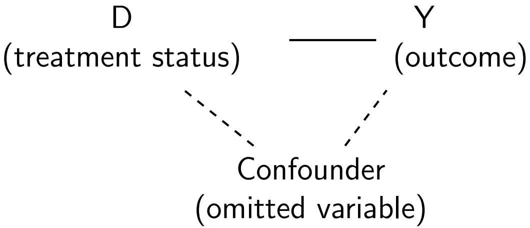


Figure 1: Map of Study Population We compare births in Northern Israel (dark gray) which were in the range of Hezbollah's missiles in 2006, to the out of reach regions in the Center, Sharon, and Jerusalem regions (pale blue), excluding the South and the West Bank (pale gray) which were subjected to a host of armed conflict events in the study years. Areas in bright blue identify bodies of water.

Selection Bias and Confounding

- Selection bias can also be viewed as the problem of **confounding**, or **omitted variable bias**. This may result in spurious association between two variables.
- The confounder induces individuals in the treatment and control group to be systematically different with regard to their potential outcome.



D
(treatment status)

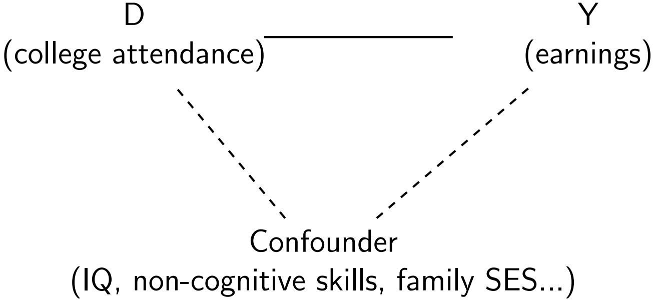
Y
(outcome)

Confounder
(omitted variable)

Selection Bias and Confounding

- If the confounder is observed, we can try to control for the confounder, so that there is no "selection bias" among units with the same value on the confounder.
- In practice, there are several ways to deal with confounders. For example, in regression, we may control for a set of covariates (potential confounders).
- For example, to account for the selection bias in college attendance (where our key research question is the causal effect of college on earnings), we may want to control for IQ, non-cognitive score, family background, ...
- See next slide...

Selection Bias and Confounding



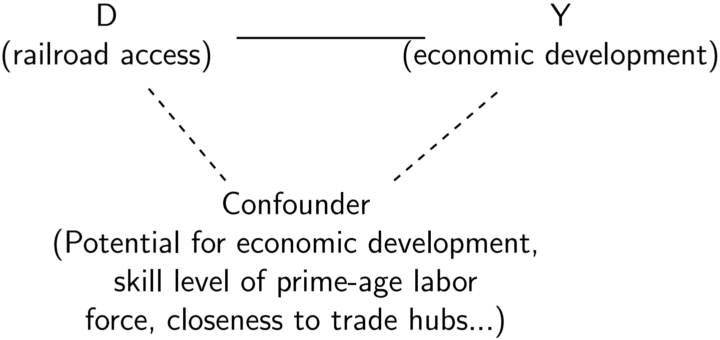
D
(college attendance)

Y
(earnings)

Confounder
(IQ, non-cognitive skills, family SES...)

Selection Bias and Confounding

Another example for railroad access and economic development in a county.



D
(railroad access)

Y
(economic development)

Confounder
(Potential for economic development,
skill level of prime-age labor
force, closeness to trade hubs...)

Matching-based methods create "**matches**" for each individual unit to other units that are very similar to this unit.

Selection Bias and Confounding

- However, in many cases, individuals may select into (or be selected into) the treatment group based on factors that are not observed by the researcher (i.e. selection on unobservables).
- One tentative solution to the problem is **before-and-after** design (e.g. fixed-effect models), which controls away individual-level time-invariant selection bias. e.g. the effect of marriage on wages; the effect of Medicaid expansion on state-level medical insurance coverage.
- Another approach is the Instrumental Variable (IV) approach, which makes use of exogeneous shocks (natural experiments) to identify the causal effect.

Omitted Variable Bias and Regression

- How does selection bias (i.e. omitted variable bias) relate to regression equations?

Omitted Variable Bias and Regression

- How does selection bias (i.e. omitted variable bias) relate to regression equations?
- Suppose the true model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{12}$$

- Now, suppose we forgot to include X_2 (X_2 becomes the omitted variable)...Then we will be estimating the following regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon. \tag{13}$$

Omitted Variable Bias and Regression

- Then, it can be shown mathematically that the expectation of our OLS estimate will be:

$$\hat{\beta}^{OLS} = \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)} \tag{14}$$

- If X_1 and X_2 are positively associated, and if X_2 and Y are also positively associated, omitting X_2 will lead to an *overstatement* of the true causal effect (i.e. β_1).

© picture-alliance/dpa/Zoo Dortmund/Marcel Stawinoga

Another Important Assumption in Causal Inference: SUTVA

- In most applications of the potential outcome (counterfactual) framework, we rely on a **Stable Unit Treatment Value Assumption (SUTVA)**.
- It basically assumes that the potential outcomes of individuals be unaffected by changes in the treatment exposures of all other individuals.

Another Important Assumption in Causal Inference: SUTVA

- Examples of scenarios in which we might want to consider the possibility that SUTVA is violated:
  1. Participation in a job training program...spill-over effects...
  2. College attendance...would college be worth the same if everyone ended up with a college degree?
  3. Effect of vaccination on the individual and community level...