150C/355C: Causal Inference for Social Science

Introduction

Jonathan Mummolo

Stanford University

Data Science Education @ Stanford Political Science

- 150A: Data Science for Politics
- 150B: Introduction to Machine Learning for Social Scientists
- 150C: Causal Inference for Social Science

What is This Class About?

- An introduction to causal inference methods in social science research

What is This Class About?

- An introduction to causal inference methods in social science research
- Methods designed to assess the impact of some potential **cause** (e.g., an intervention, a change in institutions, economic conditions, or policies) on some **outcome** (e.g., vote choice, income, election results, levels of violence)

What is This Class About?

- An introduction to causal inference methods in social science research
- Methods designed to assess the impact of some potential **cause** (e.g., an intervention, a change in institutions, economic conditions, or policies) on some **outcome** (e.g., vote choice, income, election results, levels of violence)
- We teach you the toolkit of modern causal inference methods as they are now widely used across academic fields, government, industry, and non-profits

What is This Class About?

- An introduction to causal inference methods in social science research
- Methods designed to assess the impact of some potential **cause** (e.g., an intervention, a change in institutions, economic conditions, or policies) on some **outcome** (e.g., vote choice, income, election results, levels of violence)
- We teach you the toolkit of modern causal inference methods as they are now widely used across academic fields, government, industry, and non-profits
- Causal inference skills are useful to make better decisions and valuable on the job market

What is This Class About?

- We learn all three: research design, statistics, and data analysis

What is This Class About?

- We learn all three: research design, statistics, and data analysis
- The topics include experiments, matching, regression, sensitivity analysis, difference-in-differences, panel methods, instrumental variable estimation, and regression discontinuity designs.

What is This Class About?

- We learn all three: research design, statistics, and data analysis
- The topics include experiments, matching, regression, sensitivity analysis, difference-in-differences, panel methods, instrumental variable estimation, and regression discontinuity designs.
- Applications are drawn from various fields including political science, public policy, business, health, economics, and sociology

Causal Inference

- Statistics can be used for many purposes:
  - Description

Causal Inference

- Statistics can be used for many purposes:
    - Description
    - Prediction

Causal Inference

- Statistics can be used for many purposes:
  - Description
  - Prediction
  - Causal inference
    - Relatively new subfield within statistics
    - Highly interdisciplinary, rapidly expanding

How Can we Draw Causal Inference?

- Anecdotes, Intuition, and Theory

How Can we Draw Causal Inference?

- Anecdotes, Intuition, and Theory
- Correlations

How Can we Draw Causal Inference?

- Anecdotes, Intuition, and Theory
- Correlations
- Regressions

How Can we Draw Causal Inference?

- Anecdotes, Intuition, and Theory
- Correlations
- Regressions

These methods are all severely prone to error. Causal inference is a hard problem and invalid causal reasoning is one of the most common errors in human judgment, news reporting, and scientific studies!

Anecdotes

“My grandmother Annie smoked two packs a day and lived until she was 95 years old.”

Anecdotes

“My grandmother Annie smoked two packs a day and lived until she was 95 years old.”

- For every anecdote you know, there might be many that you do not know that show the opposite pattern

Anecdotes

“My grandmother Annie smoked two packs a day and lived until she was 95 years old.”

- For every anecdote you know, there might be many that you do not know that show the opposite pattern
- We often only raise those anecdotes that we like to see to justify actions or behaviors

Anecdotes

“My grandmother Annie smoked two packs a day and lived until she was 95 years old.”

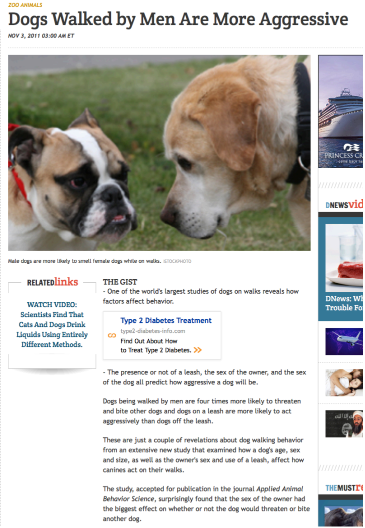
- For every anecdote you know, there might be many that you do not know that show the opposite pattern
- We often only raise those anecdotes that we like to see to justify actions or behaviors
- All that the anecdotes suggests is that Annie was prone to have a long life

Anecdotes

“My grandmother Annie smoked two packs a day and lived until she was 95 years old.”

- For every anecdote you know, there might be many that you do not know that show the opposite pattern
- We often only raise those anecdotes that we like to see to justify actions or behaviors
- All that the anecdotes suggests is that Annie was prone to have a long life
- The key question for causal inference is about the unobserved counterfactual: how long would Annie have lived had she never smoked a single cigarette?

Correlations



ZOO ANIMALS

Dogs Walked by Men Are More Aggressive

NOV 3, 2011 03:00 AM ET

Male dogs are more likely to smell female dogs while on walks. ISTOCKPHOTO

RELATEDlinks

WATCH VIDEO: Scientists Find That Cats And Dogs Drink Liquids Using Entirely Different Methods.

THE GIST

- One of the world's largest studies of dogs on walks reveals how factors affect behavior.

- The presence or not of a leash, the sex of the owner, and the sex of the dog all predict how aggressive a dog will be.

Dogs being walked by men are four times more likely to threaten and bite other dogs and dogs on a leash are more likely to act aggressively than dogs off the leash.

These are just a couple of revelations about dog walking behavior from an extensive new study that examined how a dog's age, sex and size, as well as the owner's sex and use of a leash, affect how canines act on their walks.

The study, accepted for publication in the journal *Applied Animal Behavior Science*, surprisingly found that the sex of the owner had the biggest effect on whether or not the dog would threaten or bite another dog.

Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)

Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:

Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:
  - It's not that men make dogs more aggressive, but men might simply prefer more aggressive dogs.

Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:
  - It's not that men make dogs more aggressive, but men might simply prefer more aggressive dogs.
  - Basketball players are tall, but does playing basketball make you taller?

Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:
  - It's not that men make dogs more aggressive, but men might simply prefer more aggressive dogs.
  - Basketball players are tall, but does playing basketball make you taller?
- Correlations are often driven by confounding factors: ice cream sales are correlated with murder rates throughout a typical year. Does not mean ice cream causes murders. Confounding factor: weather.

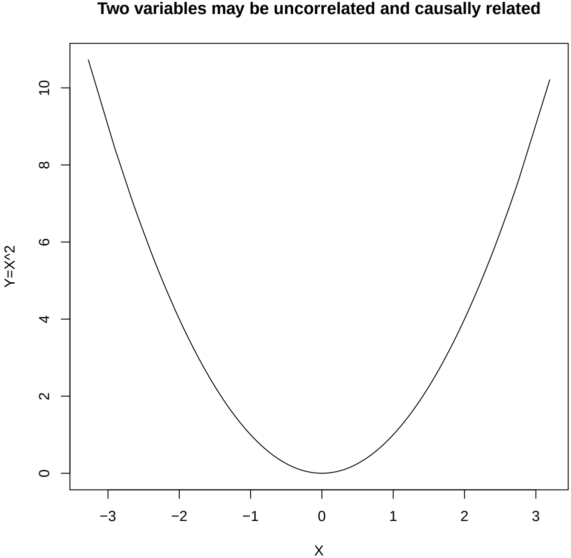
Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:
  - It's not that men make dogs more aggressive, but men might simply prefer more aggressive dogs.
  - Basketball players are tall, but does playing basketball make you taller?
- Correlations are often driven by confounding factors: ice cream sales are correlated with murder rates throughout a typical year. Does not mean ice cream causes murders. Confounding factor: weather.
- Correlations are neither a necessary nor sufficient condition for causality.

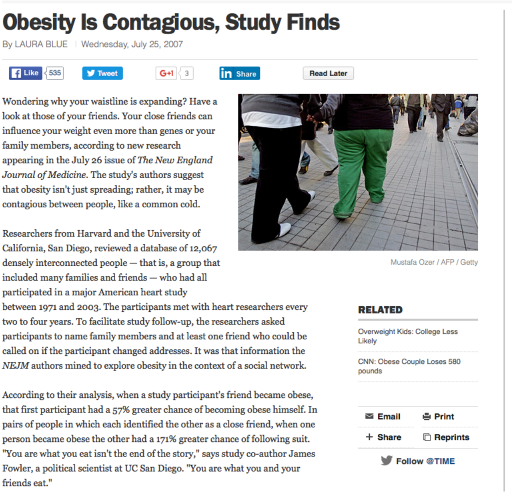
Correlations

- The problem with correlations for causal inference is that they often arise for reasons that have nothing to with the causal process under investigation (spurious correlation)
- Correlations are often driven by selection effects:
  - It's not that men make dogs more aggressive, but men might simply prefer more aggressive dogs.
  - Basketball players are tall, but does playing basketball make you taller?
- Correlations are often driven by confounding factors: ice cream sales are correlated with murder rates throughout a typical year. Does not mean ice cream causes murders. Confounding factor: weather.
- Correlations are neither a necessary nor sufficient condition for causality. Why unnecessary?

Correlations



Regressions



Obesity Is Contagious, Study Finds

By LAURA BLUE | Wednesday, July 25, 2007

Wondering why your waistline is expanding? Have a look at those of your friends. Your close friends can influence your weight even more than genes or your family members, according to new research appearing in the July 26 issue of *The New England Journal of Medicine*. The study's authors suggest that obesity isn't just spreading; rather, it may be contagious between people, like a common cold.

Mustafa Ozer / AFP / Getty

Researchers from Harvard and the University of California, San Diego, reviewed a database of 12,067 densely interconnected people — that is, a group that included many families and friends — who had all participated in a major American heart study between 1971 and 2003. The participants met with heart researchers every two to four years. To facilitate study follow-up, the researchers asked participants to name family members and at least one friend who could be called on if the participant changed addresses. It was that information the *NEJM* authors mined to explore obesity in the context of a social network.

According to their analysis, when a study participant's friend became obese, that first participant had a 57% greater chance of becoming obese himself. In pairs of people in which each identified the other as a close friend, when one person became obese the other had a 171% greater chance of following suit. "You are what you eat isn't the end of the story," says study co-author James Fowler, a political scientist at UC San Diego. "You are what you and your friends eat."

RELATED

Overweight Kids: College Less Likely

CNN: Obese Couple Loses 580 pounds

Regressions

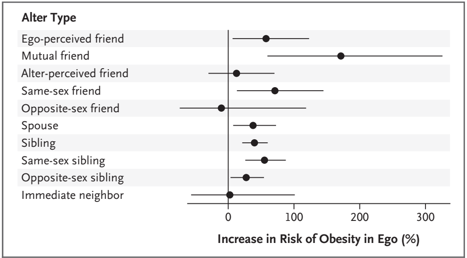


Figure 4. Probability That an Ego Will Become Obese According to the Type of Relationship with an Alter Who May Become Obese in Several Subgroups of the Social Network of the Framingham Heart Study.

The closeness of friendship is relevant to the spread of obesity. Persons in closer, mutual friendships have more of an effect on each other than persons in other types of friendships. The dependent variable in each model is the obesity of the ego. Independent variables include a time-lagged measurement of the ego's obesity; the obesity of the alter; a time-lagged measurement of the alter's obesity; the ego's age, sex, and level of education; and indicator variables (fixed effects) for each examination. Full models and equations are available in the Supplementary Appendix. Mean effect sizes

Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**

Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**
    - The list of all potential confounding factors is a bottomless pit.

Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**
  - The list of all potential confounding factors is a bottomless pit.
  - *How* to properly control for confounders is often up for debate / unknown.

Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**
  - The list of all potential confounding factors is a bottomless pit.
  - *How* to properly control for confounders is often up for debate / unknown.
- People whose friends tend to be obese might differ in many ways from those whose friends are not obese:

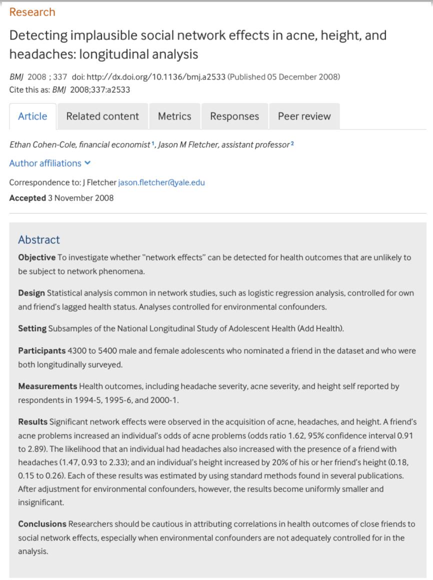
Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**
  - The list of all potential confounding factors is a bottomless pit.
  - *How* to properly control for confounders is often up for debate / unknown.
- People whose friends tend to be obese might differ in many ways from those whose friends are not obese:
  - They might be poorer economically, live in areas with easier access to healthier food, have less access to sports, different hobbies, eating habits, etc.

Regressions

- Regressions are simply refined correlations that try to control for other confounding factors. **Problems:**
  - The list of all potential confounding factors is a bottomless pit.
  - *How* to properly control for confounders is often up for debate / unknown.
- People whose friends tend to be obese might differ in many ways from those whose friends are not obese:
  - They might be poorer economically, live in areas with easier access to healthier food, have less access to sports, different hobbies, eating habits, etc.
- For causal inference we need to ask: among people who are identical in all respects, does making friends with obese persons really make them more likely to become obese?

Regressions



Research

Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis

BMJ 2008 ; 337 doi: http://dx.doi.org/10.1136/bmj.a2533 (Published 05 December 2008)
Cite this as: *BMJ* 2008;337:a2533

Ethan Cohen-Cole, financial economist[1], *Jason M Fletcher, assistant professor*[2]

Correspondence to: J Fletcher jason.fletcher@yale.edu

Accepted 3 November 2008

Abstract

Objective To investigate whether "network effects" can be detected for health outcomes that are unlikely to be subject to network phenomena.

Design Statistical analysis common in network studies, such as logistic regression analysis, controlled for own and friend's lagged health status. Analyses controlled for environmental confounders.

Setting Subsamples of the National Longitudinal Study of Adolescent Health (Add Health).

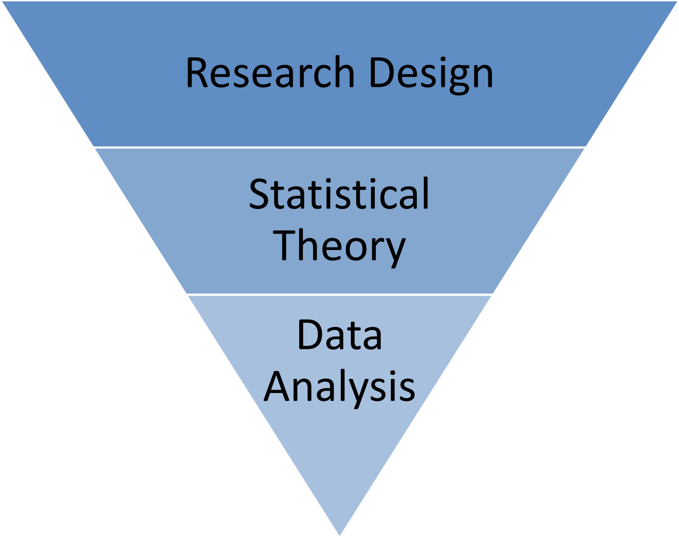
Participants 4300 to 5400 male and female adolescents who nominated a friend in the dataset and who were both longitudinally surveyed.

Measurements Health outcomes, including headache severity, acne severity, and height self reported by respondents in 1994-5, 1995-6, and 2000-1.

Results Significant network effects were observed in the acquisition of acne, headaches, and height. A friend's acne problems increased an individual's odds of acne problems (odds ratio 1.62, 95% confidence interval 0.91 to 2.89). The likelihood that an individual had headaches also increased with the presence of a friend with headaches (1.47, 0.93 to 2.33); and an individual's height increased by 20% of his or her friend's height (0.18, 0.15 to 0.26). Each of these results was estimated by using standard methods found in several publications. After adjustment for environmental confounders, however, the results become uniformly smaller and insignificant.

Conclusions Researchers should be cautious in attributing correlations in health outcomes of close friends to social network effects, especially when environmental confounders are not adequately controlled for in the analysis.

How can we conduct causal inference?

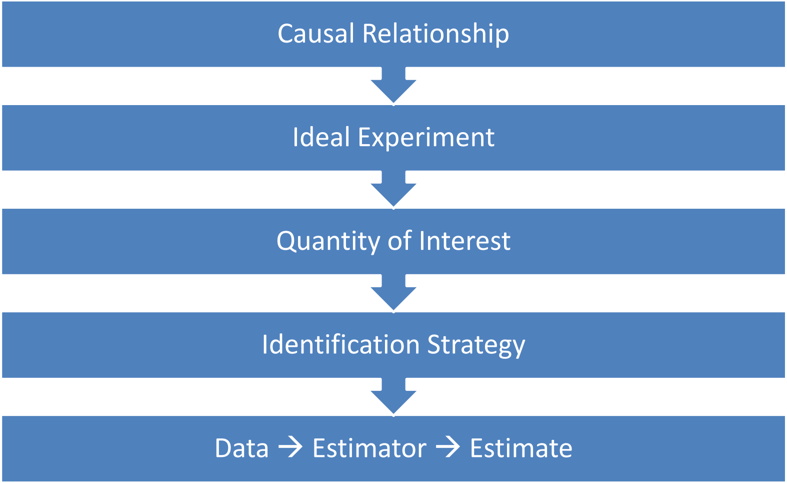


Research Design

Statistical Theory

Data Analysis

Causal Inference Workflow



Causal Relationship

Ideal Experiment

Quantity of Interest

Identification Strategy

Data → Estimator → Estimate

Roadmap for the Course

- Potential Outcomes Model

Roadmap for the Course

- Potential Outcomes Model
- Random Assignment
  - Design and Analysis of Experiments

Roadmap for the Course

- Potential Outcomes Model
- Random Assignment
  - Design and Analysis of Experiments
- Selection on Observables
  - Matching, Regression
  - Sensitivity Analyses

Roadmap for the Course

- Potential Outcomes Model
- Random Assignment
  - Design and Analysis of Experiments
- Selection on Observables
  - Matching, Regression
  - Sensitivity Analyses
- Selection on Unobservables
  - Longitudinal Research Designs: Difference-in-Differences, Panel Methods, and related methods
  - Cross-Sectional Designs: Instrumental Variables, Regression Discontinuity Design

Prerequisites

- This course assumes a undergraduate level knowledge of linear regression, probability, and statistical computing in `R` as covered in the PS 150A and B.
- A willingness to work hard on possibly unfamiliar material
  - Pride is the enemy of learning.

Prerequisites

- This course assumes a undergraduate level knowledge of linear regression, probability, and statistical computing in `R` as covered in the PS 150A and B.
- A willingness to work hard on possibly unfamiliar material
  - Pride is the enemy of learning.
  - Ask questions! Use Piazza! Come to office hours!

Prerequisites

- This course assumes a undergraduate level knowledge of linear regression, probability, and statistical computing in `R` as covered in the PS 150A and B.
- A willingness to work hard on possibly unfamiliar material
  - Pride is the enemy of learning.
  - Ask questions! Use Piazza! Come to office hours!
- A correlate of whether you might have the background:
  - `lm(Y ~ X1+X2,data=dataset)`
  - $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Prerequisites

- This course assumes a undergraduate level knowledge of linear regression, probability, and statistical computing in `R` as covered in the PS 150A and B.
- A willingness to work hard on possibly unfamiliar material
  - Pride is the enemy of learning.
  - Ask questions! Use Piazza! Come to office hours!
- A correlate of whether you might have the background:
  - `lm(Y ~ X1+X2,data=dataset)`
  - $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
  - $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$

Requirements

- Weekly readings
  - Read slow, some material should be read multiple times, and do not skip equations. All of this material can show up on a test.

Requirements

- Weekly readings
    - Read slow, some material should be read multiple times, and do not skip equations. All of this material can show up on a test.
- Bi-weekly homework assignments (32 % of the final grade)
    - Posted on Wed. ; due following Wed. before class.
    - Can work in groups, but attempt solo first.
    - Provide your own printed write-up and submit code files.

Requirements

- Weekly readings
  - Read slow, some material should be read multiple times, and do not skip equations. All of this material can show up on a test.
- Bi-weekly homework assignments (32 % of the final grade)
  - Posted on Wed. ; due following Wed. before class.
  - Can work in groups, but attempt solo first.
  - Provide your own printed write-up and submit code files.
- Midterm (32% of the final grade). May 4 in class (tentative date).
- Final exam (32% of the final grade).

Requirements

- Weekly readings
  - Read slow, some material should be read multiple times, and do not skip equations. All of this material can show up on a test.
- Bi-weekly homework assignments (32 % of the final grade)
  - Posted on Wed. ; due following Wed. before class.
  - Can work in groups, but attempt solo first.
  - Provide your own printed write-up and submit code files.
- Midterm (32% of the final grade). May 4 in class (tentative date).
- Final exam (32% of the final grade).
- Class participation (4% of the final grade)
  - “Causal Claim of the Week”: identify causal claim made in the news; summarize evidence provided; what would be ideal experiment?

Requirements

- Weekly readings
  - Read slow, some material should be read multiple times, and do not skip equations. All of this material can show up on a test.
- Bi-weekly homework assignments (32 % of the final grade)
  - Posted on Wed. ; due following Wed. before class.
  - Can work in groups, but attempt solo first.
  - Provide your own printed write-up and submit code files.
- Midterm (32% of the final grade). May 4 in class (tentative date).
- Final exam (32% of the final grade).
- Class participation (4% of the final grade)
  - “Causal Claim of the Week”: identify causal claim made in the news; summarize evidence provided; what would be ideal experiment?
  - One-page (double-spaced), discuss for a few mins. at start of class

Housekeeping

- Weekly recitations:
  Friday, 1:30 PM - 3:20 PM at Encina Hall 464
  - Material will mostly be review of lecture, but anything in section can be on a test. (Attendance is strongly encouraged!)
- Piazza course website will have slides, homework, data sets, and some additional readings:
  `https://piazza.com/stanford/spring2016/150c355c/home`
- You can sign up on the Piazza course page directly from the above address. There are also free Piazza apps for mobile devices.
- Use OHs and Piazza to ask questions about the course and homework.
- Office hours:
  - Jonathan: Thursday 2-4 pm and by appointment
  - Matt: Wed. 3:30pm-5pm and by appointment

Readings

- Books:
  - Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mastering Metrics*. Princeton University Press.
  - Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments*. W. W. Norton.
  - Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Some assigned articles
  - Will be posted on course website

Stanford's Notes on Academic Integrity

Students are held accountable for adhering to established community standards including the Fundamental Standard and the Honor Code

- Fundamental Standard:
  - Students at Stanford are expected to show both within and without the University such respect for order, morality, personal honor and the rights of others as is demanded of good citizens. Failure to do this will be sufficient cause for removal from the University.
  - Please review at: `https://communitystandards.stanford.edu/student-conduct-process/honor-code-and-fundamental-standard`
- Examples of violations of the Honor Code include:
  - Copying from another's examination paper, unpermitted collaboration, plagiarism, giving or receiving unpermitted aid on a take-home examination, etc.
  - Please review at: `https://communitystandards.stanford.edu/student-conduct-process/honor-code-and-fundamental-standard`