

# Senior Thesis

What makes research 'good'?

---

Kyle Coombs

Bates College | EC/DCS 456

# Table of Contents

- Prologue
  - Caveat
- Mostly Harmless FAQs
  1. What is causal relationship?
  2. What experiment?
  3. What is the identification strategy
  4. What is your mode of statistical inference?

# Prologue

# Prologue

- Today we're going to go over how to ask a good research question
- This is tough!
- People make careers out of this (I'm trying)
- Rest assured there are infinite interesting research questions
- But there also infinite uninteresting research questions

# Caveat

- I have no idea how to guarantee a research question is "good"
- **Spoiler:** No one does
- Sometimes the most interesting question (at first blush) turns out to be a dud
- Sometimes the least interesting question (at first blush) turns out to be a gem
- I am going to give you some tips and tricks the experts use

# Mostly Harmless FAQs

# Mostly Harmless Econometrics and FAQs

- I'm going to use the book *Mostly Harmless Econometrics* by Angrist and Pischke as a guide
- This is a great book and I highly recommend it
- It's also a great title (a play on the Hitchhiker's Guide to the Galaxy series)
- Chapter 1 details four FAQs that are useful to ask when thinking about an empirical research question
- The focus is on causal research, but the questions are useful for any empirical research

# What is causal relationship?

What is the causal relationship of interest?

- Causal relationships are super useful for making predictions about changing circumstances or policies
- Well-answered causal questions can tell us what happens in "counterfactual" worlds
- It is important to frame the core causal relationship of interest because sometimes empirical work gets hyperfocused on a specific setting
- The core causal relationship is the relationship of interest in the absence of any specific setting
- How do you generalize your results to other settings? Why do we care?
- **Policy interest:** Does your question have implications for policy decisions?
- **Theoretical interest:** Does your question derive from an economic model?



# Causality and charity

- As part of my research on charitable giving, I have studied the effect of increased government support on charitable giving
- e.g. By how much does private charitable giving change if the government increases its support of a cause by \$1?

## Policy interest

This effect is useful for evaluating whether the government should increase its support for public goods (e.g. public parks), public services (e.g. public education), or public programs (e.g. food stamps)

## Theoretical Interest

Simple models of public good provision predict 1:1 crowd-out, but more complex models predict less than 1:1 crowd-out

# Discuss with a partner

- Talk with a partner about the key causal relationship in your thesis for three minutes
- What is the bigger picture?
- Is it theoretically relevant?
- Is there a policy implication?

# What experiment?

- The second research FAQ is:

What experiment could *ideally* be used to capture the causal relationship of interest?

- Get creative!
- Imagine a world without budget constraints, ethical concerns, or time constraints
- How could you answer your question of interest?
- Often you can't logistically perform this ideal experiment
  - And in some cases, you definitely should not even consider performing this ideal experiment (e.g. what is the effect of pandemics on economic growth?)
- But the ideal experiment helps you think through your precise causal question
- It can also help you back out
  - The assumptions you need to make to answer your question
  - The unit of analysis to answer your question

# Ideal experiment and charity

- What would be the ideal experiment to look at the effect of government support on charitable giving?
- We would randomly provide government support to some charities and not others
- What if we are interested in the effect of government support on informal support systems (e.g. neighbors helping neighbors)?
- Would we randomly provide support to communities? Individuals? Both levels?
- Do we need to target support based on personal wealth? Number of friends? Network centrality?!
- As you can see, this gets complicated quickly!

# Discuss with a partner

- Discuss with the same partner what your ideal experiment would be for three minutes
- What would you do if you had infinite resources?
- What would you do if you had infinite time?
- What would you do if you had infinite data?

# Are you FUQ'd?

- Sometimes you can only think of truly impossible experiments
- Then you have a **F**undamentally **U**nidentified **Q**uestion (FUQ)
- I don't mean "impossible" in the sense that you can't do it because of budget constraints, ethical concerns, or time constraints
- I mean "impossible" in the sense that you can't do it because it is logically impossible
  - e.g. What is the effect of starting school at an older age?

# Older starting age

- Imagine an ideal experiment to measure the effect of starting school at an older age
- We could randomly assign kids to start at age 5 (group  $C$ ) or age 6 (group  $T$ )
- Then we could measure 1st grade test scores and compare them
  - But the  $C$  kids are a year younger than  $T$  kids so we would expect them to do worse on the test regardless of randomization
- Can we compare  $C$ 's 2nd grade scores to  $T$ 's' 1st grade scores? They're the same age...
  - But the  $C$  kids have had an extra year of schooling so we would expect them to do better on the test regardless of randomization
- This question is FUQ'd: We cannot disentangle the start-age effect due to maturation from the time-in-school effects due to schooling
  - They're linearly dependent while still in school
- What can we do?
- Look at wages as adults?

# Anyone have a FUQ to give?

- Talk with a partner about your research question for three minutes
- Are you worried that it is FUQ'd?
- If so, can you amend it to make it answerable?



# What is the identification strategy?

- The third research FAQ is:

What is the identification strategy?

- While the first two questions are bigger picture, this question dives into the weeds
- An *identification strategy* is the way you use observational data to approximate a real experiment
- Any *identification strategy* requires that you make some assumptions

# Common identification strategies

1. **Control variables:** Includes that control for any confounding variables between a treated and control group (generally hard to truly control for everything)
2. **Difference-in-difference:** Compare units before and after a treatment to units that are never treated
3. **Staggered difference-in-difference:** Compare units before and after a treatment to units that are treated later
4. **Instrumental variables:** Use a variable that is correlated with the treatment (relevant), but otherwise
5. **Regression discontinuity design:** Compare units just above and just below a threshold

# Identification strategy and charitable

giving

- In the case of private charitable giving, we can't randomly assign government support to charities
- But we know that government support for public parks increases by a fixed amount (e.g. \$1 million) when the population of a city passes a certain threshold (e.g. 50,000 people)
- We can use this "natural experiment" to identify the effect of government support on charitable giving
  - Option 1: Regression discontinuity design of towns on either side of the threshold
  - Option 2: Difference-in-differences design of towns that pass the threshold vs. towns that don't pass the threshold
- **Note:** This requires assumptions that the towns on either side of the threshold are similar in ways that affect charitable giving

# Discuss with partner

- What identification strategy could work in your thesis?
- What assumptions do you need to make?

# What is your mode of statistical inference?

- The fourth research FAQ is:
  - What is your mode of statistical inference?
- What is the population you are trying to make inferences about?
- How is the sample constructed and used?
- What are you assuming about the standard errors of your estimates?

# Practical terms

- Basically, this is a question about how you construct standard errors
  - e.g. Do you think your observations are independent?
  - e.g. Do you think your data exhibit heteroskedasticity?
  - These two would violate the standard assumptions of vanilla econometric methods
- Standard errors and inference can get tricky, technical, and kind of boring, but they are important to think about
- Why? Statistical inference is how we make statements about the population based on a sample
- It is how you move from a neat well-conceived project to a generalizable result that matters outside of your specific setting

# Statistical inference and charitable

giving

- Recall the example of a population threshold for increased government support of public parks
- Should we treat charitable giving of individuals in towns that pass the threshold as independent observations?
- Or should we treat charitable giving of individuals in towns that pass the threshold as dependent observations?

# Discuss statistical inference

- This is a little tricky to chat with your partner about
- What is your population of interest?
- What sample do you expect to use?
  - Is it randomly sampled?
  - Selected on some specific characteristics?
- How do you plan to construct standard errors?
  - Do you think your observations are independent?
  - Do you think your data exhibit heteroskedasticity?



Next classes: One-on-ones and Student pitches

---