SOC6707 Intermediate Data Analysis

Monica Alexander

Week 11: Research Methods and Design

Notes

- Research project analysis
- A3
- Presentations next week

Reading

Neuman, 2010. Social Research Methods: Qualitative and Quantitative Approaches.

Steps in the (quantitative) research process

1. Select a topic
   - general area of study or interest
2. Focus the question
   - a topic is too broad
   - review literature
   - develop hypotheses based on theory
   - think about operationazability
3. Design the study
   - type of sample
   - what the measure/ how to measure
   - what research technique to employ

Steps in the research process

4. Collect data
   - record
   - make useable (readable by computer)
5. Analyze the data
   - tables, graphs, statistical summaries
   - statistical methods
6. Interpret the data
   - including limitations
7. Inform others
   - know your audience

Types of studies

Within or across cases?

- Studies vary according to the number of cases we examine and the depth-intensity of investigation into features of the cases.
- A case is a unit or observation
- An individual person can be a case as can a family, company, or entire nation

Does a study primarily focus on features within cases or across cases?

Case-study research

- Case-study research examines many features of a few cases.
- The cases can be individuals, groups, organizations, movements, events, or geographic units.
- The data on the case are detailed, varied, and extensive.
- It can focus on a single point in time or a duration of time.
- Most case-study research is qualitative, but it does not have to be.

Cross-case research

- By contrast, almost all cross-case (or noncase research) is quantitative.
- Rather than carry out a detailed investigation of each case, across-case research compares select features across numerous cases.
- It treats each case as the carrier of the feature of interest.

While certain issues lend themselves to one or another approach, it is sometimes possible to study the same issue using a case study and an across-case research design.

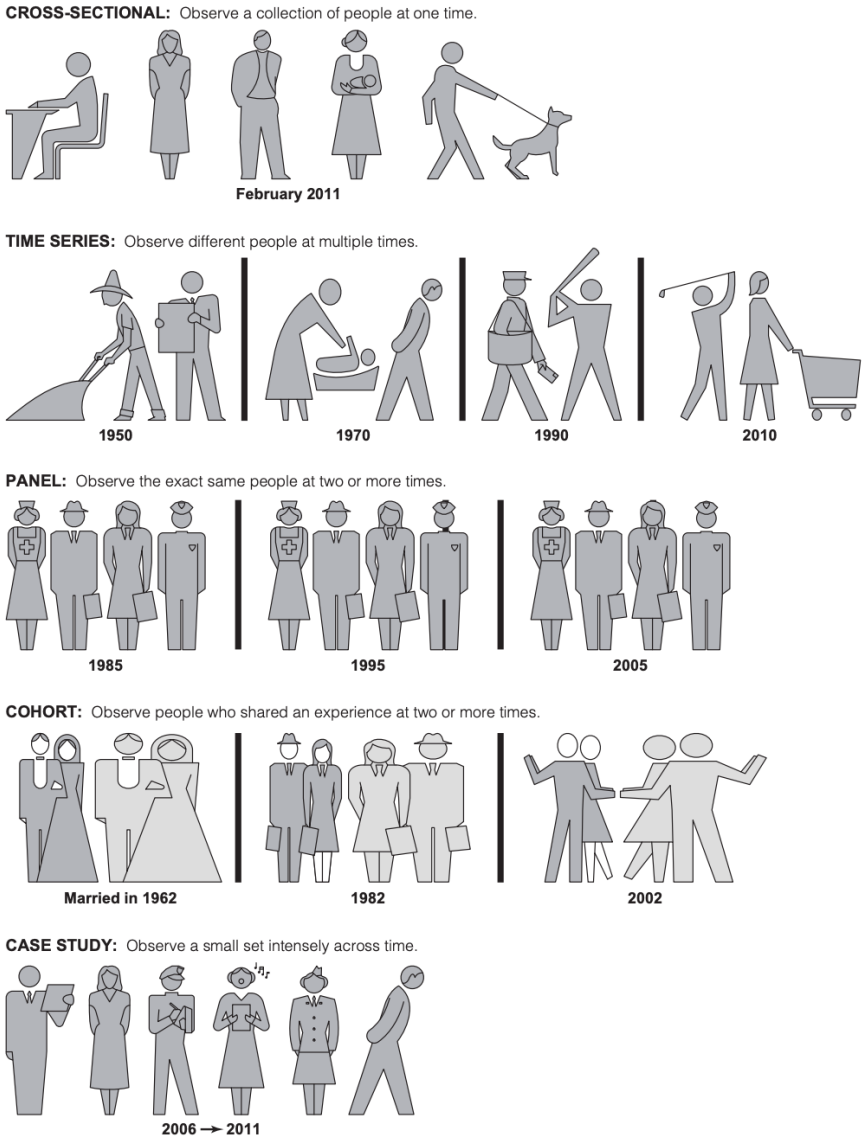
- e.g. studying why a family decides to move to Toronto

Single or multiple points in time?

- **Cross-sectional research** gathers data at one time point; a kind of 'snapshot'
  - often simplest and least costly
- **Longitudinal research** gathers data at multiple time points; a kind of 'moving picture'

Deciding whether a study is cross-sectional or longitudinal may not always be simple. Data collection takes time. But time may not be considered as a main dimension of analysis.

Types of longitudinal data



Types of quantitative data collection

- Experiments
- Surveys
- Nonreactive research
  - content analysis
  - existing data

Linking theory to empirical realities

Deductive versus inductive theorizing

In an ideal sense, you can approach the building and testing of theory from two directions:

- begin with abstract thinking and then logically connect the ideas in theory to concrete evidence or
- begin with specific observations of empirical evidence and then generalize from the evidence to build toward increasingly abstract ideas.

Deductive reasoning

- To theorize in a deductive direction, we start with abstract concepts or a theoretical proposition that outlines the logical connection among concepts.
- We move next to evaluate the concepts and propositions against concrete evidence.
- We go from ideas, theory, or a mental picture toward observable empirical evidence.

Inductive reasoning

- To theorize in an inductive direction, we begin with observing the empirical world and then reflecting on what is taking place and thinking in increasingly more abstract ways.
- We move toward theoretical concepts and propositions.
- We can begin with a general topic and a few vague ideas that we later refine and elaborate into more precise concepts when operating inductively

Forms of explanation

- Theoretical explanation: a logical argument that tells why something takes a specific form or why it occurs
- Compare to prediction: a statement that something will occur. We do not have to explain why something happens, we can just say it will happen.

Casual explanations

- A causal explanation indicates a cause-effect relationship among variables.

Requirements for causality:

1. Temporal order: the cause must come earlier in time than the effect
2. Empirical association
3. Elimination of plausible alternatives: must show that the effect is due to the causal variable, not something else.

Diagrams of relationships between variables

DAGs

- DAGs (directed acyclic graphs) are a way of visualizing relationships between variables.
- DAGs are a fancy name for a flow diagram and involves drawing arrows and lines between the variables to indicate the relationship between them.
- Below, we think X causes Y

DAGs

- Let's switch to an example you are working with on the assignment.
- Say we are interested in the relationship between income and wanting more kids

- Note the implication here is that we are interested in the causal relationship
- That is, if I were to increase someone's income, what would that do to the likelihood that they would want more kids

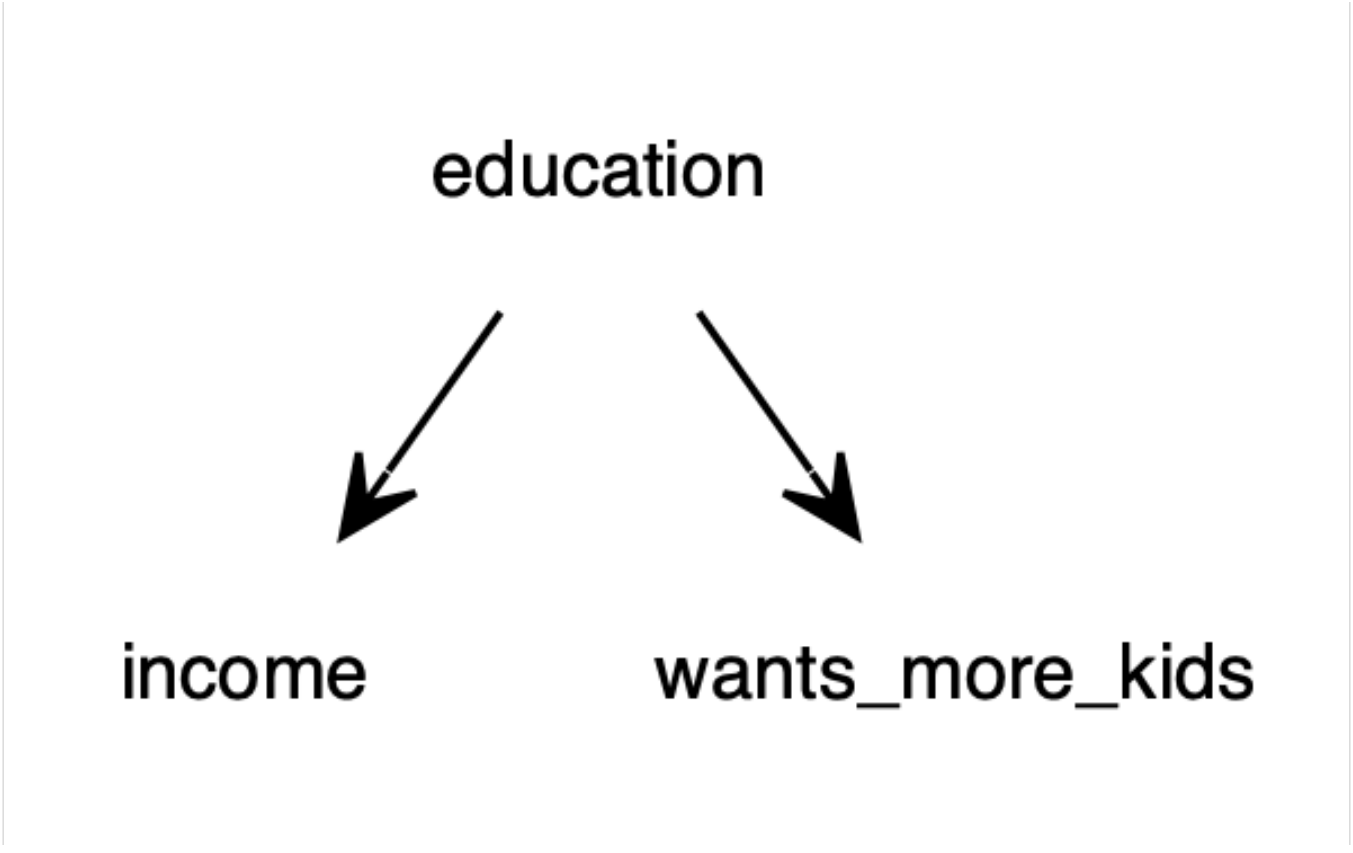
DAGs

- Now we want to build on this diagrams to consider other types of variables and how they may affect the relationship we observe between income and fertility intentions

To do

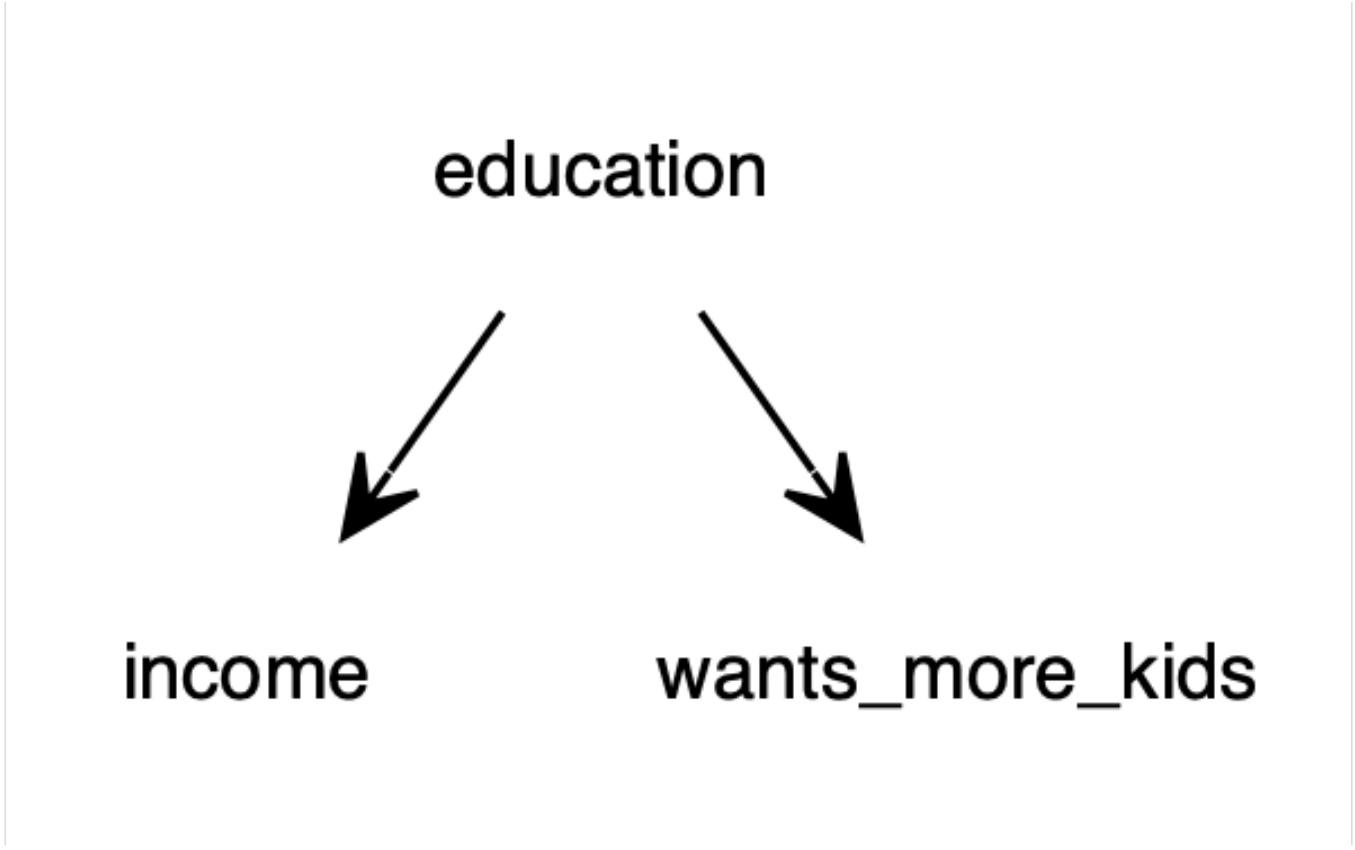
- Confounders
- Colliders
- Mediators
- Moderators

Confounders



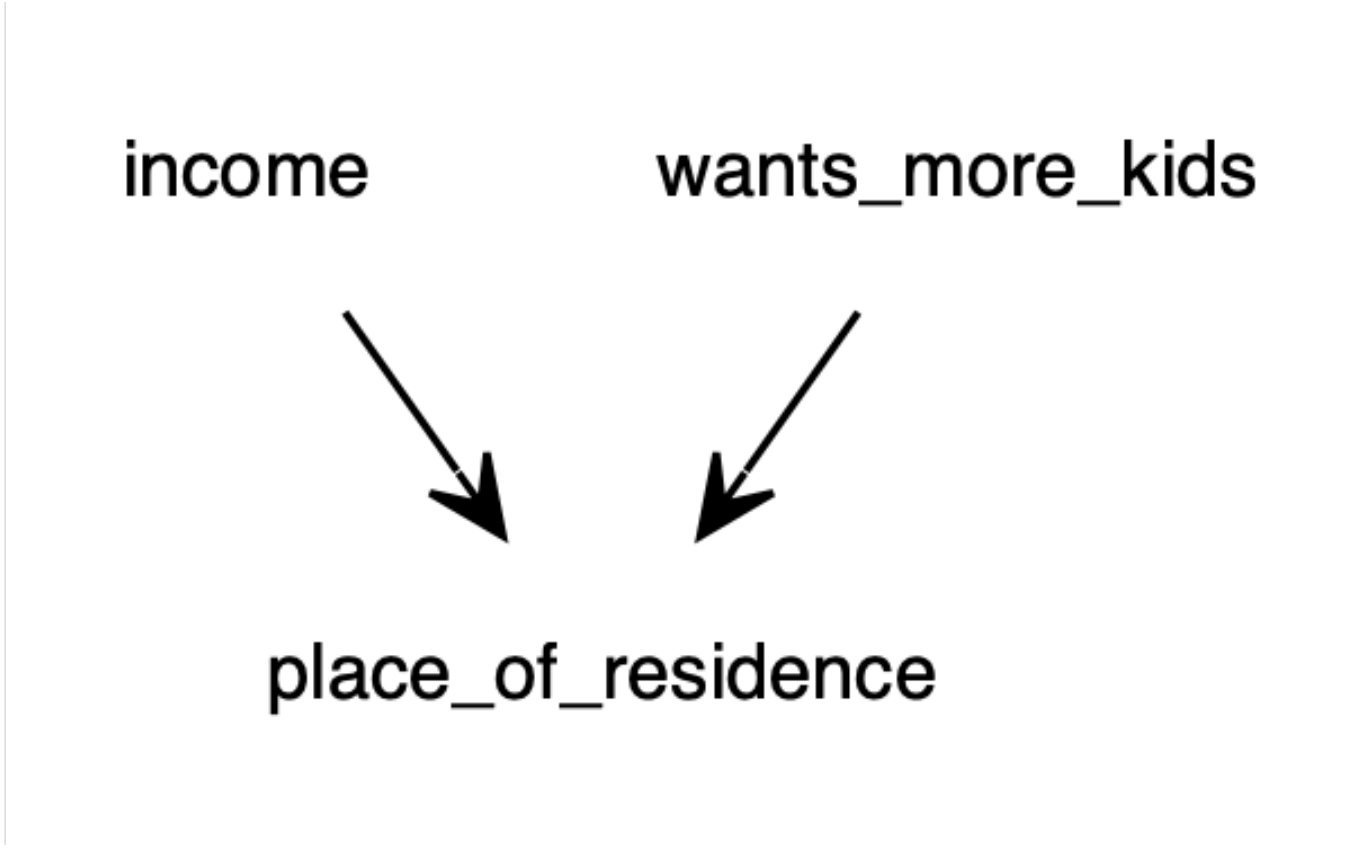
- Here, we think education causes changes in income, and also education causes changes in wanting more kids
- that is, education influences both the independent variable and dependent variable

Confounders



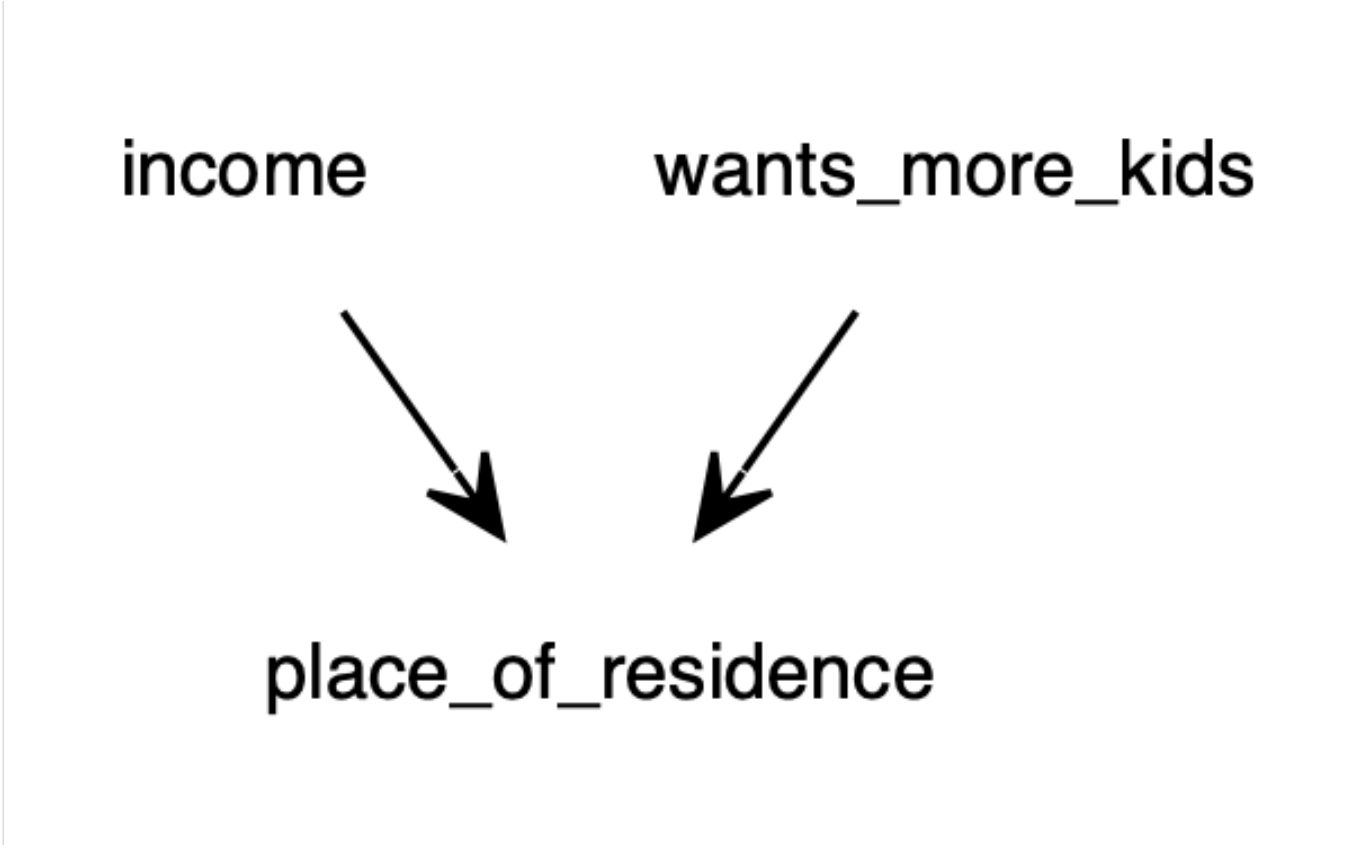
- Education is a confounding variable
- Failing to control for (condition on) education in a regression would create a spurious association between income and fertility intentions
- Or, even if there is an association, failing to control for education may overstate this association

Colliders



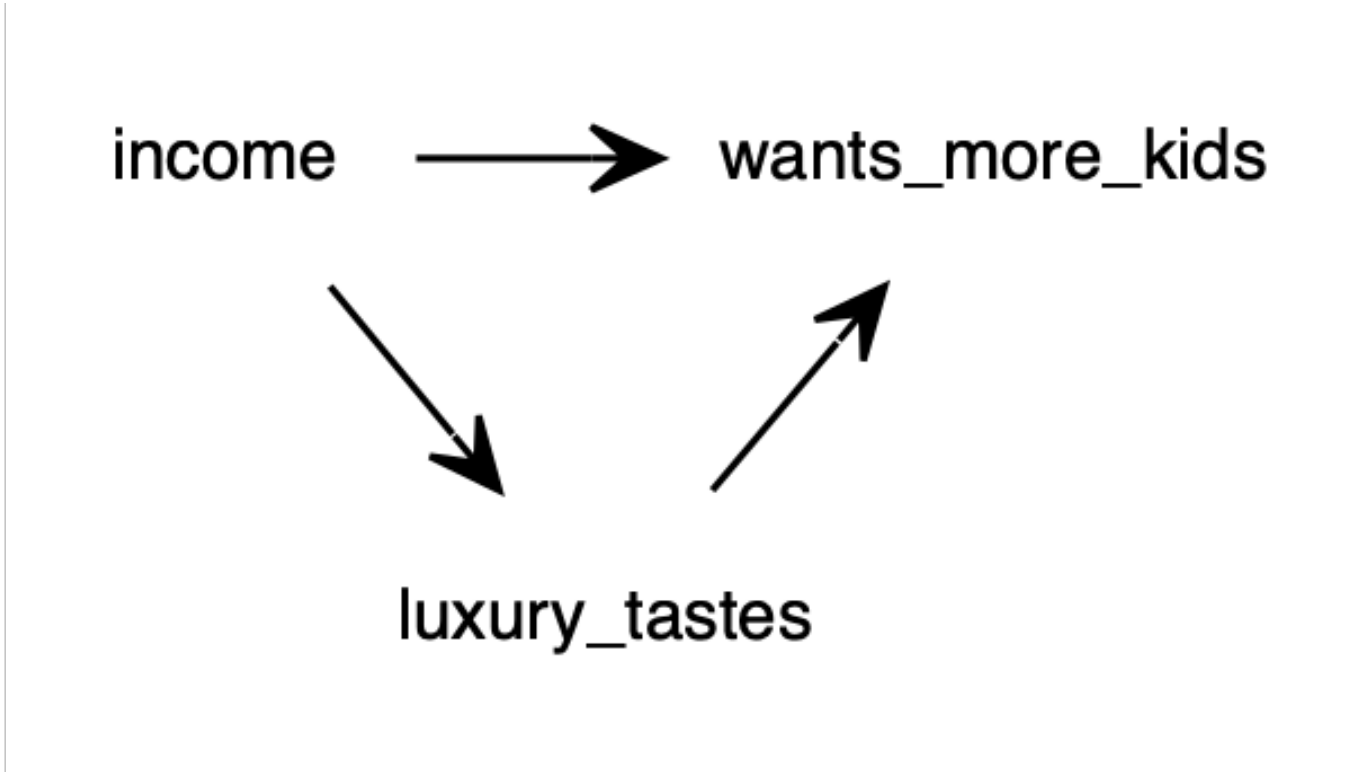
- Income influences where people live, but fertility intentions also influence where people live
- Place of residence is a collider variable, i.e. the independent and dependent variable of interest collide and both influence it

Colliders



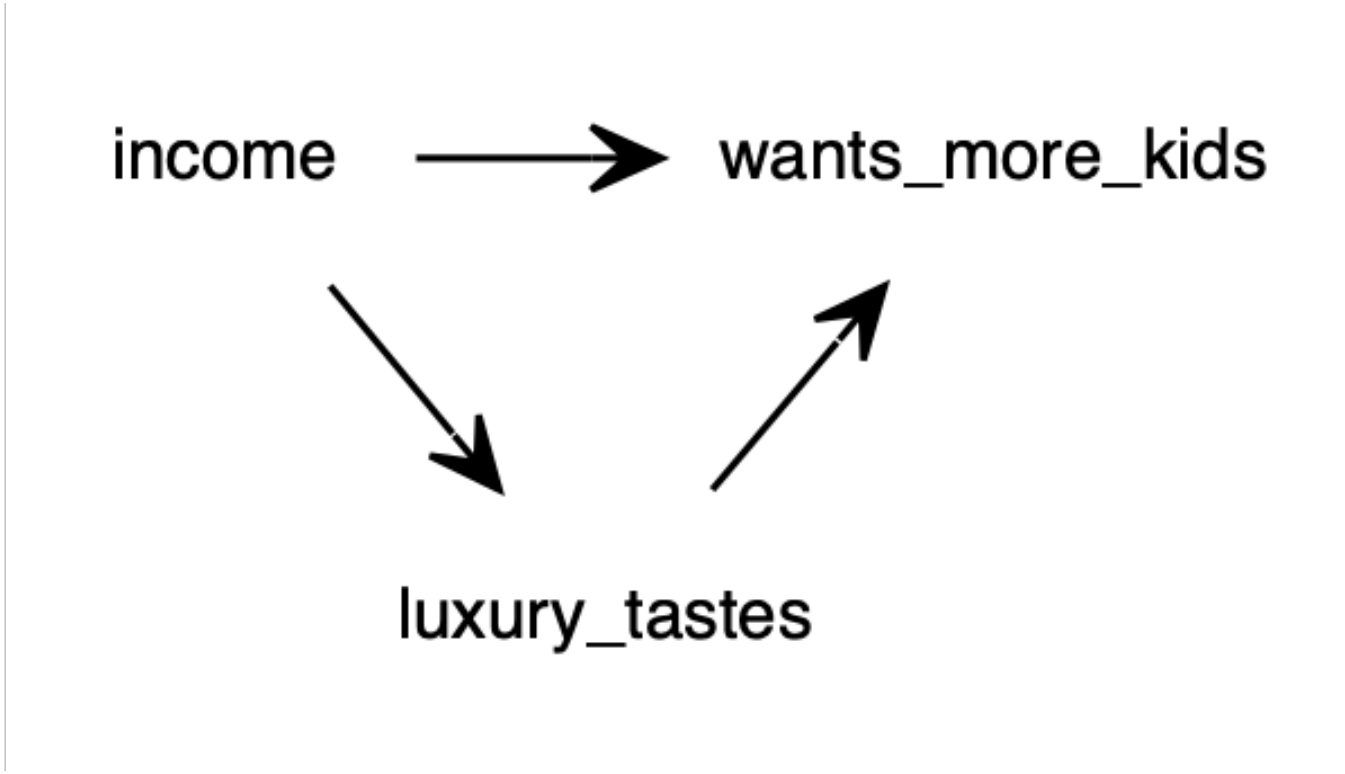
- Conditioning on (controlling for) place of residence may result in a spurious association between income and fertility intentions
- Think: income and payments example from previous assignment

Mediators



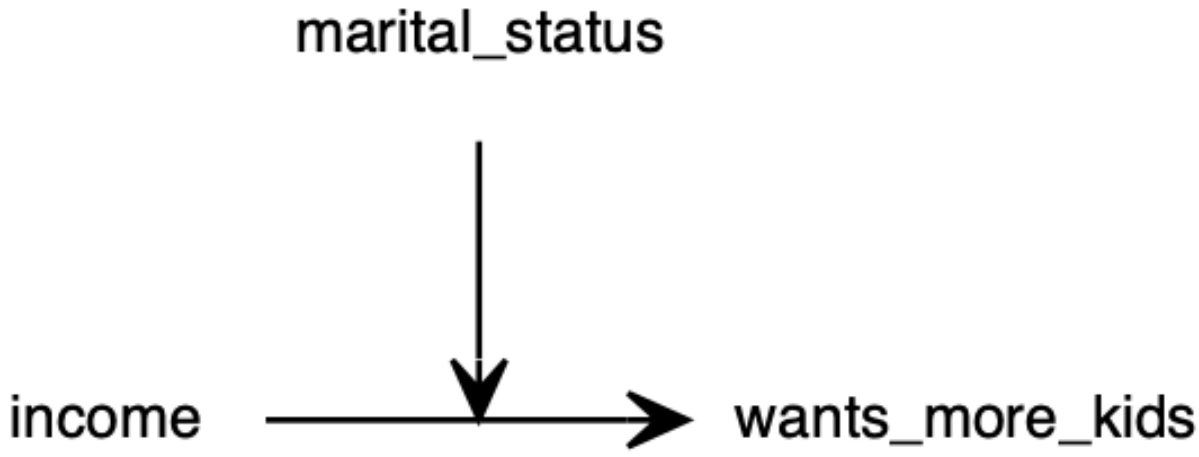
- Here we think that some degree of luxury tastes is the mechanism (or mediator) through which income influences the decision to have more children
- Increased income increases a desire for expensive things, which in turn decreases the desire for children

Mediators



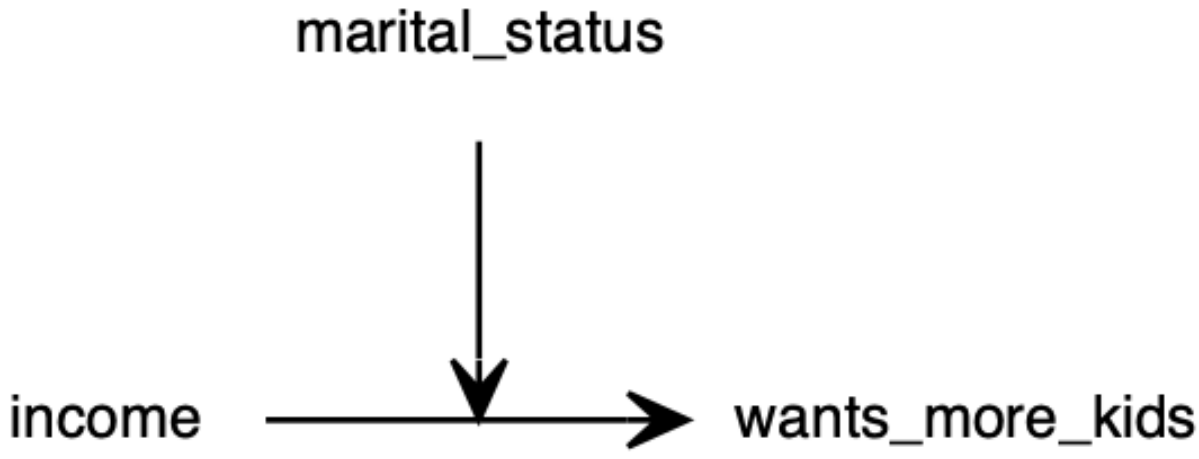
- To estimate the total effect of income on fertility intentions, we would not adjust for luxury tastes
- To estimate the direct effect of income on fertility intentions, we would adjust for luxury tastes
- Mediation Analysis (which we have not covered) decomposes the total effect of exposure X on outcome Y into the direct effect and the mediated effect transmitted through M.

Moderators



- The effect of income on fertility intentions depends on (is moderated by) marital status
- E.g. you could imagine the negative relationship between income and fertility intentions is less pronounced for married people compared to single people

Moderators



- Account for moderators by including interaction terms
- e.g. an interaction term between income and marital status would allow the affect of income on fertility intentions to be moderated by marital status

Linking back to regression methods

When to control for stuff????

We don't always want to control for as many independent variables as possible:

- If variable is a confounder, want to control for
- If variable is a collider, do not want to control for
- If variable is a mediator, control for if interested in total effect
- If variable is a moderator, include an interaction term

But how do you know?

You don't. Decisions are informed by

- Theoretical considerations, and what you're actually interested in testing
- Research design, limitations of data

DAGs are just a tool

"It is important to be clear about this: we must create the DAG ourselves, in the same way that we must put together the model ourselves. There is nothing that will create it for us. This means that we need to think carefully about the situation. Because it is one thing to see something in the DAG and then do something about it. But it is another to not even know that it is there."

From 'Telling Stories with Data'

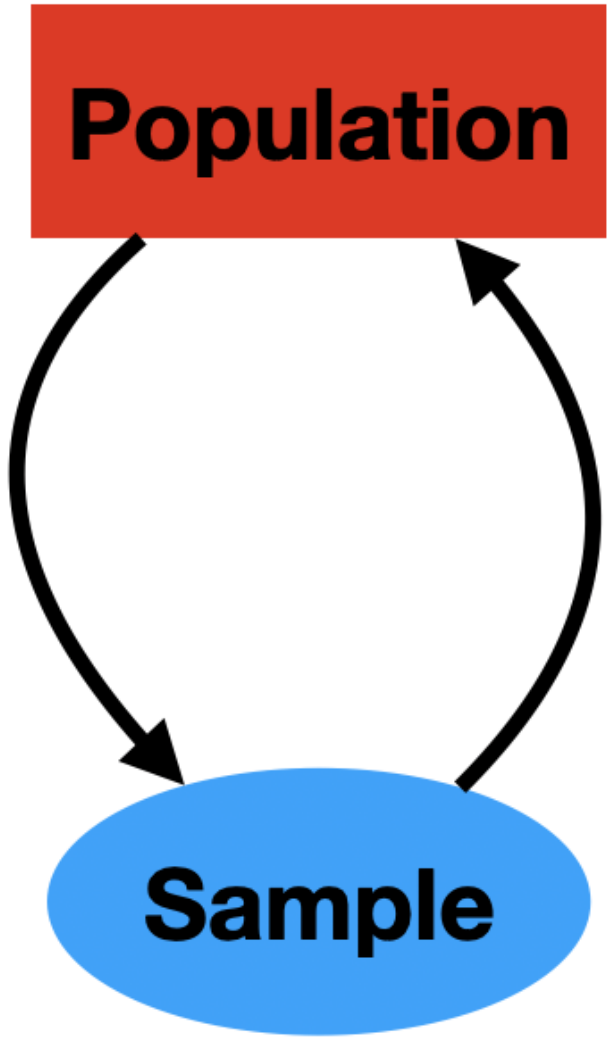
Causal interpretations versus association interpretations

- In this course I avoided causal interpretations
- Often it's just not possible given the data that we have available based on observational research
- Regression methods are still useful
  - for exploration
  - for prediction
  - in combination with other types of research
  - to inform future data collection / studies

Knowing the limits is important as reporting the model. Data and models with flaws are still useful, as long as you acknowledge those flaws

Back to the big picture

What it all comes down to



Summary

- We are interested in making inferences about a population
- **Toolkit 1**: Probability
  - when answering questions using the data we have available, there is chance/randomness involved
  - e.g. data from a sample
  - e.g. deciding whether a particular observation comes from a population
  - we can use probability to quantify uncertainty
- **Toolkit 2**: descriptive statistics and data visualizations
  - before running a model, we can get a long way by looking at key summary stats and charts
  - e.g. means by group tells us something about differences / similarities
  - e.g. key charts to visualize distributions and relationships

Summary

Toolkit 3: regression

- We are interested in explaining patterns in an outcome of interest (dependent variable) Y in relation to one or more explanatory variables $X_1, X_2, \ldots$
- i.e. how does Y vary with different levels of X_1?
  - If we consider X_2 as well, does Y still vary with X_1?
- We could explore this with graphs/ summary statistics!
- But **regression models** allow us to quantify relationships, taking into account **uncertainty** based on the data that we observe
- Can build regression models (study the association between explanatory variables and an outcome) for different types of outcome variables (continuous, binary, categorical. . .)

Focus on the data, not models

- A model is just that
- You make it up, and it can be useful or misleading
- More important to understand characteristics of your data, what it shows and what it might omit

Unsolicited advice

Unsolicited advice

- Many things that are useful after grad school are not taught in grad school
- Dissertation is only one bit
- Broad versus specific knowledge
- "Admin" stuff: Grant writing, teaching, mentoring
- Build your 'brand'
  - What's your niche?
  - What's your narrative?
  - 1/5/15min spiels of research
  - Make a website!

Websites with blogdown

- Consider making a website, if you don't have one already!
- If you are on the job market (academic or otherwise) people will Google you. It's a useful way to partially control what they see.
- Even before you're on the market, good to have, to build up a profile
- Lots of good tools, but you can also make websites in R :)

Blogdown

- Blogdown is an R package that lets you create websites in RMarkdown
- Built by people at RStudio so nicely integrated
- Builds on website templates from Hugo (https://gohugo.io/)

Example wesbites built with blogdown:

- Mine: https://www.monicaalexander.com/
- Julia Silge: https://juliasilge.com/
- Sharla Gelfand: https://sharla.party/

Follow Alison Hill's blog: https://www.apreshill.com/blog/2020-12-new-year-new-blogdown/

Presentation template

Slide 1: Introduction

- Research question
- Why you're interested/ why it's important
- Any previous studies

Slide 2: Data

- What is your dataset, where it comes from
- What is your outcome of interest
- Explanatory variables of interest
- Broad characteristics of dataset

Slide 3: a couple of interesting observations

Slide 4: Model

- What's your model

Slide 5: results

- key results from your analysis
- Are they surprising or not

Slide 6: conclusion, limitations, future

- summary
- Limitations
- what could you do in future (e.g. what data could be useful? what other analytical approaches could you try?)