Impact evaluation e Globalizzazione

21. Regression and Matching with Observational Data

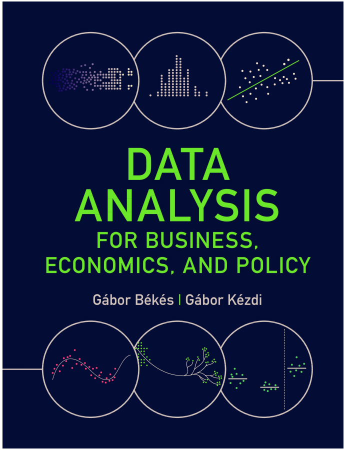
Vincenzo Lombardo

Data Analysis 4: Causality

Corso di laurea magistrale in Scienze Economiche e Finaziarie

A.A. 2023-2024

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- This slideshow is for Chapter 21

Regression and causality

- Causality – is about interpretation
- You see a pattern in the data – revealed by regression analysis
- Then, you interpret it....
- unless...
  - you get to design your own experiment
  - in that case you have a causal effect in mind and you induce controlled variation a variable
  - if all goes fine you know how to interpret patterns

Causality and regression

- You have observational data for many possible reasons.
- Experiments may be hard, expensive, unethical
- Look for great external validity
- Process of work?

Observational data approaches

- Thinking 1: Thought experiment
- Thinking 2: Variation in y - unobserved heterogeneity
- Thinking 3: Source of variation in x
- Tools 1: regression with controlling on confounders
- Tools 2: exact matching
- Tools 3: matching on the propensity score

Thinking 1: Thought experiment

- Data analysts turn to observational data for answering causal questions when they can't run an appropriate experiment.
  - Often there is not enough time or resources
  - would require controlling for too many things that would make external validity too low.
  - impossible run due to ethical concerns.
- Even when no experiment, worth to think about an experiment that could uncover the effect we are after.
- *thought experiments*: experiments that are designed in some detail but not carried out.

Thinking 1: Thought experiment

Thinking through a thought experiment when doing causal analysis on observational data has several advantages. It can:

- clarify the details of the *intervention* we want to examine and how it compares to the causal variable in the data.
- clarify the situations: what exactly it would mean for observations to be "treated" and "untreated".
- help understand the *mechanisms* through which the causal variable may affect the outcome.
- help understand how random assignment compares to the *source of variation* in the causal variable in our data.

Case study: Founder/Family Ownership and Quality of Management

- Though experiment
- We investigate whether the fact that a company is owned by its founder, or their family members, has an effect on the quality of management.
- Whether founder/family owned companies are better or worse managed than other firms, on average because of their ownership.
- This is a causal question: we are after an effect.
- Great way to understand what the intervention and the counterfactuals are.

Case study: Founder/Family Ownership and Quality of Management

- The subjects of this thought experiment are companies.
- The intervention is changing ownership of the company.
- For that we need a subject pool with the same ownership and randomly assign some of them to change their ownership.
  - To change ownership the owners would sell their stake to other investors, either directly or indirectly (stock market).
  - intervention works in one way
  - Effect of the intervention would be a form of ownership that can be the result of such sales.
  - restriction on the form of ownership after the intervention: some types of ownership are unlikely to emerge,

Case study: Founder/Family Ownership and Quality of Management

- Take all founder/family owned companies,
- Randomly chose half of them and make them sell their stakes to whoever would want that.
  - assume perfect compliance: treated companies receive offers that they don't refuse
- As a result of the intervention, untreated companies remain in founder/family ownership, while treated companies have other forms of ownership
- After some time, measure the quality of management among treated and untreated firms.
- The difference between their average quality scores would show the average effect of giving up founder/family ownership.

Case study: Founder/Family Ownership and Quality of Management

- Trick
- This thought experiment would identify the opposite of what the original question would imply.
- Instead of the "effect" of founder/family ownership it can measure the effect of giving up founder/family ownership.
  - effect identified in thought experiment = mirror image of the effect in our original question.
- Empirical work: the "effect" of founder/family ownership.
- Interpreting the results –> relate to experiment of selling stake and compare outcomes.
- There cases of family taking firm private

Variables to Condition on, Variables Not to Condition On

- Investigate sources of variation in the causal variable, two types of variation in x
  - Exogenous sources are variables that are independent of potential outcomes,
  - Endogenous sources are variables that are related to potential outcomes.
- Use exogenous sources in x, while conditioning on all endogenous sources of variation = confounders.
- Collect potential sources = thinking exercise

- Endogenous sources of variation, to condition on (confounders:
  - Common cause: the variable affects x and y.
  - Mechanism of reverse causality: y affects x through this variable.
  - Unwanted mechanism: x affects y through this variable, but we don't want to consider it when estimating the effect of x on y.

Variables to Condition on, Variables Not to Condition On

- Not condition on variables that are not part of endogenous variation
- bad conditioners: variables that data analysts should not condition on when attempting to uncover the effect of x on y:
  - An exogenous source of variation in x.
  - A mechanism that we want to include in the effect to be uncovered.
  - Common consequence: both x and y affect the variable

Variables to Condition on, Variables Not to Condition On

- Look at variables we shall have, and what we have
- List and categories
- Causal map (DAG)

- Use tools to condition on those variable we shall
  - Multivariate regression
  - Matching
  - Use smart tricks in rare settings

Conditioning, ATE, ATET

- Our usual aim is to estimate ATE
- Sometimes we also care about ATET: the treatment effect on the treated
  - ATET focuses directly on participants - sometimes this is what policy cares about
  - ATE may be driven selection or splillovers - sometimes you are interested in this
- If random assignment ATET=ATE
- With observational data, ATET may be different to ATE
  - No random assignment, treated and not treated subjects may be different (heterogeneous) in some unobserved way.
  - Example: self-selection as unobserved confounder

Case study: Founder/Family Ownership and Quality of Management

- Observational cross-sectional data
- World Management Survey = cross-section of many firms in manufacturing from 21 countries.
- The outcome variable is the management score.
- The causal variable is founder/family ownership.

- Several tasks before running regressions
  - Think about and identify sources of variation in ownership,
  - Draw a causal map,
  - Decide on observable variables to condition on

Case study: Sources of variation in ownership

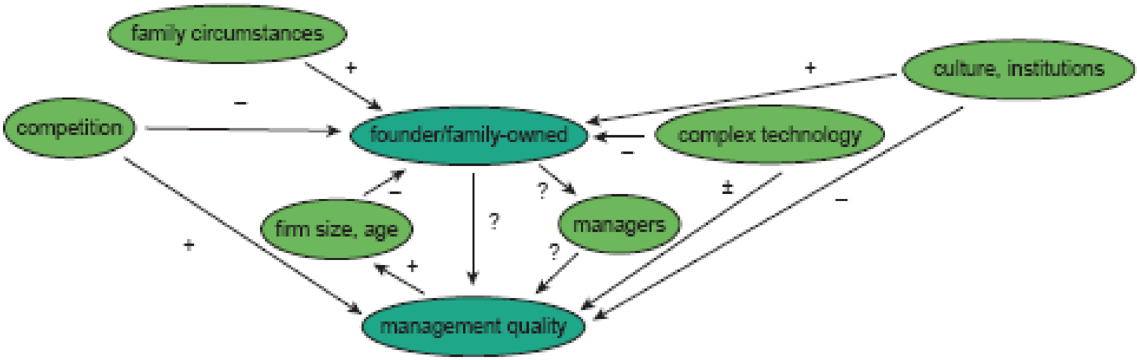
- Let us look for variation in x, ownership. Think + identify + decide.
- Firm started as founder/family-owned?
  - Alternative: spin-offs, joint ventures, multinational affiliates of other firms, including multinationals.
- Products and technology affect ownership = sources of variation in x. How about y?
- It's likely to be an endogenous source, technology correlated with management, too.

Case study: Sources of variation in ownership

- Let us look for variation in x, ownership. Think + identify + decide.
- Cultural and institutional factors, norms in a society. Affect cost of starting business, FDI. How about y?
- Likely endogenous source, culture, norms correlated with management, too.

- How about family features. Children of founders, their interests, skills. Clearly affects if ownership may be passed on. How about y?
- Likely exogenous - gender/number of kids not related to management quality
- This is the variation we need but not use as control!

Case study: Founder/family ownership: sources of variation in observational data. Causal map



Case study: Sources of variation in ownership

- Family circumstances – exogenous variation in x
- Competition – common cause confounder
- Culture and institutions – common cause confounder
- Technology, product type – common cause confounder
- Firm size, firm age – hard – may be mechanisms of reverse causality
- Feature of managers (their age, experience) – mechanism
- which ones to control on?

Case study: Sources of variation in ownership

- Family circumstances – exogenous variation in x [NO Control]
- Competition – common cause confounder [Control]
- Culture and institutions – common cause confounder [Control]
- Technology, product type – common cause confounder [Control]
- Firm size, firm age – may be mechanisms of reverse causality [Maybe Control]
- Feature of managers (their age, experience) – mechanism [NO Control]

Conditioning on Confounders by Regression

- Linear regression to condition on other variables to estimate the effect of x on y, conditioning on observable confounder variables $(z_1, z_2, \dots)$:

$$y^E = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \dots \tag{1}$$

- Note: β_1 always = estimate of average difference in y between observations that are different in x but have the same values for $z_1, z_2, \dots$ Even if not causal.

- If the $z_1, z_2, \dots$ variables capture all endogenous sources of variation, x is exogenous in the regression.
  - Conditional on $z_1, z_2, \dots$, variation in x is exogenous.
  - OLS estimate of β_1 is a good estimate of ATE of x on y.

Conditioning on Confounders by Regression

- Conditioning on all relevant confounders - very unlikely in observational data.
- z_1, z_2, ... capture some, but not all, of the endogenous sources of variation in x, x is endogenous in the regression
  - OLS estimate of β_1 is a not good estimate of the average effect of x on y.
- OLS is biased - omitted variables bias = difference between the true ATE of x on y and estimated ATE for the β_1 coefficient on x by this regression.
  - When x is exogenous in the regression, the omitted variable bias is zero.
  - Chapter 10: bias depends on how the omitted confounders are related to x and y.

Conditioning on Confounders by Regression

- OVB is positive (estimated ATE > true ATE) when the omitted confounders are correlated in the same direction with x as with y.
  - OVB negative when omitted confounders associated in the opposite direction with x and y.
- If we can speculate well, we can sign the omitted variable bias
  - Sometimes can.
- Signing OVB is often the key task - could help a great deal to see where we are re causality.

Selection of Variables in a Regression for Causal Analysis

- In practice, key question is: variable selection
  - Which z variables to add -all observed confounders or only some? Which ones?
  - What functional form? Interactions?
- Variable selection matters IF choices impact estimated ATE (coefficient estimates on x).
  - When equal: prefer simplest model, with the fewest variables, the simplest functional forms, and the fewest interactions.
- IF different regressions give substantially different coefficient estimates on x. pick one that includes more variables.
  - More variables, more flexible functional forms, or more interactions.
  - Still make sure to avoid bad conditioning variables,
- Adding variables that don't matter - usually no big deal.
  - But, in smaller dataset, it can make the effect estimates imprecise
- Often sample size determines what we can do

Case study: data

- Observational cross-sectional data
- World Management Survey.
- It is a cross-section of many firms in manufacturing from 21 countries. Representative sample of firms within countries.
- Consider a cross-section, each firm is just once in sample

Case study: outcome and causal variable

- The outcome variable is the management score.
  - Average of 18 scores that measure the quality of specific management practices.
  - Each score is measured on a 1 through 5 scale, with 1 for worst practice and 5 for best practice.
- The causal variable is founder/family ownership.
  - The ownership variable detailed
  - binary variable 1: firm is founder owned or family owned
- Other types of ownership we are interested in = could be the result of founders or their family selling their shares.
  - Drop observations that were owned by the government or a foundation or the employees. Why?
  - We also dropped observations with missing ownership data and "other" ownership type.

Case study: Summary of confounders

- List of confounders: suggested by causal map + available data
- Technology - industry dummy; share of college-educated workers (outside senior management).
- Customs, law - country dummy, product competition
- Firm size - not sure if confounder or bad control.
  - will try with and without
- Other variables that we'll use in our analysis: employment, college share, competition, industry, country

Exact matching

- Linear regression is an approximation
  - the difference in average y between observations with different x but the same values for the other right-hand-side variables z_1, z_2, ...
- Why do approximation when can compare observations with the same z_1, z_2, ... values?
- Could we take those variables and find observations with the exact same values?
- This is idea of matching: compare the outcomes between observations that have the same values of all of the other variables and different values of the x variable.

Exact matching

- Ideal case exact matching - not an approximation.
- It matches observations on exact values
- Aggregation: observations = different value-combinations of all confounders

- With $z_1, z_2, \dots$ variables, each cell would have a particular value-combination $z_1 = z_1^*, z_2 = z_2^*, \dots$.
- Within each cell, Compute the average y for all treated observations and the average y for all untreated observations, and we take their difference:

$$E[y|x = 1, z_1 = z_1^*, z_2 = z_2^*, \dots] - E[y|x = 0, z_1 = z_1^*, z_2 = z_2^*, \dots] \qquad (2)$$

Exact matching

- ATET = number of treated observations in the cells as weights
- Matching gives a good estimate of ATET when selection is based on observables
  - This is often the default
- ATE = can calculate by some re-weighting - average of differences weighted by the number of observations in cells.
- If ATE and ATET is very different - something problematic is going on.
  - Strong self-selection, a confounder we did not take into account.

Exact matching

- It is feasible when many observations, few variables or variables with few values.
- In practice, exact matching is rarely feasible.
  - unlikely to find exact matches for all z values.
- In practice, in some cells have $x = 1$ observations only, others, $x = 0$ only.
- For ATE: both are problem
  - For ATET, need cells in which we have $x = 1$ observations

Exact matching

- In practice, in some cells have $x = 1$ observations only, others, $x = 0$ only. Two possible reasons:
- Substantive problem: $x = 1$ and $x = 0$ observations differ so much that some values of some confounder variables exist only in one of the two groups in the population.
- Data problem. A value combination is not there in our sample, but could be, and could very well be in the population
  - Larger sample can help
- Can we know which one we face?

Coarsened exact matching

- Coarsening qualitative variables means joining categories to fewer, broader ones and creating binary variables for those broader categories (e.g., groups of countries, less refined industry categories).
- Coarsening quantitative variables means creating bins (e.g., bins for age of individuals or size of organizations).
- Fewer binary variables and fewer bins of quantitative variables make matches mode likely by reducing the number of variables.
- Coarsening is based on a trade-off: it makes exact matches more likely but it reduces variation in the confounder variables used for the matching

Exact matching: summary

- The interpretation of this estimate is intuitive: it is the average difference in y between treated and untreated observations that have the exact same $z_1, z_2, \ldots$.
- Recall that the linear regression gives an approximation to this average difference.
- In contrast, exact matching is not an approximation.
- If matching is successful for all $x = 1$ observations, it gives exactly the average difference in the data.
- The key problem is feasibility: could be too many values. Aggregation is arbitrary.

The idea of the common support

- Exact matching may fail for a substantive reason = there is a lack of common support.
  - "Support" = the set of values a variable can take.
- Common support = confounders can take the same values among treated and untreated observations.
- In the population or general pattern, our data represents.
- When we don't have common support, we can't estimate the effect for all subjects in the data.

The idea of the common support

- Consequence is general not just for matching
- We shouldn't (cannot) estimate ATE when have no common support.
- Instead, we shall estimate the effect of x on the part of the dataset with common support
- Compare distributions with histograms, tabulate key categorical variables, even interactions
- Drop ranges of observations when no common support

Matching on the Propensity Score

- Idea = creating a single quantitative variable from the many confounder variables.
- Matching is then done by finding similar observations in terms of this single quantitative variable.
- Similar observations = nearest neighbors.

- Most widely used method is called matching on the propensity score.
- The propensity score is a conditional probability: it is the probability of an observation having $x = 1$ as opposed to $x = 0$, conditional on all the confounder variables z.
- The propensity score is a single quantitative variable (the probability) that combines all confounder variables (the conditioning variables)

Matching on the Propensity Score

- The propensity score is not something we know. It is something we need to estimate it.
- That means estimating, or, more precisely, predicting, the probability of $x = 1$ for each and every observation in the data, based on what values they have for the z variables.
- The usual procedure is to estimate a probability model, most often a logit, for the probability of $x = 1$, as a function of the confounder variables.

Using a logit, we get the propensity score, *pscore*,

$$pscore = P[x = 1|z_1, z_2, ...] = x^P = \Lambda(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + ...) \tag{3}$$

Matching on the Propensity Score

- With the propensity score at hand, we can match $x = 1$ and $x = 0$ observations that are close to each other.
- The most widely used matching procedure is *nearest neighbor matching on the propensity score*.
- This procedure takes each $x = 1$ observation, matches it to the $x = 0$ observation with the nearest value of the propensity score.
- If many $x = 0$ observations are nearest neighbors, all are picked and average outcome taken.
- Once a match is found, take difference of y values between the matched $x = 1$ and the $x = 0$ observation.

Matching on the Propensity Score

- Matching and then difference taking is repeated for all $x = 1$ observations.
- The estimated effect of x on y is then the average of those differences.
- If all confounders are included, the propensity score incorporates all endogenous sources of variation in the causal variable.
- In practice, many possible decisions...

Case study: variables

- The outcome variable is the management score: range in the data is 1 to 4.9, its average is 2.88, standard deviation 0.64
- The causal variable is whether the firm is owned by its founder or their family: 45%
- Direct comparison: 2.68 vs 3.05
- Founder/family owned firms – management score is -0.37 points lower, on average.
  - Difference a little more than half SD of outcome variable (0.64) - so large in magnitude
- Causal statement would be like: The quality of management in founder/family owned firms would increase by 0.37 points, on average, if the ownership of their firm were transferred to other investors.
  - Transferring ownership away from founder/family would make management quality improve

Case study: Estimates of the effect of founder/family ownership on the quality of management. Multiple regression results

Variables	(1) No confounders	(2) With confounders	(3) With confounders interacted
Founder/family owned	-0.37**	-0.19**	-0.19**
	(0.01)	(0.01)	(0.01)
Constant	3.05**	1.75**	1.46**
	(0.01)	(0.05)	(0.22)
Observations	8,440	8,439	8,439
R-squared	0.08	0.29	0.37

Note: Outcome variable: management quality score. Robust standard error estimates in parentheses.** $p<0.01$, * $p<0.05$. Source: wms-management-survey dataset.

Case study: Add variables

- When adding confounders, coefficient drops from -0.37 to -0.19
- The quality of management is lower, on average, by 0.19 points or about 30% of a standard deviation, in founder/family-owned firms than other firms of the same country, industry, size, age, with the same proportion of college-educated workers, and with a similar number of competitors.
- Adding confounders with interactions, quadratic forms, does not matter
  - causal variable + up to 745 variables in the regression

Case study: Causality and signing the bias

- When adding confounders, coefficient is -0.19.
- Biased? Yes. But how?
- Most omitted confounders are correlated with founder/family ownership and the quality of management in opposite directions.
- the estimated effect of founder/family ownership is biased in the negative direction.
- Thus the true effect is probably weaker (less negative).
  - As did confounders we have already added.
- True effect could be zero. Or even positive.
- What can we do to increase belief in causality?

Comparing Linear Regression and Matching

- ATE (and ATET) make sense only with common support.
- Regression and matching uncover, deal lack of common support differently.
- Exact matching automatically drops observations (no matching).
- Matching on the propensity score, also detects the lack of common support.
  - If PS close to 0 or 1 – not be matched by nearest neighbor matching.
- Linear regression not detect the lack of common support. Uses all observations to produce its coefficients.
  - This would include observations without common support.
- Lack of common support -> estimate a biased average effect of x on y.
  - Estimated regression line affected by observations that are not supposed to count.

Comparing Linear Regression and Matching

- When estimating ATE by regression, we need to make sure that the support is common before the estimation.
- The lack of common support means OLS may under or over-estimate the effect of x on y.
- Extra step of data analysis.

Case study: Common support

- We argued that common support is needed to avoid biased ATE
- While matching is designed to do that, we can check it with regressions
- Checked statistics of the distributions of each included confounder among founder/family owned vs other ownership.
- Concluded: common support assumption OK in our data
- Main reason why similar results from regression and matching

Case study conclusions

- We estimated an average treatment effect, fairly precisely.
- Is this the "true" effect of founder/family ownership of a company on the quality of management?
- Probably not, more likely an upper bound in magnitude
  - Most likely other confounders, negative bias - overestimated size of the effect

Case study conclusions

- Did conditioning on observable confounders matter?
  - Yes
  - When we conditioned on what we could, the difference halved
- Did the way we condition on them matter?
  - No
  - Regression estimates were essentially the same as the estimates from matching on the propensity score
  - Including many interactions among the confounder variables didn't matter, either
- What matters is what we can condition on
  - The causal map helped outline what we would want to condition on
  - Our data had a small subset of those variables
- If we want a better estimate need to measure more of those potential confounders
- Or isolate exogenous variation in x in some other way

Review of advanced methods to help read papers

- Introduce two ways to isolate exogenous variation in x to uncover its effect on y
  - instrumental variables
  - regression-discontinuity.
- Alternative to condition on all confounders
- Make sure that we use only the exogenous part of variation in x for estimating its effect.
- Can be used under specific circumstances.

Instrumental variables

- Instrumental variables (IV) is a method to estimate the effect of x on y
- By directly isolating an exogenous source of variation in x

- Under ideal circumstances the IV method can give a good estimate of the effect
- In observational data
- Even if there are endogenous sources of variation in x, too

Instrumental variables main idea

- There is a variable in the data that is an exogenous source of variation in x
- This is called the instrumental variable, IV, or simply the instrument
  - The IV is independent of potential outcomes
  - The IV affects x
  - The IV has no direct effect on y
- Compare y across observations that are different in the IV
  - If there is a difference in observed y
  - That must be the effect of the IV
    - Because the IV is exogenous (independent of potential outcomes)
  - And the effect of the IV is only through x
  - Thus, that difference in observed y is because of the effect of x on y

Instrumental variables example

- What is the effect of having more than two children (x, binary) on whether the mother works for pay (y, binary), in the USA?
- The IV is whether the first two children have the same sex
  - It's one of the many sources of variation in x
    - It does affect x: the proportion of women with more than two children is 6 percentage points higher ($+0.06$) if the first two children have the same sex (USA).
  - The IV is likely exogenous
  - The IV likely has no effect on y except through x
- Women whose first two children have the same sex are less likely to work for pay
  - Difference is 0.8 percentage point (-0.008)
- That difference must be the effect of those women being more likely to have more than two children

Instrumental variables example

- So we established that having more than two children leads to a lower likelihood of work for pay
- But by how much?
- Answer: adjust the effect of same-sex first children on y (-0.008) by its effect on x (0.06)
- The effect of having more than two children (x) on working for pay (y) is then negative 13 percentage points
  - $-0.008/0.06 = -0.13$

Instrumental variables formula

$$\hat{x}^E = \hat{\pi}_0 + \hat{\pi}_1 IV \quad (4)$$

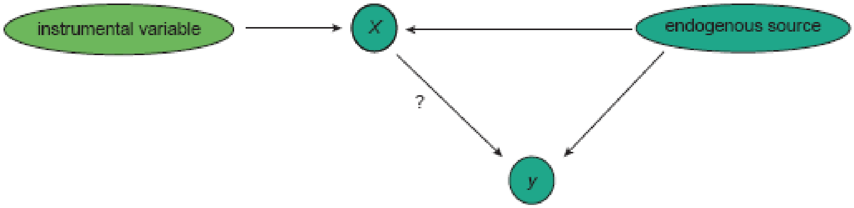
$$\hat{y}^E = \hat{\phi}_0 + \hat{\phi}_1 IV \quad (5)$$

$$\hat{\beta}_{IV} = \hat{\phi}_1 / \hat{\pi}_1 \quad (6)$$

- First equation is the effect of the IV on x
  - Called the first stage
  - In the example $\hat{\pi}_1 = 0.06$
- Second equation is the effect of the IV on y
  - Called the reduced form
  - In the example $\hat{\phi}_1 = -0.008$
- Third equation is the instrumental variables estimate of the effect of x on y
  - In the example $\hat{\beta}_{IV} = -0.13$

Causal map with an instrumental variable

- This causal map illustrates a situation in which the IV works even though there is endogenous source of variation in x
- As long as the IV is an exogenous source



Instrumental variables summary

- When applicable, IV is a powerful method to estimate the effect of x on y
- When is it applicable?
- The key assumption is exogeneity
  - The IV should be independent of potential outcomes
  - It can affect y only through x
  - This is an assumption that we can't verify
- The other assumption is that the IV should affect x
  - This we can easily check in the data

- It's usually difficult to find an IV that fits the requirements
- When the requirements are not met, the IV estimate is biased
  - And the IV estimate doesn't necessarily get us closer to the true effect

Regression-discontinuity

- Regression-discontinuity (RD) is another method to estimate the effect of x on y
- By directly isolating an exogenous source of variation in x even in the presence of endogenous variation, too

- It is applicable under very specific circumstances
- When there is a threshold value of a variable that determines treatment
  - This is called the running variable
  - For example, an age threshold (age is the running variable)
- Main idea: subjects on the two sides of the threshold are very similar to each other
  - The closer they are to the threshold the more similar they are
  - In their potential outcomes, too
- So it's almost like random assignment

Regression-discontinuity example

- Subjects are unemployed people
- Intervention is a compulsory program that helps job search (x)
- Outcome is whether they find a job in 3 months (y)
- Subjects below age 25 are required to participate in the program
- Subjects 25 or older cannot participate in the program
- Compare the outcome of 24-year-old subjects and 25-year-old subjects
  - If average y differs between the two groups that's because of the effect of the program
  - Because the job finding rate with or without the program (potential outcomes) should be similar

Regression-discontinuity extensions and caveats

- A version of RD allows for both sides of the threshold to be treated with some probability
  - In the simple version above the probability was one for one group and zero for the other
  - In the general version all is needed is a noticeable difference in the treatment probabilities at the threshold of the running variable
- Caveats
  - The threshold of the running variable would determine the intervention probability only
    - Nothing else related to potential outcomes
  - Subjects should not be able to manipulate the running variable
  - The method can give a good estimate of the effect for the group of subjects around the threshold value of the running variable

Main takeaways

- We need exogenous variation in x to uncover its effect on y, but that's hard to achieve with cross-sectional observational data
  - We can rarely condition on all confounders, so our effect estimates are almost always biased
  - By conditioning on what we can, we may decrease this bias
  - We may be able to sign the bias
- Linear regression and matching on the propensity score are alternative ways to condition on observable confounders
- With common support, regression and matching tend to give similar results
- With experience and luck, we may find another, more direct way to isolate exogenous variation in x
  - Instrumental variables method
  - Regression-discontinuity design