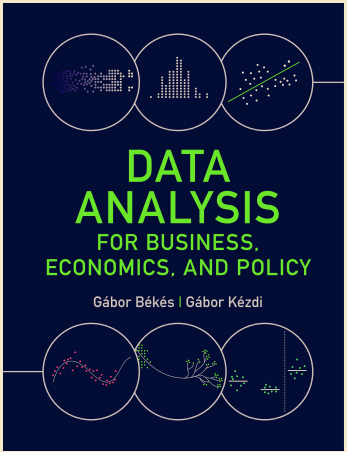
Békés-Kézdi: Data Analysis, Chapter 21: Regression and Matching with Observational Data



Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

gabors-data-analysis.com

Central European University

Any comments or suggestions:
gabors.da.contact@gmail.com

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
  - have a causal effect in mind
  - you induce controlled variation a variable
  - 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: Variables to Condition on, Variables Not to Condition On
- Tools 1: regression with controlling on confounders
- Tools 2: exact matching
- Tools 3: matching on the propensity score

Think and design

Thinking 1: Thought experiment

- observational data for answering causal questions when no appropriate experiment.
  - Not enough time or resources
  - possible but external validity too low
  - impossible 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 vs causal variable in the data.
- clarify the situations: "treated" and "untreated".
- help understand the mechanisms .
- 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.

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.
- Thought experiment?

Case study: Founder/Family Ownership and Quality of Management

- Subjects – companies.
- The intervention is changing ownership of the company.
- Subject pool with the same ownership and randomly assign some of them to change their ownership.
  - Change ownership = owners would sell their stake
  - intervention works in one way
  - Effect of the intervention would be a form of ownership that can be the result of such sales.
- Details of the process?

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

Regression design

Variables to Condition on, Variables Not to Condition On

- Sources of variation in the causal variable, 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

Variables to Condition on, Variables Not to Condition On

- 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.

Confounders in practice: Self-selection

- In business, economics and policy applications some confounder variables represent selection.
- Self-selection - when subjects themselves decide/affect being treated or not
  - that decision is related to confounder variable z that affects the outcome y as well.
  - Could be common cause or unwanted mechanism

- Example: People living healthier lives are more likely to take vitamin supplements, and they are more likely to stay healthy even if vitamin supplements have no effect on health.
  - Self-selection: people decide on healthy lives = self-select into behavior

Variables to Condition on, Variables Not to Condition On

- Not condition on variables that are not part of endogenous variation (bad conditioners)
  - 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 should like to have, and what we actually 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

Average Effects in Subgroups and ATET

- Remember heterogeneity in ITE

Average Effects in Subgroups and ATET

- Remember heterogeneity in ITE
- ATE = average of te_i across all subjects in the population that we defined.
- Consider a key subgroup - those that are treated
- ATET = the average treatment effect on the treated
  - subgroup = all subjects that end up being treated.
- ATET sometimes equals ATE, but other times it does not

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 is different to ATET when treated and not treated subjects are different
  - Often in some unobserved way.
  - Example: self-selection as unobserved confounder: people had some say in which group they belong.

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

- 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?

Case study: Sources of variation in ownership

- 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?

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, and norms 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, and norms correlated with management, too.

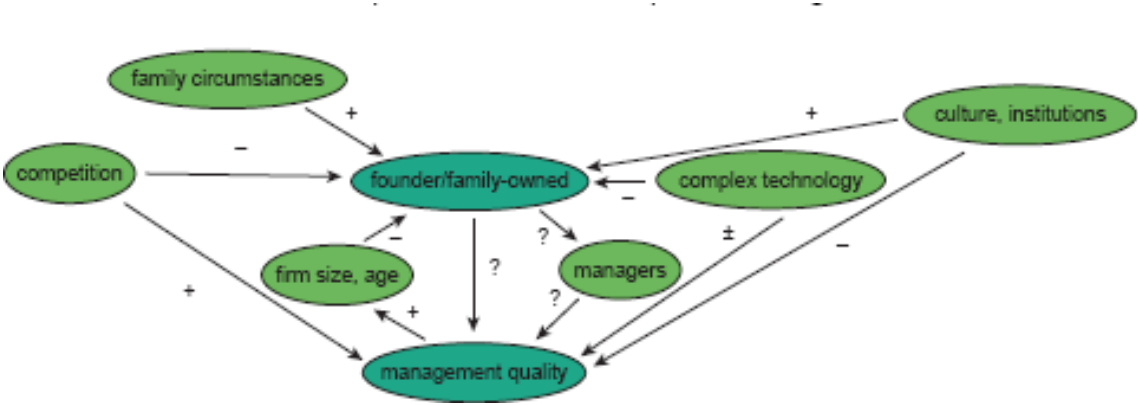
- How about family features? Children of founders, their interests, skills. Clearly affects if ownership may be passed on. How about y?

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, and 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 do 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, ...):

$$y^E = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + ... \quad (1)$$

- 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, ... Even if not causal.

- If the z_1, z_2, ... variables capture all endogenous sources of variation, x is exogenous in the regression.
  - Conditional on z_1, z_2, ... , 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.

Case study on food and health

- $\beta = 0.0040$ (100 more grams of fruit and vegetables - 0.4p lower bp)
- We missed health-consciousness. In reality, where would β be?
- We missed doctor's advice. In reality, where would β be?

Selection of Variables in a Regression for Causal Analysis 1

- 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.

Selection of Variables in a Regression for Causal Analysis 2

- 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

Selection of Variables in a Regression for Causal Analysis

- How to proceed in practice?
- Make sure to include confounders, and exclude bad controls
- Try out a few simpler models, see what happens.
- If sample is large enough, have all possible measures of confounders.
  - In smaller samples, need to pick key ones.
- Most often, complicated functional forms, interactions do not matter much.
  - Could matter for the top few confounders at best. Such as Income and gender.

Selection of Variables in a Regression for Causal Analysis

- There are algorithms to help with this.
- Not trivial procedure.
- Frontier of research
  - Double robust machine learning

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
  - highest proportion of family firms in India and Brazil (72% and 60%), lowest in Sweden and Poland (14% and 24%).
- Firm size - not sure if confounder or bad control.
  - will try with and without

Case study: data re confounders

- Other variables that we'll use in our analysis:
  - employment,
  - proportion of employees with college education (except for management),
  - whether the firm faces moderate or strong competition in the market of its main product,
  - in what industry the firm operates (20 categories of standard industrial classification with 2 digits),
  - Country of the firm.
- Drop if these vars missing; less than 50 employees or with more than 5000 employees.

Case study: data cleaning

Discuss prep steps from code

- How to create key variables
- Look for errors and weird stuff: tabulate, look at histograms:
  - drop if error
- Condense (merge) categorical vars when low variation - reduce noise
- Sample selection - before analysis
  - Why do it?
- look at descriptive stats, correlation patterns - help design functional form for confounders.
- Start with 10,282 observations, final work set is 8439

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
  - What does this mean? Why think in terms of SD?
- The causal variable is whether the firm is owned by its founder or their family.
- 45% of the firms are founder/family owned in the data.
- The rest: private equity: 4%, Dispersed (including stock market): 29%. Non-founder private: 22%
- Think back to thought process...

Case study: simple difference

- Direct comparison: 2.68 vs 3.05
- When comparing founder/family owned firms with the other firms in the data we find that their 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
  - Simple tabulation vs regression. Same result? Why regression?

Case study: simple difference

- Direct comparison: 2.68 vs 3.05
- When comparing founder/family owned firms with the other firms in the data we find that their 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
  - Simple tabulation vs regression. Same result? Why regression?
- Causal statement would be like:
  - The quality of management in founder/family owned firms would increase by 0.39 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: Add variables

- proportion of non-management employees with college education (a quantitative variable) - we created four bins the proportion of college-educated workforce
- whether the firm faces weak, moderate or strong competition in the market of its main product (qualitative variable with 3 categories)
- in what industry the firm operates (20 categories of standard industrial classification)
- in what country the firm operates (24 countries)

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?
- Being in big city. Being close to Stanford.

Case study: Causality and signing the bias

- When adding confounders, coefficient is -0.19.
- Biased? Yes. But how?
- Being in big city. Being close to Stanford.
- 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?

Matching

Exact matching

- Linear regression is a linear 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 approximate 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?

- Matching: compare the outcomes between observations that have
  - the same values of all z_i variables
  - 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 $z_1, z_2, ...$ variables, each cell would have a particular value-combination $z_1 = z_1^*, z_2 = z_2^*, ...$
- Within each cell, Compute the average y for all treated observations and the average y for all untreated observations, and we take their difference:

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

Exact matching: ATE and ATET

- ATE = the average of these differences weighted by the number of observations in the cells.
- ATET – the number of treated observations in the cells as weights

- 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.
  - Example: 5 confounder variables, each with 5 possible values. $5^5 = 3125$ value-combinations and thus 3125 cells.

- 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 for exact matching: $x = 1$ 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 for exact matching. A value combination is not there in our sample, but could be in the population
- Can we know which one we face?

Coarsened exact matching

- Coarsening variables = joining categories or values to fewer, broader ones
  - creating new categorical variables for those broader categories
  - Fewer bins - more likely match.

- Coarsening exact matching = exact matching on coarsened 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

Coarsened exact matching

- Could be by hand (domain knowledge)
- Could be by algorithm (variance reduction)
  - Packages in Stata and in R (CEM)

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 $z_1, z_2, ...$
- 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.
- So what?

The idea of the common support

- Consequence is general not just for matching
- We shouldn't (cannot) estimate ATE when have no common support.
- So what?

- 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
- Example for quantitative variable, example for categorical variable?

Matching on the Propensity Score

- Idea = creating a single quantitative variable from 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.
- Estimating (=predicting), the probability of $x = 1$ for each and every observation, based on values for the z variables.
  - Estimate a logit, for the probability of $x = 1$, as a function of the confounder variables.

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

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

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.
- Nearest neighbor matching on the propensity score.
- 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

- Match + take difference in y for all $x = 1$ observations.
- The estimated effect of x on y is then the average of those differences.

- If internal validity is high (selection is on observables), we have a good estimate of ATET.
  - With some manipulation, we can get an estimate of ATE, but it's the ATET that we trust more.

Matching on the Propensity Score

- If all confounders are included, the propensity score incorporates all endogenous sources of variation in the causal variable.
- Differences in the z variables with the same propensity score don't make a difference for the estimated effect <– they are not related to the causal variable.

Matching on the Propensity Score

- In practice, many possible decisions...
  - Logit? Probit?
  - Which confounders and in which functional form?
- But once modeling choices are done, it is easy
  - Incl estimate appropriate standard errors by bootstrap.

Technical issues

- We say: Nearest neighbor matching on the propensity score
  - Economists say: Propensity score matching
  - Medical /other say: Nearest neighbor matching
- Stata vs R. Not same results.
  - Defaults vary.
  - Differences due to randomization / bootstrap

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.

- When very different –> likely problem
- If an issue–> matching can help - it takes care of common support by design.

- IF common support exists –> matching, regression: similar results.

Comparing Linear Regression and Matching

- Start analysis by checking common support
- Focus on part of data where it may be found
- Do OLS and matching
- If very different...

Comparing Linear Regression and Matching

- Start analysis by checking common support
- Focus on part of data where it may be found
- Do OLS and matching
- If very different...
- Functional form matters
- Common support not taken care of

Case study: Ownership and management - exact matching

- Take our variables, after condensing some
  - degree nm bins: 4, agecat: 4, competition: 3
  - empbin: 5, industry: 20, countrycode: 24
- theoretical combinations: 4*4*3*5*20*24 = 115,200
- combinations in the data: 6,976
- Firms in both treated and untreated: 1207
- Thus, exact matching is not feasible for about one third of the founder/family owned firms in the data. That's a problem.

- Exact matching yields a difference of -0.155
- If we believe we captured all confounders, this is ATE.
- But: on a very small subsample that is unlikely to be representative.
- Coarsened matching could be a solution. Exercise: try it.

Case study: Ownership and management - NN matching on the propensity score

- Estimated the effect by matching on the propensity score
- Using the same potential confounder variables + same form as for the linear regression.
- Step 1: estimate (predicted) the propensity score by logit
- matched each “treated" (founder/family owned) firm with its nearest neighbor “non-treated" (other owned) firm in terms of the estimated propensity score.
- The average effect estimate is the average of the differences of the matches.

Case study: Ownership and management - NN matching on the propensity score

	(1) All confounders	(2) All confounders interacted with industry and country
ATE estimate	-0.18**	-0.18**
	(0.02)	(0.03)
ATET estimate	-0.20**	-0.21**
	(0.02)	(0.03)
Observations used by logit	8,439	8,223
Number of matched observations	5751	5528
Propensity score model	Logit	Logit

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 : results overview

- We have four point estimate (without employment).
  - Simple difference is -0.39.
  - Regression with controls is -0.19
  - Exact matching is -0.155
  - Propensity score is -0.18
- Overall: no matter which method, once we partial out confounders, estimated ATE is pretty close
  - What does it imply

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.
- For binary variables, proportion of observations =1. Common support = never zero
  - True for baseline regression. (Not when adding interaction)
- Quantitative variables: check range with: min, 5th percentile, the 95th percentile. Common support = range same among the $x = 1$ and the $x = 0$ observations.
  - True in this data.
- Conclude: common support assumption OK in our data
- Main reason why similar results from regression and matching
- What if not

Case study: Ownership and management

- Discussion of findings, causality

Case study: Ownership and management

- Key methodology conclusion (for life):
- if we want to uncover the effect of ownership, different methods won't help much when common support is there and relationships are simple / data is not huge.
- We may need even better data, with more and better-measured confounder variables.
- We may need a different setup, finding exogenous variation in x - come back on week 6.
- Could we do an experiment?

Advanced tools for special situations

- Sometimes data and the setup are such that it allows some nice tricks.
- This is rare *in practice*.

Instrumental variables: the idea

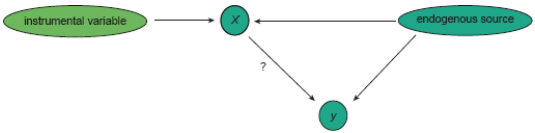
- Suppose there is a common cause confounder, z that is an endogenous source of variation
- If we can observe it, we condition on it (partial it out), and we have a causal identification
- But often, we cannot. If we don't observe it, we cannot condition on it.
- So we look for a trick

Instrumental variables: the idea

- The trick is to find a new variable
- We observe a variable that is an exogenous source of variation in x: Instrumental variable (IV, instrument).
- The IV affects x, but it is not related to y in any other way.
- The IV may be related to observed y through x because it affects x, and x may affect y.
- If y is different across observations that are different in the IV, we know that it's because x affects y.

Instrumental variables

- The IV will uncover the effect of x on y even if x has endogenous sources of variation besides exogenous sources.
- Observe at least one variable that is an exogenous source of variation in x.



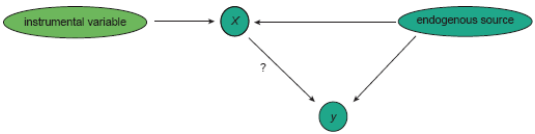
Instrumental variables

- If we can find a variable that is indeed the exogenous source of variation – this is a great way to isolate causality
  - Very elegant
- In practice: this is very hard
  - Very popular method earlier, many academic paper use
  - We have grown more skeptic: hard to find an IV that we believe

Regression discontinuity design

- The idea is close to matching, but applicable for a very narrow set of cases
- Consider a binary causal variable x
- Regression discontinuity design (RDD): comparison of treated or untreated units as determined by a clear rule based on a quantitative variable, called the running variable.
- Treatment is determined by a threshold of the running variable: all subjects are treated on one side of the threshold, and no subject is treated on the other side.

Regression discontinuity design



Note: The question is the effect of x on y; there is an exogenous instrumental variable and another variable that is an endogenous source of variation in x.

Regression discontinuity design

- Example: a tax break given to small enterprises to recover their innovation expenditures if they have, 250 or fewer employees, with no enterprise eligible for the tax break if they have more than 250 employees.
- So: basically compare firms 251-260 vs 240-250 employees
- Very similar firms, on either side of the cut-off

Review of advanced methods to help read papers

- Review of advanced methods to help read papers
- Very hard to them well
- Technically easy
- Hard to find a good application
- Academic use

A lesson

- The data we can use to estimate and effect tends to be more important than the particular method we use.
- Comparing methods could be a good way to find weak spots in the data.

Take-aways

- It is hard to establish causality with cross-sectional observational data.
- Thinking is key, and so is collection a great deal of potential confounders.
- We can rarely condition on all confounders, so our effect estimates are almost always biased.
- Linear regression, various matching methods are alternative ways to condition on observable confounders.
  - With common support, regression and matching tend to give similar results.