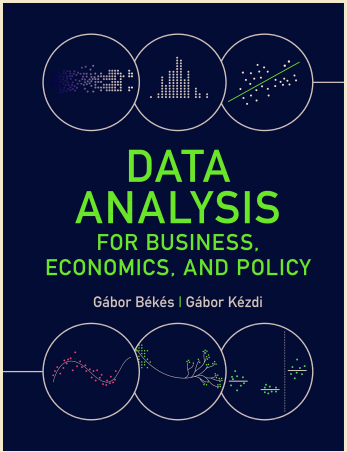
Békés-Kézdi: Data Analysis, Chapter 19: A Framework for causal analysis



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

Causal analysis

- Patterns – associations in the data

- This course: causality
- Most philosophical part
- Model choice is not driven by fit
- Instead: thinking: how to capture the causal link

Causal questions

- How does having a major industrial investment affect house prices?
- Do vitamins have a beneficial health effect?
- Does better management yield greater revenues?
- Does a better diet makes you live longer?
- Does a merger between very large companies cause prices to rise?

Causal questions

- Causal words emphasized.

- How does having a major industrial investment affect house prices?
- Do vitamins have a beneficial health effect?
- Does better management yield greater revenues?
- Does a better diet makes you live longer?
- Does a merger between very large companies cause prices to rise?

Causality implies a role of action 1

- Example for a correlation where there is a causal relationship

- Consider a university classroom in winter
- Correlation:
  - With a window open, the room temperature is lower, on average.
- Causal claim
  - Having an open window make the room temperature be lower than it would be without the window being open
- Causal claim $\rightarrow$ room for action (intervention, policy)
  - Opening the window will cool room temperature.

Causality implies a role of action 2

- Example for correlation but no causal effect

- A barometer shows atmospheric pressure; a drop is associated with low pressure = rain
- The observation of a drop increases the probability of rain. But cannot explain the occurrence of a storm
- How do we know this?
- Consider an intervention on the barometer, putting it into a vacuum chamber (ie no pressure)
  - Set a value randomly
  - Whatever its value, it will not affect the weather
  - The correlation is driven by a common cause, a drop in atmospheric pressure

Causality require variation

- So causality requires the presence of a possible intervention
- Causality also requires variation
- How does taking vitamins effect health?
  - We need people who take and people who do not take vitamins.
- Does better management yield greater revenues?
  - We need firms to have a variation in the quality of management.

Correlation and Causality

- Correlation is not causality

- But we actually can say more.
- If we see evidence of correlation, it's because
  - There is actual causality
  - There is no direct causality, something in the background is going on
  - It's just randomness, and really, no correlation...

The setup: Intervention, treatment, subjects, outcomes

- Intervention describes a decision that aims changing the behavior or situation of people, firms. Also called Treatment.
- Subjects of an intervention are those that may be affected. Treated or untreated.
- Outcome variables, or outcomes, are variables that may be affected by the intervention.
- Causal variables, or treatment variables are the variables that indicate the intervention.
- Need idea why the intervention may affect an outcome variable. Mechanisms by which an intervention exerts an effect on a particular outcome variable or variables.
  - Other names for mechanisms: pathways or mediator variables

The causal question

Most important elements of a precise causal question are

- What's the outcome (Y) variable?
- What's the causal (X) variable?
  - Binary (intervention Y/N) or a quantitative (amount of intervention).
- What are the subjects (the outcome for whom?)
- What is the specific intervention (who, and how, would manipulate the cause to alter the outcome?)
- What is or could be the mechanism (why should one expect an effect of the intervention on the subject?).

Data to answer causal questions

- Use already existing data
- Collect data – run experiments

- Throughout lecture, will discuss several examples
- One small case study at the end

Example 1 Dietary supplements and health

- Consider the potential effect of dietary supplements on long-term health.
  - Dietary supplements are vitamins, minerals and other chemicals
- Example: beta-carotene, an anti-oxidant found in abundance in carrots.
  - Long-term health may be measured by how long one lives.

Example 1 Dietary supplements and health

- Research question: how much longer people can expect to live if they take beta-carotene supplements compared to how long they can expect to live id they don't take them.

Aspect	Dietary supplement and health example
Outcome variable	The length of life, or, to relate it to existing data, the probability of survival to a certain age
Causal variable	Taking beta-carotene supplement (pill) regularly
Subjects	A person
Intervention	Analyst deciding on asking people to take the pill.
Mechanism	anti-oxidants neutralizing free radicals in our bodies (anti-cancer).

Example 1 Dietary supplements and health

- Research question: how much longer people can expect to live if they take beta-carotene supplements compared to how long they can expect to live id they don't take them.
- Turns out beta-carotine supplement makes no difference

https://www.mountsinai.org/health-library/supplement/beta-carotene

Example 2: TV ads and product purchase

- A firm that wants to place online advertising to induce viewers of such advertising to buy its product.
- We look at this from the point of view of the individuals.

- Each individual either sees the ad or does not see the ad.
- If the individual sees the ad she may click on a link, visit the website of the firm, and may or may not purchase the product.
- Of course the individual may purchase the product without seeing the ad, too.

Example 2: TV ads and product purchase

- Research question: Does showing an ad makes people more likely a particular product (in a short time period)

Aspect	TV ads and product purchase example
Outcome variable	Indicator of the person purchasing the product or not.
Causal variable	Indicator of a person is presented with the ad or not (treated units are people who are presented with the online advertising. Untreated people are not presented with the ad.)
Subjects	Persons
Intervention	Placing the advertisement
Mechanism	By seeing an ad people may be induced to think that they need to buy that particular product here and now.

Modeling causality: PO and DAG

Potential outcomes framework

- Potential outcomes framework is a structure to study causal questions.
- Thinking in this framework will make defining the effect of an intervention straightforward.
- The outcome variable Y, may be
  - Binary: whether an individual buys the product or not
  - Quantitative: the sales value of a house.

Potential outcomes framework

- Binary interventions: subjects may be either treated or untreated.
  - The outcome may be anything, including binary or multi-valued variables.
- Can always think about two potential outcomes for each subject:
  - what their outcomes would be if they were treated (their treated outcome),
  - what their outcomes would be if they were untreated (their untreated outcome).

Potential outcomes framework

- Of these two potential outcomes, each subject will experience only one: that's their observed outcome.
  - Treated subject: Observed outcome = their treated outcome.
  - Not treated subject: Observed outcome = their untreated outcome.
- The other potential outcome, unobserved, is their counterfactual outcome
  - what could have been observed had the subject experienced what did not happen.

Potential outcomes framework

- Each subject has two potential outcomes before the intervention, both unobserved.

- Then each subjects gets assigned to be treated or untreated.
- The intervention reveals one of their potential outcomes, the one that conforms their assignment.
- Their other potential outcome remains unobserved = counterfactual outcome.

The Individual Treatment Effect

- The individual treatment effect for subject i is the difference between their two potential outcomes: the value of the potential treated outcome for the subject minus the value of the potential untreated outcome:

$$te_i = y_i^1 - y_i^0 \tag{1}$$

  - $y_i =$ observable outcome
  - $y_i = y_i^1$ for subjects that end up being treated
  - $y_i = y_i^0$ for subjects that end up being not treated

Individual treatment effects

- te_i= the value of the treated outcome for the subject minus the value of the untreated outcome for the same subject i.
- te_i may be 0, positive or negative
- Consider binary outcomes (0 or 1), so the ITE=[0,-1,1].
  - $te_i = 1$ if the treated outcome is one and the untreated outcome is zero.
  - $te_i = -1$ if the treated outcome is zero and the untreated outcome is one.
  - $te_i = 0$ if both the treated outcome and the untreated is one, or both of them is zero.

Individual treatment effects

- Individual treatment - think cause and effect without observing them.
- The individual treatment effect is never observable.

- There is no way to know
  - what the outcome of untreated subjects would have been if they were treated,
  - what the treated outcome of untreated subjects would have been.
- Analyst cannot uncover individual treatment effects by observation.

Heterogeneous treatment effects

- Individual treatment effects will vary, of course.
- heterogeneous treatment effects.
- Can't observe te_i – will not know if indeed heterogeneous among the subjects we care about.

- ITE may vary across groups
  - Men vs women
  - Small vs large markets
- We can actually look at it (Case study on Week 4)

Example 1 Dietary supplements and health

- Example of taking vitamins or supplements regularly
- What is the te_i in this example, and can it be heterogeneous?

Average treatment effect

- Instead of ITE (te_i), we can observe the average
- ATE = average treatment effect = average of the ITE across all subjects.
- For binary outcomes (e.g. buy product or not)
  - average outcomes are probabilities
  - ATE are differences in probabilities.

ATE as average / expected ITE

- ATE is the expected (=average) difference between potential outcomes
  - Expectation operator ($E[]$)

$$ATE = E[te_i] = E[y_i^1 - y_i^0] \tag{2}$$

- The average of the differences is equal to the difference of the averages.
- –> ATE is also the difference between the average of potential treated outcomes and the average of potential untreated outcomes:

$$ATE = E[y_i^1] - E[y_i^0] \tag{3}$$

Average treatment effect

- Policy: "The effect" –> ATE in mind
- ATE = expected effect of the intervention for a subject randomly chosen from population.
- ATE = total effect of the intervention if multiplied by the population size

- ATE may be estimated in the right setup

Average Effects in Subgroups

- Heterogeneity may be hidden behind the ATE.
- Consider $ATE = 0$:
  - all individual treatment effects are all zero.
  - the intervention has positive effects on some subjects and negative effect on other subjects but those cancel out.
- Any value may conceal a division of groups of subjects with very high and low effect.

ATE when Quantitative Causal Variables

- Examples of interventions that lead to quantitative causal variables
  - setting prices of products or services;
  - deciding on the budget to be spent on advertising through a social media platform.
- PO framework - designed binary interventions.
- Concepts apply to quantitative causal variables
- But more complicated

Quantitative Causal Variables

- A quantitative causal variable - the intervention is not binary (happens to you or not), but the effect size varies by subject
- Many individual treatment effects beyond (0,1).
- Many potential outcomes for each subject (beyond -1,0,1)

ATE and Quantitative Causal Variables

- Quantitative causal variables (numeric x) lead to not one individual treatment effect but a series of them,
- One more step: average individual treatment across possible values of x before taking the average across subjects for ATE.
- Difficult to think about average effects of quantitative causal variables.
- But the idea is fundamentally the same.

- Often use quantitative variable and create a binary: low vs high

Example: Price and sales in a pub

- You run a pub, question is: Pricing a new IPA (beer), effect of price.
- Change price p monthly, observe monthly sales q
- Pub level sales at different months are the "subjects" here

- PO: in a specific month t, you can sell q_t you could sell if charged various prices p_t.
- ITE: how much more beverage the company would sell in a particular month if it charged a particular price, compared to how much it would sell if it charged another price.
- Using a simple benchmark: the difference in q (in percentage terms) if p were different by one percent. = Price elasticity.

Ceteris Paribus: Other Things Being the Same

- What we really mean by potential outcomes.
- The difference between treated and untreated outcome is the intervention and only the intervention.
- All other things that may affect the outcome variable are the same.
  - relevant things that may cause the outcome variable to change besides the intervention.
- "all other (relevant things) being the same" = "ceteris paribus".

Ceteris paribus vs multivariate regression

- Remember DA2 (Chapter 10), with outcome y, causal variable x

$$y^E = \beta_0 + \beta_1 x + \beta_2 z \tag{4}$$

- In regression we condition on z
- Compare two observations that have the same z but are different in x by one unit. The observation with a one unit higher x is expected to have β_1 units higher y.

Ceteris paribus vs multivariate regression

- Can we condition on all potential confounders in regression?

Ceteris paribus vs multivariate regression

- Can we condition on all potential confounders in regression?
- That would be ceteris paribus analysis
- Probably not
  - Can include only what we observed in data
  - Can never be sure we if left out an important confounder

- In a regression, we compare observations that differ in x and are same in all other RHS variables that we observe and include in the regression

Average treatment effect

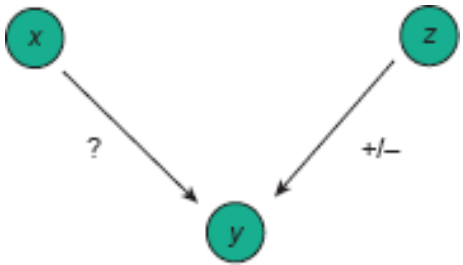
- How to calculate ATE - main issue for this course
- Because te_i cannot be calculated and averaged
- Because ceteris paribus exists as a theoretical concept and need to work hard to get close

Causal maps to uncover causal structure

- Start to think through cases, understand how causality may work
  - Summarizing our assumptions about how variables affect each other.
- Causal maps: key tool to think about causality
- A causal map = graph connecting variables, showing direction of causality
- Another name for causal map is directed acyclic graphs, DAG - graph of nodes and arrows.
  - also: causal map, graph, diagram

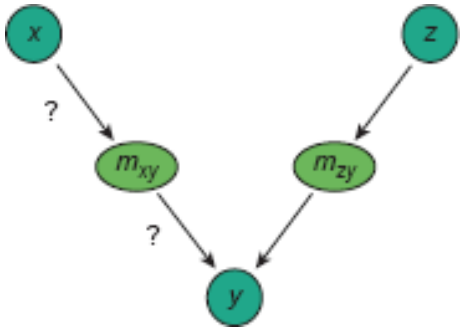
DAG: simplest case

- Example: an outcome variable is caused by the intervention of interest (x) but also other variables like z
  - x is causing y,
  - z is causing y.
  - z is unrelated to x



DAG: mechanisms

- Add variables that measure the mechanisms (m) through which x and z affect y.
- m_{zx} = through which x affects y
- m_{zy} = through which z affects y.



Example 2: TV ads and product purchase

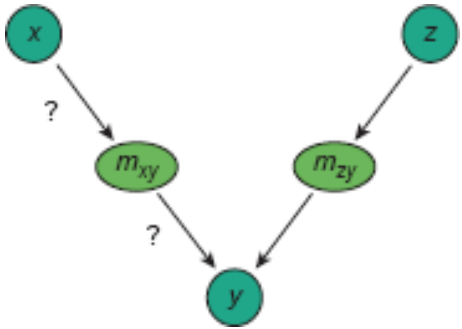
- Potential outcomes = 0 (no purchase) or 1 (purchase).
- $te_i = -1, 0, 1$
- ATE = is change in the likelihood of purchasing the product due to seeing the ad.
  - That is the combination of the three possible treatment effects $(1, 0, -1)$.
- The higher the proportion of people with treatment effect 1 the more positive the average effect.

Example 2: TV ads and product purchase

- Without being presented the ad, 10% of the subjects would buy the product.
  - The untreated PO=1 for 10% of the subjects and 0 for 90%.
- If presented the ad, 11% of the subjects would buy the product.
  - the treated PO= 1 for 11% of the subjects and 0 for 89%.
- The average treatment effect here is 1 percentage point: $ATE = 0.01$.

Example 2: TV ads and product purchase

- Causal variable, x?
- Outcome variable, y
- Mechanism m_{xy}
- Confounder variable, z?



Randomization to get ATE

Comparing Different Observations to Uncover Average Effects

- PO, DAG frameworks - think more precisely about the effect we want to measure.

- But: te_i cannot be measured
- = Counterfactual outcome ("what would have been") is never observed

- Observable are:
- The (potential) treated outcome (y_i^1) for subjects treated.
- The (potential) untreated outcome (y_i^0) for subjects not treated.

ATE, Potential outcomes, observed outcomes

- ATE = what we are interested in
- ATE = Difference between the average of potential treated outcomes and the average of potential untreated outcomes

- We have average difference between observed outcomes

Comparing Different Observations to Uncover Average Effects

- Uncover average potential outcomes from the average observable outcome IF two good approximations.
  - Average of the observed outcomes for treated subjects ($E[y_i|\, iistreated]$) $\approx$ the average of the potential treated outcomes across all subjects.
  - Average of the observed outcomes for untreated subjects ($E[y_i|\, iisnottreated]$.) $\approx$ the average of the potential untreated outcomes across all subjects.

$$E[y_i|\, iistreated] \overset{?}{\approx} E[y_i^1] \tag{5}$$

$$E[y_i|\, iisnottreated] \overset{?}{\approx} E[y_i^0] \tag{6}$$

Comparing Different Observations to Uncover Average Effects

- Message: Data helps uncover ATE the closer observed groups represent theoretical concepts of PO.

Random assignment

- How can we get data where these assumptions would hold?

Random assignment

- How can we get data where these assumptions would hold?
- The random assignment condition = assignment is independent of potential outcomes
  - whichever subject ends up being treated or untreated is independent of their potential outcomes
- Random assignment == independence of potential outcomes.
  - Not about how the data was collected (unfortunate name)

Random assignment and ATE

- Independence = treated and untreated groups are similar in terms of their potential outcomes, on average
- If yes –> simple way to get a good estimate for ATE.
- If assignment is random, the difference between average observed outcomes of treated versus untreated subjects is a good estimate of ATE.
- Random assignment – theoretical concept – aspiration to get good ATE estimate.

Random assignment, ATE and ATET

- Random assignment: observed difference is good estimate of ATE as well as ATET.
- Because, in this case, ATE and ATET are equal.
  - ATET = ATE on treated
- Random assignment –> those end up being treated = (in terms of their potential outcomes) = entire population.

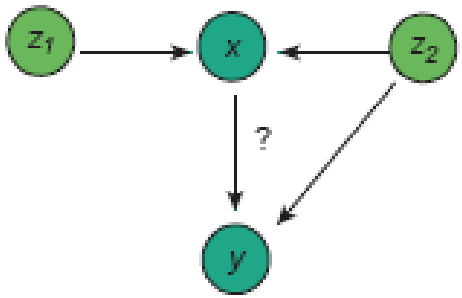
Opportunity and threat to causal identification

Sources of Variation in the Causal Variable

- Sources of variation in the causal variable - thinking task
- An endogenous source of variation is when the source of variation in x is also related to y.
- An exogenous source of variation is when a source of variation that affects x is independent of y.

An exogenous and an endogenous source of variation in x

- Assumption 1: z_1 is an exogenous source of variation in x;
- Assumption 2: z_2 is an endogenous source of variation in x.



Sources of Variation in the Causal Variable

- Random assignment and exogeneity in the source of variation are close concepts.
- When assignment is random, there are only exogenous sources of variation in x.
- When assignment of x is not random, there are likely to be endogenous and exogenous sources of variation

Good and bad sources

- For the question of the effect of x on y, we need to assess all things that may make x vary across observations, and then divide them into
  - good ones (exogenous) and
  - bad ones (endogenous).
- To uncover the effect we'll need to keep the good ones and get rid of the bad ones.
- Next bits + most of the course is about how to do that.

Experimenting versus Conditioning

- Several ways to uncover causality
- Experimenting versus Conditioning

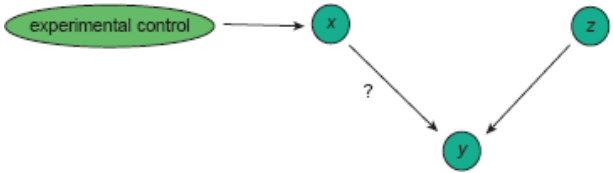
Experimenting versus Conditioning: 1 Controlled experiments

- Controlled experiments allows for controlling variation in the causal variable
- Variation in the causal variable x is controlled by assigning values of x to the observations.
- The intervention is hence done by the analyst

- This practice is called controlled assignment.
  - attempts to make sure that the value of x observations "receive" is not affected by the decisions of people who may be interested in the outcome.
  - It can also help avoid reverse causality by not letting the outcome y affect x in any way.
- If binary treatment x variable observations are assigned to a treated and an untreated ("control") group by the analyst.

Controlled experimental variation in x

- Experimental control is the only source of variation in x.
- Other variables,summarized by z, may affect y but are unrelated to x.



Experimenting versus Conditioning

- Sometimes controlled experiments are impossible, impractical, or would produce uninformative results,
- This is when data analysts will have to resort to using observational data.

Experimenting versus Conditioning: 2 Natural experiments

- In natural experiments - may assume that variation in x in observational data is exogenous,
  - ... as if it came from a controlled experiment.
- Natural experiments do not have experimenters who assign treatment in a controlled way.
- Assume that assignment in a natural experiment took place as if it were a well-designed controlled experiment.

- Key is indeed exogenous variation in x
- Example: Natural disasters, geography

Experimenting versus Conditioning: 3 Conditioning on observables

- Most often, no natural experiment situation
- Conditioning on endogenous sources of variation in the causal variable.
  - conditioning on the values of variable z when comparing the values of y by values of x.
  - Let exogenous sources vary AND, not let endogenous sources vary.
- Comparing observations that are different in terms of exogenous sources of variation in x, while having similar values for the variables that are endogenous sources of variation.
- Why need difference in exogenous sources of variation in x?

Experimenting versus Conditioning: 3 Conditioning on observables

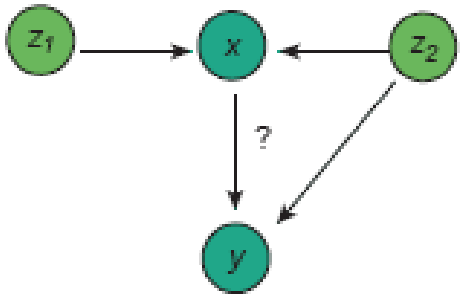
- Most often, no natural experiment situation
- Conditioning on endogenous sources of variation in the causal variable.
  - conditioning on the values of variable z when comparing the values of y by values of x.
  - Let exogenous sources vary AND, not let endogenous sources vary.
- Comparing observations that are different in terms of exogenous sources of variation in x, while having similar values for the variables that are endogenous sources of variation.
- Why need difference in exogenous sources of variation in x?
- Conditioning = isolating exogenous sources of variation in x

Confounders in Observational Data

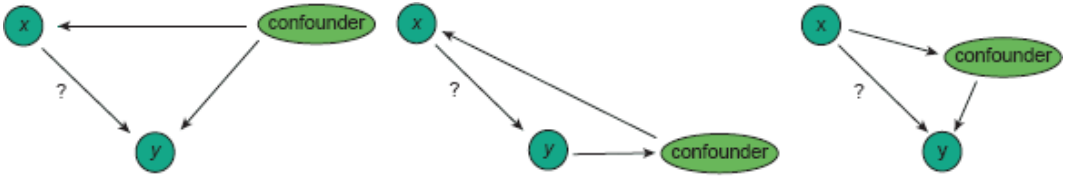
- Confounding variables (confounders) in observational data
  - endogenous sources of variation in a causal variable
- The key issue to think about when doing causal analysis with observational data

Confounders in Observational Data

- z_2 is an endogenous source of variation in x.
- Makes y and x correlated even though x not cause y and y not cause x.



Three types of confounders



(a) Common cause confounder

(b) Mechanism of reverse causality

(c) Unwanted mechanism confounder

Common cause confounder

- When we speak of confounders we often mean common cause confounders
- z affects y
- z also affects x
- Examples could be income, education affecting several choices and conditions of people

Mechanism of reverse causality

- The outcome variable y itself may affect the causal variable x: reverse causality.
- Here y affects x when, instead, we are interested in the effect of x on y.
- This reverse causality operates via the mechanism of z.
- Example, if sales are going down the management of the firm may want to reverse that negative trend by advertising more.
- How would a DAG look like?

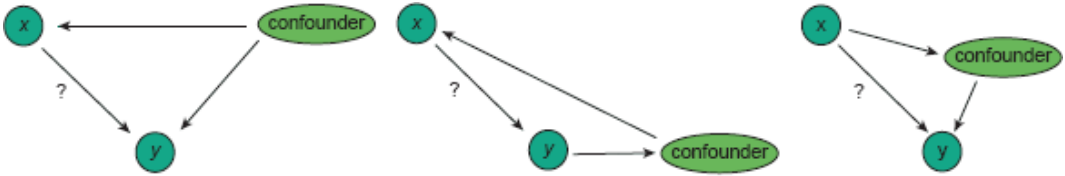
Reverse causality

- Even more complicated: feedback loop
- That may induce feedback loops: x affecting y, then y affecting x in turn, and so forth.
- Positive feedback loops reinforce the original effect of x; negative feedback loops diminish its effect.

Unwanted mechanism

- Unwanted mechanism confounder: a mechanism through which x affects y, but one that we want to exclude.
- Not actually a source of variation in x, but we want to condition on it nevertheless.
- It could be a mechanism of selection, that we want to exclude
  - Hard, more later...

Three types of confounders (repeated)



(a) Common cause confounder

(b) Mechanism of reverse causality

(c) Unwanted mechanism confounder

Bad Conditioners: Variables Not to Condition On

- So we have great deal of variables in our data
- To be on the safe side, we should condition on all available variables in our data?

Bad Conditioners: Variables Not to Condition On

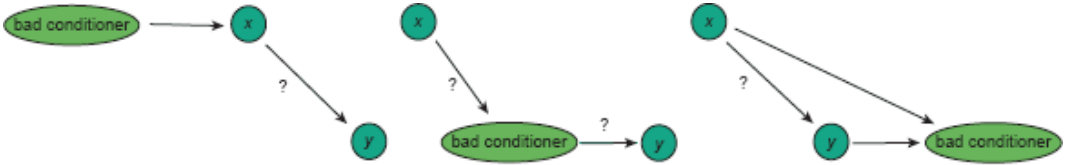
- So we have great deal of variables in our data
- To be on the safe side, we should condition on all available variables in our data?
- No
- There are variables that we should not condition on when trying to estimate the effect of x on y. Bad conditioning variables.

The three types of bad conditioning variables

- exogenous source of variation in the causal variable x.
- part of the mechanism by which x affects y – that is of course if we want to include that mechanism in the effect we want to uncover
- collider variable: a common effect, or common consequence, of both x and y;

- How to know if we should condition on a variable or not?
- Analyst must think and decide
- Causal map (DAG) helps

The three types of bad conditioning variables



(a) Exogenous source of variation (b) Mechanism variable (c) Common consequence (collider variable)

The three types of bad conditioning variables

- exogenous source of variation in the causal variable x.
- part of the mechanism by which x affects y – that is of course if we want to include that mechanism in the effect we want to uncover
- collider variable: a common effect, or common consequence, of both x and y;

- If you believe you have such variables, do NOT add them to a regression

From Latent Variables to Measured Variables

- From Causal map to data: latent and missing variables
- Causal map to data: two problems: (1) hard to measure, (2) not available.
- Confounders that we want to condition on are not directly measurable = latent variables.
- Variables in real data are often imperfect measures of the latent variables that we want to consider.

From Latent Variables to Measured Variables

- Real data rarely includes variables that measure all of the confounders.
- Failing to condition on some of the confounders, or conditioning on imperfect measures of them, leads to a biased estimate of the effect.
- This is the Omitted variable bias

Case study: Food and health: data

- You are what you eat
- causal statement: some kinds of food make you healthier than other kinds of food.
- Does eating more fruit and vegetables help us avoid high blood pressure?

- Case study briefly in lecture, please read details

Case study: Food and health

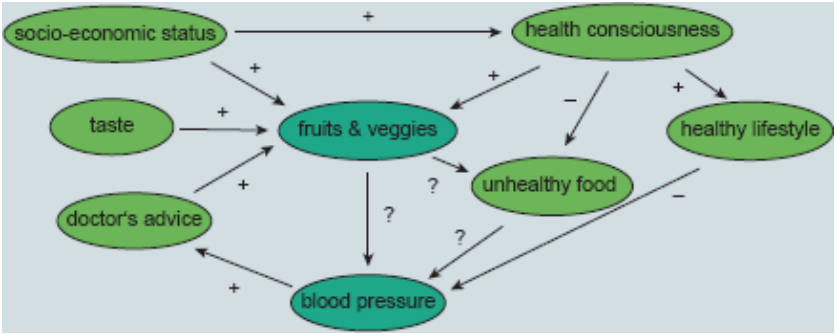
- The food-health dataset we use comes from the National Health and Nutrition Examination Survey (NHANES) in the United States.
- The amount of fruit and vegetables consumed per day and blood pressure
  - Measured by an interview that asks respondents to recall everything they ate in two days.
- Blood pressure is sum of systolic and diastolic measures.
- Fruit and vegetables is the amount consumed per day (g)
- Source: food-health dataset, USA,
- ages 30–59, 2009–2013. N=7358.

Case study: Food and health – descriptive statistics

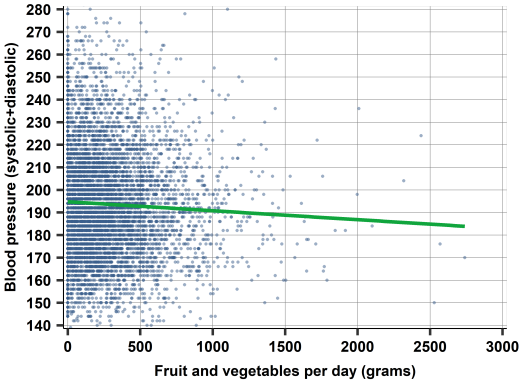
	Mean	Median	Std.Dev.	Min	Max	Obs
Blood pressure (systolic+diastolic)	194	192	24	129	300	7359
Fruit and vegetables per day, grams	361	255	383	0	3153	7359

Source: food-health dataset, USA, ages 30 to 59, 2009–2013.

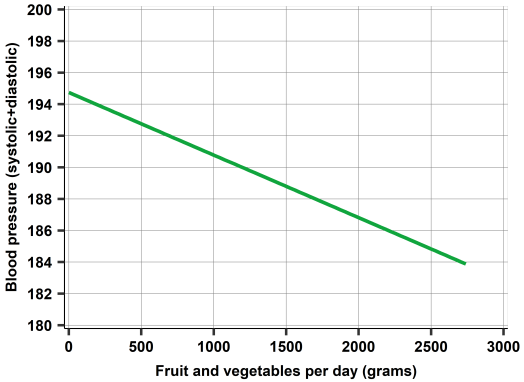
Case study: A causal map - effect of fruit and vegetables on blood pressure



Case study: Food and health- Correlation

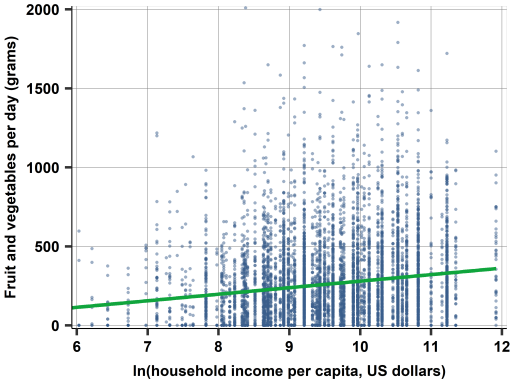


Scatterplot and regression line

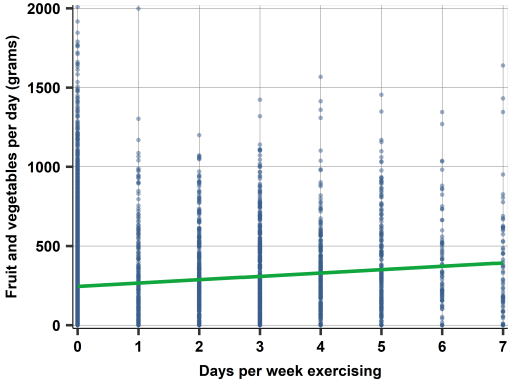


Regression line only

Case study: Food and health- two sources of variation in eating veggies



Log household income and amount fruit + vegetables



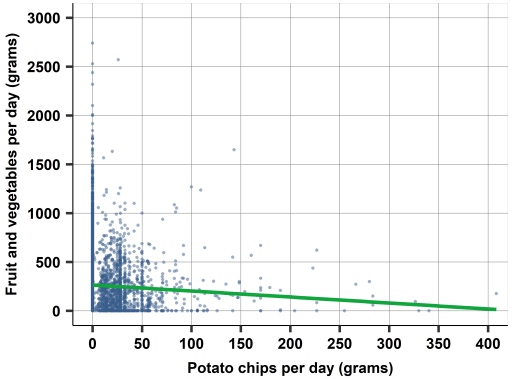
Days/week exercising and amount fruit + vegetables

Case study: Food and health- Consumption of an unhealthy food item

- Chips consumption. Should we condition on?
- Yes. Chips eating is a common cause. Chip eating signal unhealthy diet could affect chance of veggies and health

Case study: Food and health- Consumption of an unhealthy food item

- Chips consumption. Should we condition on?
- Yes. Chips eating is a common cause. Chip eating signal unhealthy diet could affect chance of veggies and health
- No. A potential bad conditioning variable: Veggie eating causes less chips that causes better health. Unwanted mechanism.



Summary

- Food and health correlated
- Many potential confounders
- Never be really causal
- But can offer insight and prompt experiments
- Can be informative - more likely causally true than not.

Comparing pros and cons of approaches

- Causality can be established
  - Controlled experiment = great confidence
  - Natural experiment = good confidence, but work is needed to prove it
  - Conditioning on confounders = never be certain.
- This is about internal validity
  - The extent of which we can be certain that indeed, we uncovered a causal relationship

External validity

- However, there is another aspect
- External validity is measure of confidence about generalization
  - Will the causal relationship work in the future
  - Will the causal relationship work in other markets, countries
- Key issue throughout the course is discussing internal and external validity
  - Often a trade-off

Constructive skepticism

- No analysis is perfect
  - Weigh pros and cons of different approaches
- One can still learn from a well-designed analysis
  - Be that a controlled experiment or an observational study
- Solid knowledge from many studies
  - With different approaches
  - Pointing to similar conclusion if biases well understood
    - some studies mar be more biased than others
    - Need to take into account when summing up evidence from multiple studies

It pays to be specific - Contrast question to data

- The outcome and cause variables in the data and in the question
  - the more similar they are the more relevant our findings and the less we have to worry about measurement errors and the like.
- The observations in the data may be close measure of the subject or not
  - The more similar they are the more relevant the findings may be for the question.
- The reason for the causal variable to vary in the data may or may not be due to interventions one is interested in;
  - the more similar they are the more likely that they represent the causal relationship in question.
- Focusing on a specific mechanism
  - helps assess whether the relationship found in the data may be caused by such a mechanism.