Causal Data analysis

Chapter 21: Regression and Matching with Observational Data

Álmos Telegdy

Corvinus University of Budapest

Causality with observational data

Most of the analysis is based on **observational** (real) data

Experimental analysis has better internal validity, but such data are very hard to produce (sometimes impossible)

- Expensive, time consuming, unethical
- External validity open to doubt

How to estimate causal relationships with observational data?

- It is about the interpretation of the results ⇒ **under what conditions can we interpret the estimated relation as causal?**

Think before running running regressions!

Make a **thought experiment**

- Who are the subjects?
- What is the treatment, how good a proxy is the treatment variable?
- What is the treatment mechanism?
- What kind of endogeneity may arise?
  - What are the mechanisms, which move both x and y?
  - Is there a proxy, which may (partially) control for this endogeneity?

Methods to control for endogeneity

The most obvious solution: **multivariate regression analysis**, where one **controls** for the endogenous variables

- What control variables should we use?
    - In what functional form?
- What to do when the control and treated groups are very different?
    - Matching (always applicable)
    - Go for better data (panel data regressions)
    - Other methods (excellent internal consistency, but rarely applicable)
        - Instrumental variables
        - Regression discontinuity design

Measured effect

- Linear regression

$$y_i = \alpha + \beta x_i + u_i$$

- $u_i \rightarrow$ contains the unobserved heterogeneity

- Why y differs across subjects?
  1. Causal relationship between x and y: the variability of y is partially determined by the variability of x
  2. Other reasons: *unobserved heterogeneity* ("selection")

MEASURED EFFECT = CAUSALITY + SELECTION

Unobserved heterogeneity

- Unobserved heterogeneity is present in **every** regression analysis
  - The variability of the outcome variable is caused by many variables

- The aim of causal analysis is to separate the effect of interest from all other effects → **identification**

Control variables (recap)

Aim: measure the effect of exogenous variation of x on y

- Control for all possible endogenous variation (never possible in practice)
    - Similar cause: $z \to x, y$
    - Reverse causality: $y \to x$
    - Unwanted mechanism: $x \to y$, but not through the desired mechanism
- Do not control for the following variables:
    - z has an exogenous effect on x
    - z is part of the mechanism, which explains the relation between x and y
    - z that is affected by both x and y

How to identify adequate control variables?

Find the exogenous and endogenous variability in the treatment variable and translate it to control variables

Causal map → confounding variables → latent variables → actual variables

Unobserved heterogeneity cont.

Unobserved heterogeneity has three components:

Some variables affect y...

1. ...and they are in the data (*known knowns*)
2. ...but are not in the data (*known unknowns*)
3. ...but we did not even think about them (*unknown unknowns*)

Example: food and health

The relation between fruit and vegetables consumption and high blood pressure. What is the effect of health consciousness?

- Proxy variables for health consciousness: smoking, daily exercise
  - Not included: hours of sleep, level of stress (good levels in the health conscious group)

Example: food and health

The relation between fruit and vegetables consumption and high blood pressure. What is the effect of health consciousness?

- Proxy variables for health consciousness: smoking, daily exercise
    - Not included: hours of sleep, level of stress (good levels in the health conscious group)
    - Excluded variable: → **negative bias**
        - The measured effect is stronger than the real effect
- Medical advice
    - High blood pressure → eat more healthy food

Example: food and health

The relation between fruit and vegetables consumption and high blood pressure. What is the effect of health consciousness?

- Proxy variables for health consciousness: smoking, daily exercise
  - Not included: hours of sleep, level of stress (good levels in the health conscious group)
  - Excluded variable: → **negative bias**
    - The measured effect is stronger than the real effect
- Medical advice
  - High blood pressure → eat more healthy food
  - Reverse causality → **positive bias**
    - the measured effect is weaker than the actual effect

We do not know whether healthy lifestyle induces a positive or negative bias

Functional form

How to add control variables to the regression (e.g., level, logged, dummy, polynomial, interactions)

1. If the inclusion does not affect the coefficient of x
   - Use the simplest specification (e.g., experimental data)
2. If the inclusion changes the coefficient of x
   - Take out as much unobserved heterogeneity as possible
   - Use many control variables
   - Functional form should be parsimonious
   - BUT: do not use bad controls
   - BUT: too many controls may decrease the precision of the estimation

Case study: family-owned companies

Most firms (even the large ones) are under family ownership

General question: is this ownership form more or less efficient, than other forms (e.g., company ownership, dispersed ownership)?

Good causal question: do family-owned firms have better management than firms under other types of ownership?

First best: experimental data

Is this feasible?

- Impossible (econspeach: prohibitively expensive): owners will not sold their companies randomly
  - → Random assignment not feasible
- Quasiexperimental data: takeovers
  - These are rare events → analysis of ownership change not realistic

Realistic option: compare firms under family and non-family ownership

Data

- Source: World Management Survey
- Cross sectional firm-level data from 21 countries
  - Representative within country
- Sample: 7,506 firms

Variables, sample

- Outcome variable: **management index**
  - The average of 18 management techniques scores
  - Each technique was given a score (1 – very bad; 5 – excellent)
- Treatment variable: **family ownership**
  - Dummy variable
- Other variables: Employment, share of collage grads, medium/high competition on product market, industry, country
- Most firms are born as family-owned, but some are sold to outside investors
  - State-, foundation-, worker-owned firms excluded
  - Firms with 50–50,000 employees kept in the sample
  - Final sample: 6,137 firms

Potential control variables

- Past of the firm – founded as a family firm or not

Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital

Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital (end.)
- Culture, institutions

Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital (end.)
- Culture, institutions (end.)
  - Some countries have more family-owned firms than others
- Firm characteristics (age, size)

Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital (end.)
- Culture, institutions (end.)
  - Some countries have more family-owned firms than others
- Firm characteristics (age, size) (end., reverse causality)
  - Older firms have a higher chance to become outside-owned
  - Large companies more likely needed outside capital
- Family characteristics

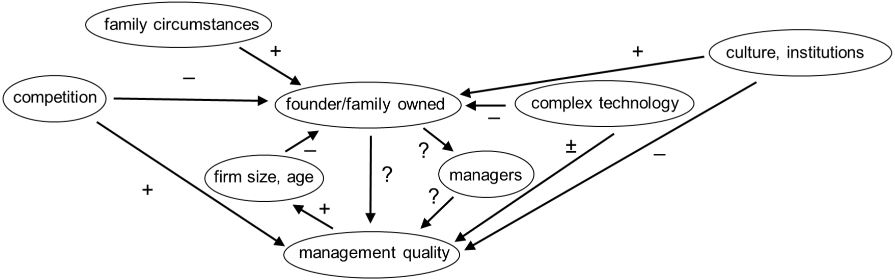
Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital (end.)
- Culture, institutions (end.)
  - Some countries have more family-owned firms than others
- Firm characteristics (age, size) (end., reverse causality)
  - Older firms have a higher chance to become outside-owned
  - Large companies more likely needed outside capital
- Family characteristics (ex.)
  - Number/gender of children affects inheritance
- Product market competition

Potential control variables

- Past of the firm – founded as a family firm or not (endogenous)
- Technology – some production modes need a lot of capital (end.)
- Culture, institutions (end.)
  - Some countries have more family-owned firms than others
- Firm characteristics (age, size) (end., reverse causality)
  - Older firms have a higher chance to become outside-owned
  - Large companies more likely needed outside capital
- Family characteristics (ex.)
  - Number/gender of children affects inheritance
- Product market competition (end.)
  - High competition increases the likeliness of takeover

Causal map


family circumstances
+
culture, institutions
+
–
competition
founder/family owned
complex technology
–
–
?
±
–
firm size, age
?
managers
+
?
+
management quality

Latent variables → proxy variables

- Technology → **industry** (20 cat.), **share of collage grads** (level+squared)
- Culture, institutions → **country** (24)
- Firm attributes → **employment** (log), **age** ($\leq$ 30, 30-80, $\geq$ 81), missing for 14%
- Product market competition → # **competitors** (4 cat.)

Bad controls

- CEO experience – this is part of the mechanism
- Exports' share in sales – measures success

Regression results

	No conf.	With conf.	With conf. 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)
N	8440	8439	8439
R-squared	0.08	0.29	0.37

$\beta = -0.19$ – approx. 30% of the depvar's variance

Can we interpret the result as causal?

No. Data far from ideal → we know very little of the selection mechanism

Can we interpret the result as causal?

No. Data far from ideal → we know very little of the selection mechanism

Direction of bias? (= what we don't control for?)

- Underperforming firms more likely sold to outside owners → reverse causality
    - Better firms remain in family ownership → we overestimate the effect of family ownership

Values of control variables

Ideal regression analysis: compare the average value of the outcome variable in the control and treated groups such that *all control variables have identical values in the two groups* (ceteris paribus)

- BUT nothing prevents the regression to compare firms which do not have similar control values
- e.g., average firm size may be different between the control and treated firms (domestically-owned; treated = foreign-owned)

In principle, it is possible to compare subjects which **have exactly the same attributes** (but the treatment)

- In practice not always possible

Exact matching

We compare subjects with control variables having the same values

- Take *treated* subject i: $x = 1$, control variables $z_{1i}, z_{2i}, \ldots$
- Find a *control* subject j with the following attributes: $x = 0$, $z_{1i} = z_{1j}, z_{2i} = z_{2j}, \ldots$
- Compute the difference in the outcome variable y between i and j
- Do this for all treated subjects
- Average out the differences. This is the estimated difference between the treated and control groups *exactly matched* on the $z_1, z_2, \ldots$ variables

Example: foreign ownership of soccer clubs

What is the effect of foreign ownership on European soccer clubs?

- Control variables: country, division, city size

Exact matching

- Compose groups by country (20), division (2), city size (5)
- 200 groups

Example cont'd.

- Compute club performance by ownership type *within* these groups:

$$E[y|x=1, z_1 = z_1^*, z_2 = z_2^*, ...] - E[y|x=0, z_1 = z_1^*, z_2 = z_2^*, ...]$$

- Count the number of treated and control observations
- Exact matching estimation
  - **ATE** = the average difference weighted by the number of observations in a cell (we want to compute the effect in the whole data)
  - **ATET** = the average difference weighted by the number of treated observations in a cell (we want to compute the effect on the treated observations)

When is exact matching feasible?

- Each treated group must have a corresponding untreated group
  - Does not need to be true on the other way around: there may exist untreated groups without a corresponding treated group
- This condition doesn't always hold: the data generating process may not allow it
  - E.g., German soccer clubs cannot be foreign owned
  - In general, treated and control groups' $z's$ can be very different $\rightarrow$ the data may not have such subjects
  - E.g., 5 controls, each taking 5 values: $5^5 = 3125$. Even in large datasets some of the cells are empty

Coarsened exact matching

- **Coarsening qualitative variables:** 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 firms
- 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

P-score matching

Exact matching it not always feasible, especially if there are many z_k control variables.

Solution:

- Create a **single quantitative variable from all the z_k confounder variables**
  - We reduce a multidimensional problem to one-dimensional problem
- We execute matching with this new variable: we match treated and control subjects that have similar values of this variable
- The most popular matching method: **propensity score matching**
- P-score is a *conditional probability*: the probability that an observation is treated ($x = 1$), conditional on all z_k confounders
  - The p-score is a scalar (probability), which embeds the effect of all $z's$ on whether a given observation is treated or untreated

Obtaining the P-score

P-score is unknown – we need to estimate it

- Estimate a probability model (logit or probit) with the dependent variable x, explanatory variables $z_1, z_2, ...$
- Predict $x^P \Rightarrow \hat{x}^P$ is the estimated probability of treatment

The estimation function (logit):

$$x^P = \Lambda(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + ...) \tag{1}$$

Identification assumption: **p-score is the probability of the treatment** and it embeds the endogenous variation of the causal variable

Nearest neighbor matching

Most commonly used method: **nearest neighbor matching**

- Find treated and control observations with similar p-score
- Assign to each treated subject the control subject which has the closest p-score
- If many control subjects have similar p-scores, take all into account and compute the average of the outcome variable
- Compute the difference in y within groups
- Average out the differences

Common support

- **Common support:** the $z_1, z_2, \dots$ control variables range over the same values in the treated ($x = 1$) and untreated ($x = 0$) subjects
- Outside the common support, for given values of $z_1, z_2, \dots$, subjects are either treated ($x = 1$) OR untreated ($x = 0$)
  - Soccer example: in some countries, clubs cannot be owned by foreigners
  - FDI example: foreign-owned firms are usually larger, more productive, and pay higher wages than domestically-owned firms
- The effect of x on y **should be estimated on the common support**

Common support cont'd.

- Uncovering the common support is the important difference between OLS and matching.
- For this reason, matching is superior to OLS: it finds the common support by construction while the OLS does not
  - Exact matching: drops observations outside the common support
  - P-score matching: drops observations outside the common support during NN matching
- Lack of common support may *bias the estimated effect*
  - OLS: we can manually drop observations outside the common support

Matching: limitations

Matching: limitations

We can only match on observable characteristics

- **Unobserved heterogeneity may still bias the estimation**
- We may be aware of this (known unknown) but it is also possible, that we did not think of the mechanism leading to endogeneity (unknown unknown)

Case study: exact matching (coarsened)

- Matching variables
  - Prop of employees with college: 4 bins
  - Firm age: 4 bins
  - Level of competition: 3 bins
  - Employment: 10 bins
  - Industry: 20 bins
  - Country: 21 bins
- 5,040 bins in total
- In the data we observe only 2,435 cells
  - E.g., there is only one furniture making company in Japan
- Exact matching can be performed only with about one-third of the subjects
  - This sample may be non-random

Exact matching: results

- 766 cells with both treated and control subjects
- Estimated effect with exact matching: -0.16
- $ATE = -0.16$ – if we believe that we solved all endogeneity issues

NN matching

	(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: summary

- Four estimations
  - Simple OLS: -0.37
  - OLS with controls: -0.19
  - Exact matching: -0.16
  - NN matching: -0.18
- Which is the causal effect?
  - Simple OLS almost certainly upward biased
  - Exact matching: we lose most of the sample. Internal validity better, external validity worse
  - P-score matching gives the same result as OLS with controls: OLS simpler, I would use that
  - **Unobserved heterogeneity may still bias the estimation!!!** It is hard to establish causality with cross-sectional observational data