Applied Quantitative Methods II

Lecture 6: Matching, Standard Errors

Klára Kalíšková

Outline

1. Matching: Idea
2. Matching: Theory
3. Matching Example: HISP
4. Matching Example - Piped Water
5. Standard errors

Motivation

Example

Question: does part-time employment increase life satisfaction?

- DiD approach?

- What if we don't have a single assignment rule for the treatment?

Examples

- Does probation period instead of sentencing detter recidivism?
- Does university degree increase wages?
- ...

Motivation

Example

Question: does part-time employment increase life satisfaction?

- If we observe many people; some in T (part-time) while others in C (full-time)
  - And have many observable characteristics: demographics, abilities, health data ...
- We may assume observables fully characterize person/unit
  - And no role of unobservables!
- And find two similar units (twins), one in T, other in C
  - and compare their outcomes: difference in life satisfaction between similar T and C
  - "MATCH THEM BASED ON OBSERVABLE CHARACTERISTICS AND COMPARE"

Motivation & Intuition

Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	2	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

Outline

1. Matching: Idea
2. Matching: Theory
3. Matching Example: HISP
4. Matching Example - Piped Water
5. Standard errors

Theory

- Counterfactuals: what would have happened to treated subjects, had they not received treatment?
  - Potential outcomes vs real outcomes
- Matching = pairing treatment and comparison units that are similar in terms of observable characteristics
  - construction of artificial counterfactual

Assumption 1 – Unconfoundedness

Conditional on observables (X_i) we can take assignment to treatment (T_i) as "random" :

$$(Y_{i0}, Y_{i1}) \perp T_i \mid X_i$$

- Implicitly, we assume that unobservables do not play role in treatment assignment –they are similar among groups
- Strong assumption!

Matching methods

1. Exact matching
2. Propensity score matching
    1. Nearest neighbor
    2. Kernel matching
    3. Radius matching

1. Exact matching

- Each group of treated has the counterpart with exactly the same characteristics

Procedure:

1. We define cells for combinations of observables
   - E.g.: Sex x age group x education x region
   - must be baseline chars (not affected by T)
2. We compare average outcome of treated and untreated in each cell (combination of characteristics)
3. Total effect: weighted average of cells (weights are frequencies of observed cells)
   - Example: Payne, Lissenburgh, White a Payne (1996) Employment training, Employment Action in Great Britain

Exact matching – Problems

- Problem: To create cells, only few X's can be used and they cannot be continuous measures
- If we use more X's , we will not have enough matches (the cell size will be small)
  - called "curse of dimensionality"
- Few X's might not fully explain selection process $\implies$ main assumption of matching would be violated
- We need a tool that "merges" more dimensions into one 1 number – score, that would measure how much similar are treated and untreated
  - Solution = propensity score matching

2. Propensity score matching

- Propensity score = probability that an individual is treated based on his/her pre-treatment characteristics:
  $P(X) = P(T = 1 \mid X) = E(T \mid X)$
  - regress T on chars, get fitted values (score)
  - compare outcomes of units with similar propensity score (procedure in Stata)
- When can we use propensity score (p score or $P(X)$) instead of X?
  1. **Balancing property** – for given range of propensity score, distribution of characteristics of treated and untreated is the same (testable!)

  - Therefore, we can only use propensity score matching (instead of exact matching) if the Xs that are used to calculate p score are chosen so that **individuals with similar p score also have all their characterisctics similar**
  - You can check this property by looking at means of characteristics (Xs) within a given range of p score and check if they are indeed similar.

Common support

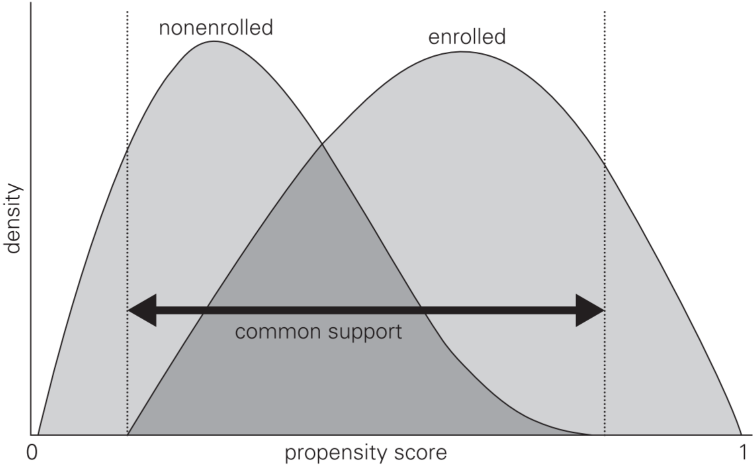
- Matching can only work if there is a region of "**common support**"
  - People with the same X values are in both the treatment and the control groups
  - Example: If treatment group only includes people with high-school education and control group only includes people with elementary education, we cannot use matching!

Assumption 2 – Common support

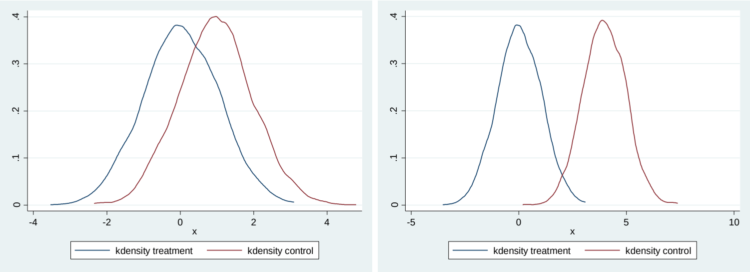
Let S be the set of all observables X, then $0 < Pr(T = 1 \mid X) < 1$ for some S^* subset of S

- Intuition: Someone in control group has to be close enough to match to treatment unit
- For common support, we need to have enough overlap in the distribution of characteristics of treated and untreated individuals.

Common support



Common support



Motivation & Intuition

Example

Question: does part-time employment increase life satisfaction?

- Problems: I may not have similar units in T and C
  - only less educated in C, only more educated in T
  - cannot compare them
  - "lack of common support"
  - (or only a small overlap)
- Have to restrict the sample to the "comparable ones" (with similar characteristics)
- **Matching estimates impact only for comparable people**, while OLS uses all observations!!

Propensity score matching – procedure

1. Data: needs large and detailed dataset with a lot of observable characteristics
2. Define T and C groups
3. Estimate propensity score by logit / probit / LPM from the pooled sample of T and C
4. Check balancing property (test means of X within strata of $p(X)$)
5. Choose matching method (see below)
6. Compare outcomes (means of outcome for T and matched C)
7. Check sensitivity

Matching methods

- 1. **Nearest neighbor matching**
  - Searching for the most similar unit in C for each unit from T
    - the one with the closest propensity score
    - outcome of this unit is used as a counterfactual
  - Distance (difference of p score) between treated and control unit is not always the same
  - All matches are weighted the same in final average effect
- 2. **Radius matching**
  - We define distance and match with all controls within this distance – average of the outcome of these units is used as a counterfactual
- 3. **Kernel matching**
  - We put some type of distribution (e.g. normal) around the each treatment unit and use it to weight closer control units more and farther control units less (this weighted average of outcomes of the control units is our counterfactual)
  - We can set “bandwidth” - limiting the maximum distance in p score that is allowed

Matching – potential problems

- "as good as the X's"
- Choice of matching algorithm – no "perfect" solution, depends on the properties of sample
  - Rule of thumb – if all give the same results it is ok, if not – look for problem
- Standard errors: Estimated variance of treatment effect should include additional variance from estimating $p(X)$
  - Typically people "bootstrap" – estimate your coefficients over and over until you get a distribution of those coefficients—calculate SD from this distribution + 95% confidence interval
- Sensitivity analysis
  - change the parameters of matching algorithm and check if results are the same

Ok, why don't we just use OLS and be done with it?

- Common support
  - OLS extrapolated treatment effect also on the regions outside of common support
  - i.e. OLS uses also "uncomparable" individuals from the treatment and control groups to estimate the impact
  - matching only looks at comparable units, which is often preferable
- OLS is imposing functional form, while PSM is nonparametric

Advantages of matching

- Works where other methods fail
  - no randomization, RDD or diff-in-diff
- Good at evaluating obligatory programs
  - or if participation is based on some clearly defined observed characteristics
- Non-parametric method
- Good in combination with other methods

Disadvantages of matching

- Questionable assumption about irrelevance of unobservables in participation decision
  - non-testable
- Sensitive to what Xs we choose
- Required to have large and rich data
  - a lot of pre-treatment characteristics

Outline

1. Matching: Idea
2. Matching: Theory
3. Matching Example: HISP
4. Matching Example - Piped Water
5. Standard errors

Matching – example 1

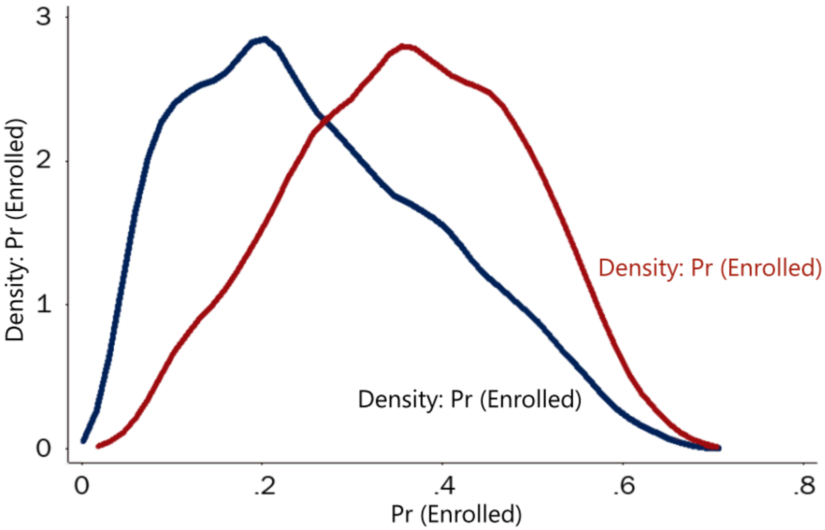
- Policy: Health Insurance Subsidy Program (HISP), which subsidized health insurance for poor households in rural areas, covering costs of primary care and basic drugs
- Aim: decrease health-expenditures and improve health status of poor families
- Let's use matching for evaluation:
- 1. estimate probability of enrolling into HISP
- 2. restrict sample to those that have counterparts
    - for each enrolled household, find non-enrolled subgroup with similar p score
    - compare the outcomes of the two groups

Matching – example 1

Table 7.1 Estimating the Propensity Score Based on Observed Characteristics

Dependent Variable: *Enrolled* = 1	
Explanatory variables / characteristics	**Coefficient**
Head of household's age (years)	−0.022**
Spouse's age (years)	−0.017**
Head of household's education (years)	−0.059**
Spouse's education (years)	−0.030**
Head of household is female = 1	−0.067
Indigenous = 1	0.345**
Number of household members	0.216**
Dirt floor = 1	0.676**
Bathroom = 1	−0.197**
Hectares of land	−0.042**
Distance to hospital (km)	0.001*
Constant	0.664**

Matching – example 1



Source: Gertler et al. Impact Evaluation in Practice

Matching – example 1

Table 7.2 Case 7—HISP Impact Using Matching (Comparison of Means)

	Enrolled	Matched comparison	Difference	*t*-stat
Household health expenditures	7.8	16.1	−8.3	−13.1

Source: Authors.

Table 7.3 Case 7—HISP Impact Using Matching (Regression Analysis)

	Multivariate linear regression
Estimated impact on household health expenditures	−8.3** (0.63)

Source: Authors.

Source: Gertler et al. Impact Evaluation in Practice

Outline

1. Matching: Idea
2. Matching: Theory
3. Matching Example: HISP
4. Matching Example - Piped Water
5. Standard errors

Case study: Piped Water in India

- Jalan and Ravaillion (2003): Impact of piped water for children's health in rural India
- Research questions of interest include:

1. Is a child less vulnerable to diarrhoeal disease if he/she lives in a HH with access to piped water?
2. Do children in poor or poorly educated HHs have smaller health gains from piped water?
3. Does income matter independently of parental education?

Case study: Piped Water in India

Design

- Classic problem for infrastructure programs:
  - randomization is generally not an option
  - (although randomization in timing may be possible in other contexts)
- The challenge: observable and unobservable differences across households with piped water and those without
  - What are differences for such households in India?
- Jalan and Ravallion use cross-sectional data 1993-1994 nationally representative survey on 33,000 rural HHs from 1765 villages

Case study: Piped Water in India

PSM in practice

- To estimate the propensity score, authors used logit with these vars:
  - Village level characteristics Including:
    - Village size, amount of irrigated land, schools, infrastructure (bus stop, railway station)
  - Household level variables Including:
    - Ethnicity / caste / religion, asset ownership (bicycle, radio, thresher), educational background of HH members
- Are there important variables which can not be included?
- Poor households may be less able to benefit from piped water b/c they do not properly store water
  - Can we check that?

Case study: Piped Water in India

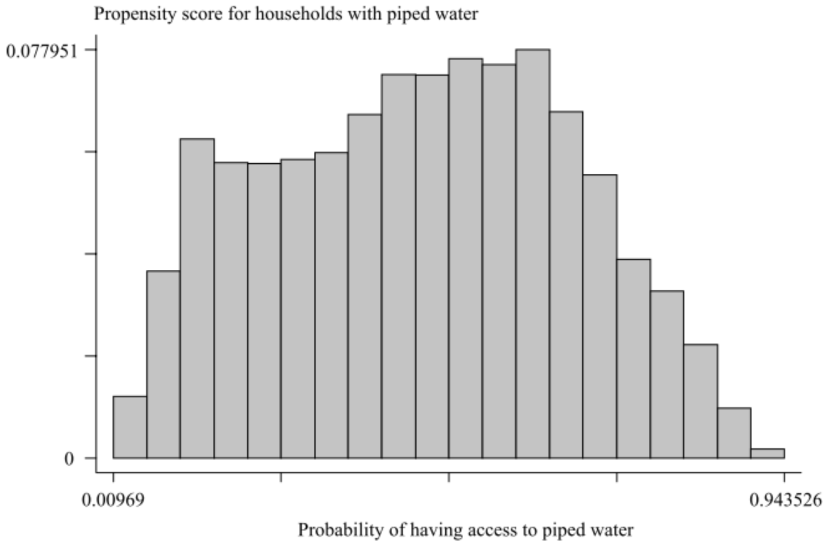
Results - logit (only a small part)

Logit regression for piped water

	Coefficient	t-statistic
Village variables		
Village size (log)	0.08212	4.269
Proportion of gross cropped area which is irrigated: > 0.75	−0.04824	−1.185
Proportion of gross cropped area which is irrigated: 0.5–0.75	0.19399	4.178
Whether village has a day care center	−0.07249	−2.225
Whether village has a primary school	−0.08136	−1.434
Whether village has a middle school	−0.09019	−2.578
Whether village has a high school	0.26460	7.405
Female to male students in the village	0.10637	3.010
Female to male students for minority groups	−0.07661	−2.111
Main approachable road to village: pucca road	0.19441	3.637
jeepable/kuchha road	−0.00163	−0.033
Whether bus-stoop is within the village	0.11423	2.951
Whether railway station is within the village	0.00920	0.179

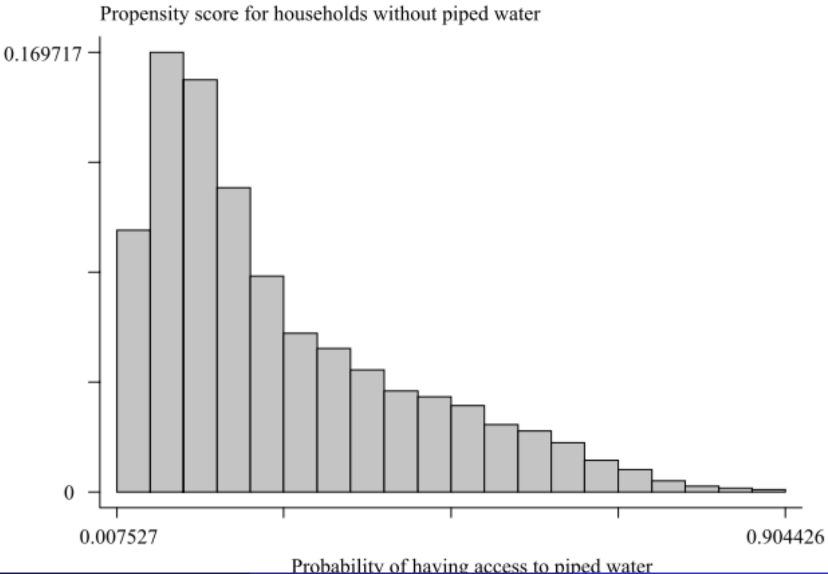
Case study: Piped Water in India

Results - common support



Case study: Piped Water in India

Results - common support



Case study: Piped Water in India

Results - nearest neighbor

Table 3
Impacts of piped water on diarrhea prevalence and duration for children under five

	Prevalence of diarrhea		Duration of illness	
	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)
Full sample	0.0108	−0.0023*	0.3254	−0.0957*
	(0.046)	(0.001)	(1.650)	(0.021)
Stratified by household income per capita (quintiles)				
1 (poorest)	0.0155	0.0032*	0.4805	0.0713
	(0.055)	(0.001)	(2.030)	(0.053)
2	0.0136	0.0007	0.4170	0.0312
	(0.051)	(0.001)	(1.805)	(0.051)
3	0.0083	−0.0039*	0.2636	−0.1258*
	(0.038)	(0.001)	(1.418)	(0.042)
4	0.0100	−0.0036*	0.3195	−0.1392*
	(0.044)	(0.001)	(1.703)	(0.048)
5	0.0076	−0.0068*	0.1848	−0.2682*
	(0.042)	(0.001)	(1.254)	(0.036)

Case study: Piped Water in India

Conclusion

- Disease prevalence among those with piped water would be 21% higher without it
- Gains from piped water exploited more by wealthier households and households with more educated mothers
- Even find counterintuitive result for low income, illiterate HH: piped water is associated with higher diarrhea prevalence
- need to combine infrastructure investment with promotion of knowledge about health

Outline

1 Matching: Idea

2 Matching: Theory

3 Matching Example: HISP

4 Matching Example - Piped Water

5 Standard errors

Standard errors – OLS residuals

- Recall the Gauss-Markov property of OLS
  - error term has a constant variance and is uncorrelated with each other
- When this assumption is violated:
  1. Heteroskedasticity - variance of error term is different for different observations (people with higher income have also higher variance in unexplained part of the income)
  2. Error terms are related to each other within certain groups (unobservables of children within one class will be correlated more than across classes)
- How to colve these problems?
  1. Heteroskedasticity-robust standard errors (White)
  2. Clustered standard errors (assume correlation of error term within cluster, but independence across clusters)

Robust vs. clustered SE

- Heteroskedasticity-robust standard errors assume independent observations of the error term
- Is this a good assumption?
  - Example: Imagine a policy that is focused only on certain villages
  - Can we assume that unobservable characteristics (ε) within one village are uncorrelated?
- The solution is to allow for a different correlation of unobservable characteristics (ε) within villages than across villages
- This solution is called clustering - Stata command: , vce(cluster clustervar)
- Clustering can only be used if number of clusters is reasonably large (20+)

Clustered standard errors

A haiku by Keisuke Hirano

T-stat looks too good
Try clustered standard errors–
Significance gone

Bootstrap

- A method for estimating the distribution of an estimator or test statistic by resampling data
- If an estimator is difficult to calculate
- Bootstrap standard errors are used when other methods (robust, cluster) fail
  - usually with small samples and small number of clusters

Bootstrap – general procedure

- Let's say we want to estimate SEs of a regression coefficient $\hat{\beta}_1$:

1. Generate a bootstrap sample b of size n by randomly selecting observations with replacement from your original data – this is called also nonparametric bootstrap
2. Estimate coefficient $\hat{\beta}^*_{1b}$ coming from your model fitted to this bootstrap sample
3. Repeat 1.+2. B times to gain a set of $\hat{\beta}^*$
4. Compute SEs:

$$SE_{\hat{\beta}_1,B} = \left(\frac{1}{1-B} \sum_{b=1}^{B} \left(\hat{\beta}^*_{1b} - \overline{\hat{\beta}^*_1} \right)^2 \right)^{1/2}$$

where $\overline{\hat{\beta}^*_1}$ is the mean of bootstrap estimates of β, i.e. $\overline{\hat{\beta}^*_1} = (1/B) \sum_{b=1}^{B} \hat{\beta}^*_{1b}$

- We can then use $SE_{\hat{\beta}_1,B}$ in a typical Wald test
- For block bootstrap, we select clusters with replacement instead of

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY