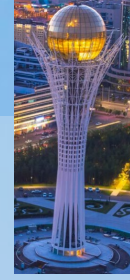
unicef | for every child



Implementing Causal Machine Learning in

Online Workshop 4: Data for the hands-on-project & useful descriptive statistics

August / September / October 2024

Michael Lechner
Professor of Econometrics | University of St. Gallen | Switzerland

unicef | for every child

The workshop series | 1

4 online workshops

- August 27, 13-15: Correlation & causation
- September 4, 13-15: Available identification strategies & classical estimation
- September 10, 13-15: Machine Learning
- September 11, 13-14: The data for the Astana workshop & useful descriptive statistics

unicef | for every child

The workshop series | 2

The Astana Workshop September 30 to October 4, 2024 (10-13, 14:30-17:30)

- Monday
  - Morning: Identification with experiments & selection on observables
  - Afternoon: Discussion of potential programmes to be evaluated
- Tuesday: Causal Machine Learning (theory)
- Wednesday
  - Morning: 2 empirical examples
  - Afternoon: The mcf package – how to use it & how to interpret the results
- Thursday: Doing an empirical study in groups with the data introduced in online workshop 4
- Friday: Discussion of programmes to be evaluated continued (core team only)

unicef | for every child

Quick introduction

Participants

- Professional background?
- Knowledge in the estimation of causal effects, machine learning, Python?

Myself

- Professor of Econometrics at the University of St. Gallen
- Co-head of The Swiss Institute for Empirical Economic Research at the University of St. Gallen
- Empirical Economic Research | SEW-HSG | University of St.Gallen (unisg.ch)

  www.michael-lechner.eu
- Research interest in Causal Machine Learning, AI, programme evaluation, …

unicef | for every child

Plan for today's workshop

The data for the Astana workshop & useful descriptive statistics

- Information on the data provided
- Basic descriptive statistics
- Selectivity in the data
- Common support
- Your take-home task
  - Do descriptive analysis of your region of choice
  - Estimate the programme effects with OLS on the common support

unicef | for every child

Data & context description

Evaluate an Active Labour Market Policy with 4 different programmes of country XXX

- 2 different training programmes
- 2 different employment programmes
- Non-participant

The file ***study_description.pdf*** provides …

- Information on programmes, selection-into-the programmes & other institutional details
  - This information should be sufficient to decide on a credible research design (i.e., identifying assumption)
- Code book for the data

Data

Data comes from 5 different regions of that country

- *Central.csv, East.csv, North.csv, South.csv, Southwest.csv, West.csv*
  - These data files may contain missing values
- Two additional (somewhat cleaned) data sets are available for test purposes
  - *Large_testdata_noNaN.csv* (N=100'000)
  - *Small_testdata_noNaN.csv* (N=4'000)
- All data files
  - contain the same variables
  - are comma separated with variable names in the first row
  - *Python*: Easy way to read this data is putting the data into a Pandas DataFrame
    - *E.g., data_df = pandas.read_csv(file_name)*

The subsequent statistics are based on *Large_testdata_noNaN.csv*

unicef | for every child

Standard descriptive statistics

Min, max, mean, median, # of NaN's

	count	mean	std	min	25%	
id_mcf	100000.0	49999.500000	28867.657797	0.000000	24999.750000	49999.50
earnx9_4	100000.0	4160.564153	2416.457958	0.000000	2737.982457	4246.57
ptype	100000.0	1.294620	1.420922	0.000000	0.000000	1.00
earnx2_1	100000.0	801.317523	1644.385062	0.000000	0.000000	0.00
earnx2_2	100000.0	445.165093	1246.100281	0.000000	0.000000	0.00
earnx2_3	100000.0	225.939558	876.453603	0.000000	0.000000	0.00
earnx2_4	100000.0	90.896164	542.867278	0.000000	0.000000	0.00
age	100000.0	40.020600	6.056986	30.000000	35.000000	40.00
sex	100000.0	1.600260	0.489847	1.000000	1.000000	2.00
school	100000.0	10.202370	1.326859	8.000000	9.000000	10.00

Purpose (smell tests)

- Identify inconsistencies with the data codebook ➔ SEX cannot have a minimum of 1!
- Check credibility of variables
  - Example: Is the mean of earnings plausible?
- Check if it is plausible that data is representative for a meaningful population or highly selective
- …

Descriptive statistics by treatment status | 1

Check if all treatments are large enough to be formally evaluated

0	1	2	3	4
43494	19469	12264	13627	11146

Check unadjusted outcome difference (causal effect incl. selection)

- This would be the treatment effect if there is no selection bias

Mean					
ptype	0	1	2	3	4
earnx9_4	4200.189603	5037.850116	4709.193401	4237.840594	1775.422853

  - Either programme 4 is highly selective, or there is a huge negative (!) programme effect for P4

unicef | for every child

Descriptive statistics by treatment status | 2

Get descriptive impression about possible variables leading to selection effects

Mean					
ptype	0	1	2	3	4
earnx9_4	4200.189603	5037.850116	4709.193401	4237.840594	1775.422853
earnx2_1	838.503235	923.466058	1080.963458	599.546689	381.838625
earnx2_2	474.259527	522.187779	600.051272	311.598112	189.970869
earnx2_3	237.787964	277.673447	300.062270	159.681803	88.788273
earnx2_4	95.299385	110.116460	117.284814	69.438827	37.339293
age	40.585529	40.737737	38.052104	40.551992	38.079760
sex	1.603164	1.589399	1.519651	1.607764	1.687421
school	10.482365	10.330012	10.264188	9.687239	9.448591
voc_deg	1.105854	0.968411	0.827462	0.767741	0.695765

- Pre-programme earnings & schooling for participants in P4 are much lower
  - Most likely at least partly responsible for negative unadjusted effect (seen in previous slide)

Balancing tests | 1

Balancing tests are a more formal way to check if some variables have an unequal distribution between the treatments

Pairwise comparisons between programmes

Two common approaches

- Formal two sample t-test for mean differences
  - Rejects (almost) always if sample is large enough, even if differences are too small to matter
- Standardized difference
  - Independent of sample size
  - Values > 10/20 are seen as indication or stronger selection

$$\frac{mean_{P1}(X) - mean_{P0}(X)}{\sqrt{\frac{\text{var}_{P1}(X) + \text{var}_{P0}(X)}{2}}} \times 100$$

Balancing tests | 2

As example, let's look at P4 again

Comparing treatments 4 and 0

Variable	Mean	Std	t-val	p-val (%)	Stand.Difference (%)
earnx9_4	-2424.76675	19.81864	122.35	0.00	-136.05
earnx2_1	-456.66461	11.85361	38.53	0.00	-33.09
earnx2_2	-284.28866	8.71399	32.62	0.00	-27.38
earnx2_3	-148.99969	5.92108	25.16	0.00	-20.82
earnx2_4	-57.96009	3.72092	15.58	0.00	-13.02
age	-2.50577	0.06183	40.53	0.00	-42.36
sex	0.08426	0.00498	16.93	0.00	17.68
school	-1.03377	0.01185	87.25	0.00	-84.63
voc_deg	-0.41009	0.00537	76.32	0.00	-72.41

Clear indication of strong selectivity!

unicef | for every child

Selectivity for P4 much stronger than for, e.g., P1

Comparing treatments 1 and 0

Variable	Mean	Std	t-val	p-val (%)	Stand.Diffe
earnx9_4	837.66051	20.47156	40.92	0.00	38.18
earnx2_1	84.96282	15.30547	5.55	0.00	4.82
earnx2_2	47.92825	11.67235	4.11	0.00	3.56
earnx2_3	39.88548	8.36253	4.77	0.00	4.17
earnx2_4	14.81707	5.20282	2.85	0.44	2.49
age	0.15221	0.05199	2.93	0.34	2.52
sex	-0.01377	0.00423	3.25	0.12	-2.81
school	-0.15235	0.01149	13.26	0.00	-11.32
voc_deg	-0.13744	0.00531	25.89	0.00	-21.92

Comparing treatments 4 and 0

Variable	Mean	Std	t-val	p-val (%)	Stand.Difference (%)
earnx9_4	-2424.76675	19.81864	122.35	0.00	-136.05
earnx2_1	-456.66461	11.85361	38.53	0.00	-33.09
earnx2_2	-284.28866	8.71399	32.62	0.00	-27.38
earnx2_3	-148.99969	5.92108	25.16	0.00	-20.82
earnx2_4	-57.96009	3.72092	15.58	0.00	-13.02
age	-2.50577	0.06183	40.53	0.00	-42.36
sex	0.08426	0.00498	16.93	0.00	17.68
school	-1.03377	0.01185	87.25	0.00	-84.63
voc_deg	-0.41009	0.00537	76.32	0.00	-72.41

unicef | for every child

Intro

The common support conditions means in this data that every unit in the data must be able to end up in all 5 possible states

If the common support condition is violated, many causal evaluation methods will not work

How to check for common support violation | 1

Descriptive statistics by treatment state

- Are there any features that do not appear in certain treatments? → check means by treatment
  - Not obvious from the descriptive analysis so far
    - But: Consider more variables in descriptive analysis
    - But: Recode categorial variables as dummies to make detection easier

How to check for common support violation | 2

Check formal identification condition (propensity score)

$$0 < P(D = j \mid X = x) < 1, \quad j = 1, 2, 3, 4$$

However, even if this is fulfilled, it might be in the data there are no comparison observations that have similar propensity scores in the different treatment groups

How to check for common support violation | 3

Treatment sample		Treatment probabilities in %				
		Upper limits				
D =	0	100.00%	60.79%	59.94%	42.01%	74.90%
D =	1	100.00%	59.79%	63.33%	39.06%	75.19%
D =	2	100.00%	49.49%	63.41%	43.81%	70.27%
D =	3	100.00%	50.01%	58.02%	42.61%	72.84%
D =	4	100.00%	47.39%	48.34%	41.48%	71.52%
		Lower limits				
D =	0	0.00%	4.95%	0.04%	0.00%	0.00%
D =	1	0.00%	5.68%	0.00%	0.00%	0.00%
D =	2	0.00%	5.28%	1.79%	0.74%	0.00%
D =	3	0.00%	4.33%	1.29%	1.82%	0.00%
D =	4	0.00%	4.58%	1.11%	4.76%	0.59%

Rule

- Largest acceptable upper value: Smallest maximum in any treatment sample
- Smallest acceptable lower value: Largest minimum in any treatment sample

Upper limits used:	100.00%	47.39%	48.34%	39.06%	70.27%
Lower limits used:	0.00%	5.68%	1.79%	4.76%	0.59%

Remedies

Delete observation that are off-support

```
Observations deleted: 9391 (18.78%)
-----------------------------------

Observations kept, by treatment
            Obs.  Share in %
ptype
0      15148       37.30
1       7592       18.70
2       5597       13.78
3       6687       16.47
4       5585       13.75
-   -   -   -   -   -   -   -   -
Observations deleted, by treatment
        Obs.  Share in %
ptype
0      6594       70.22
1      2155       22.95
2       494        5.26
3       140        1.49
4         8        0.09
```

Implications

Has the population changed in an important way?

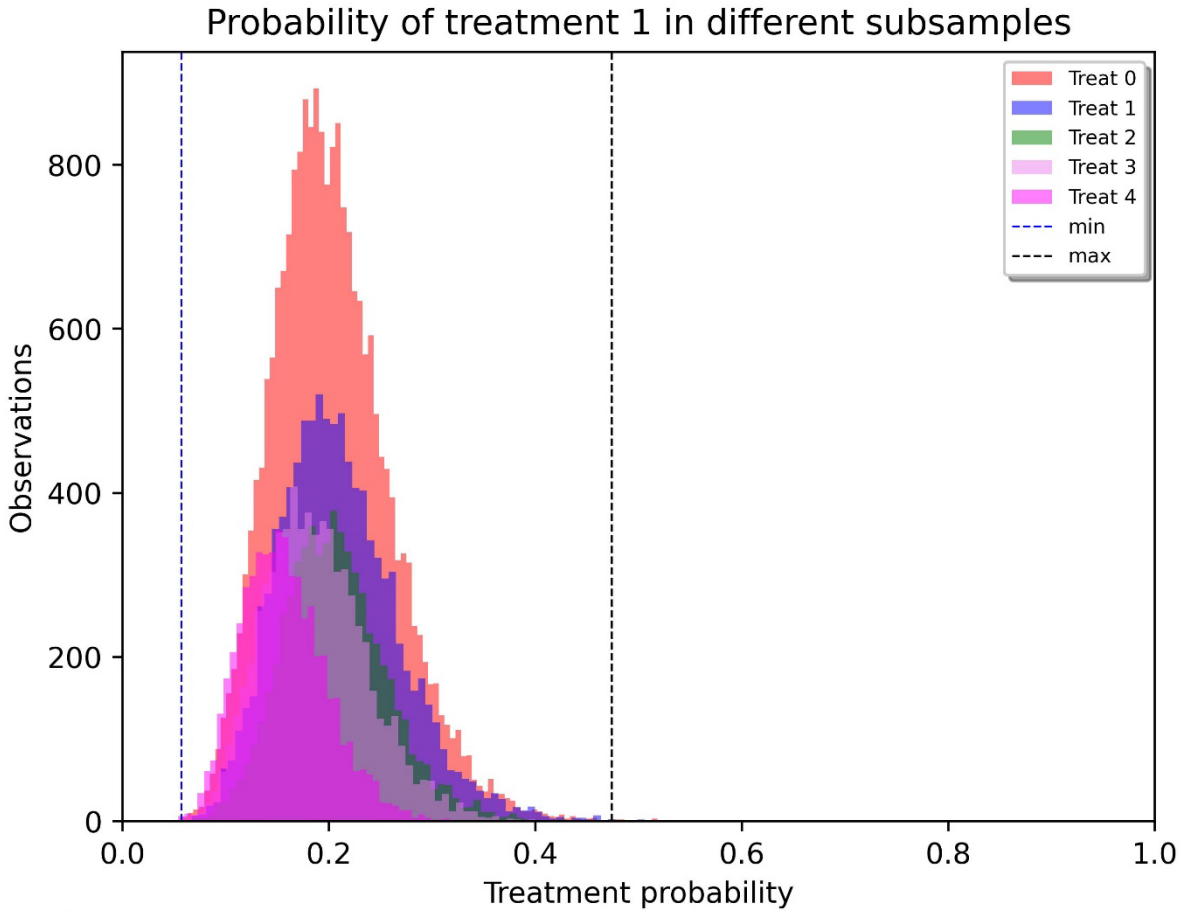
Full sample (Data ON and OFF support)

	count	mean	std	min	25%	50%	75%	max
age	50000.0	40.017100	6.049888	30.000000	35.000000	40.000000	45.000000	50.000000
sex	50000.0	1.598960	0.490114	1.000000	1.000000	2.000000	2.000000	2.000000
school	50000.0	10.205120	1.323964	8.000000	9.000000	10.000000	12.000000	12.000000
voc_deg	50000.0	0.951100	0.588214	0.000000	1.000000	1.000000	1.000000	2.000000
reg_al	50000.0	11.546772	4.360950	5.029994	7.246800	11.031909	15.077459	19.806119

Data ON support

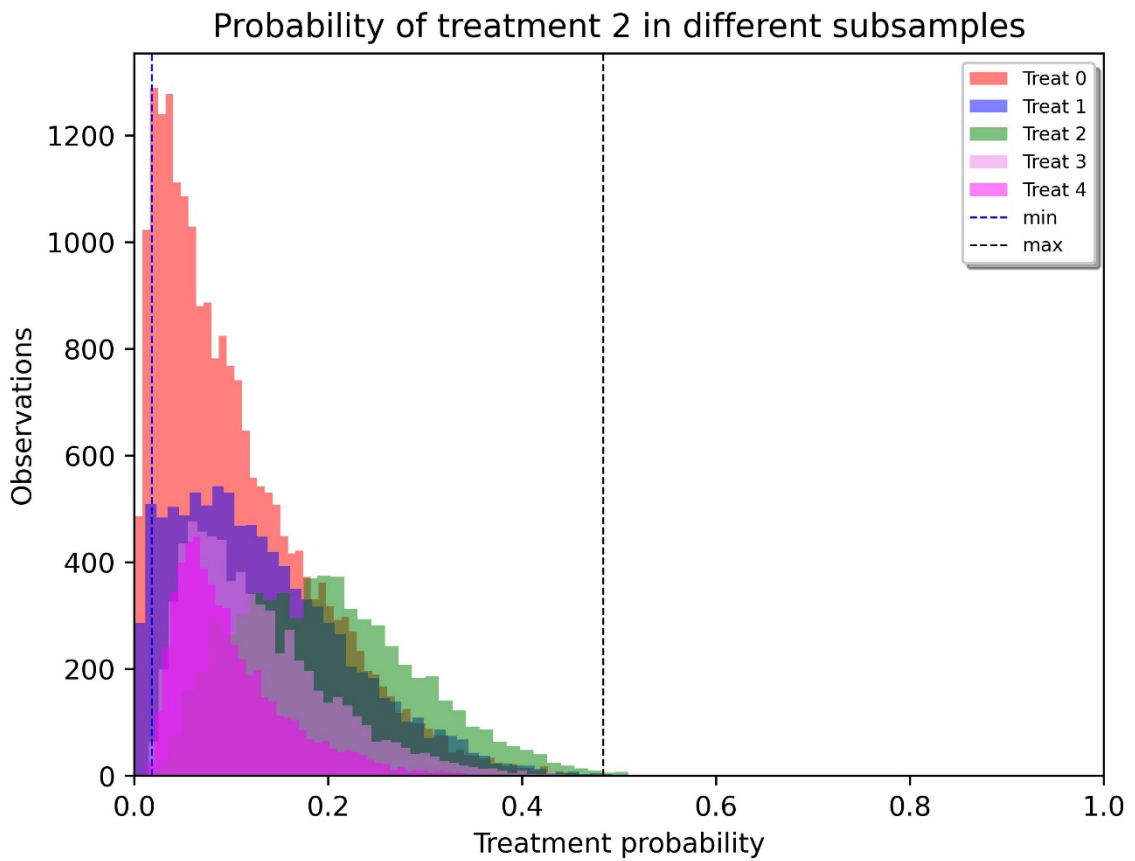
	count	mean	std	min	25%	50%	75%	max
age	40609.0	39.846930	6.025594	30.000000	35.000000	40.000000	45.000000	50.000000
sex	40609.0	1.610480	0.487647	1.000000	1.000000	2.000000	2.000000	2.000000
school	40609.0	9.836613	1.157735	8.000000	9.000000	10.000000	10.000000	12.000000
voc_deg	40609.0	0.762269	0.429557	0.000000	1.000000	1.000000	1.000000	2.000000
reg_al	40609.0	11.545319	4.362435	5.029994	7.246800	11.031909	15.077459	19.806119

Visualisations | P1



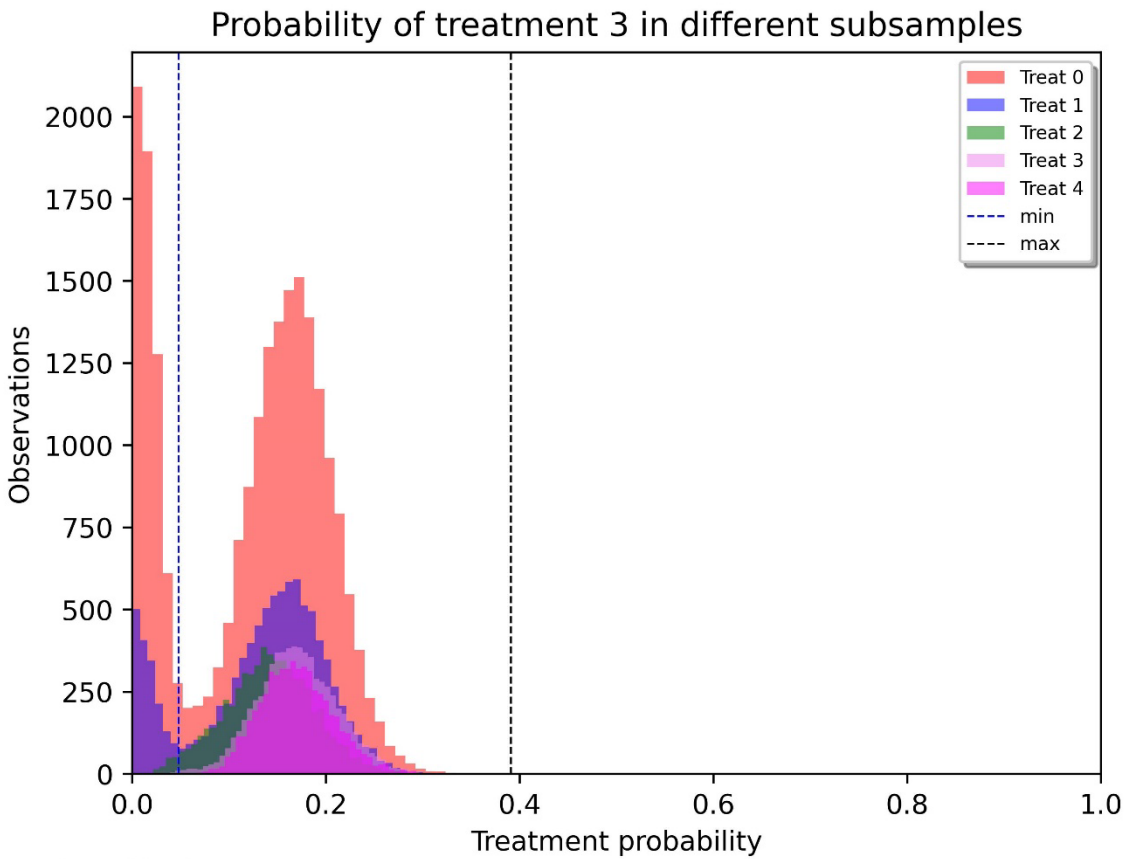
Data: Training - fill mcf with y data

Visualisations | P2



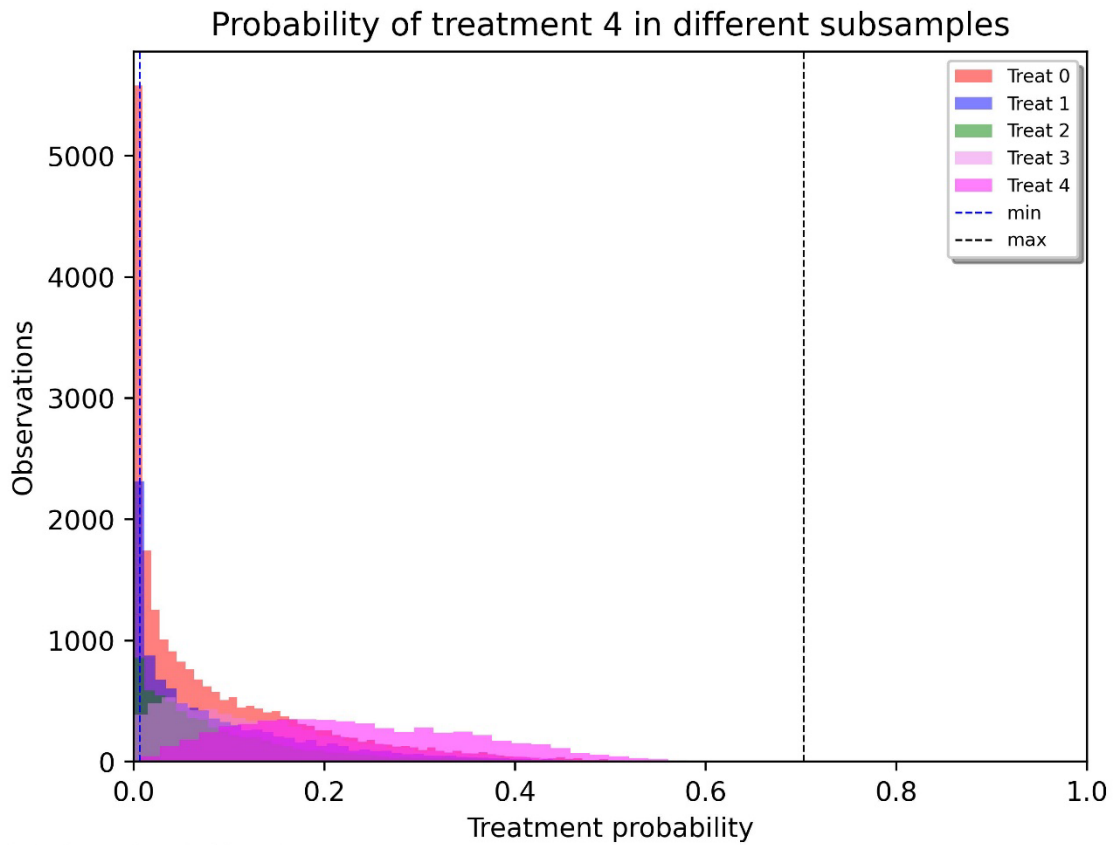
Data: Training - fill mcf with v data

Visualisations | P3



Data: Training - fill mcf with v data

Visualisations | P4



unicef | for every child

unicef | for every child

Tasks

Provide the descriptive statistics discussed in this lecture for one of the regions in country XXX

Analyse the common support

Provide a regression (OLS) based estimation of the average causal effect of all 4 programmes compared to non-participation

- Use the common support as data
- Specify the treatment with 4 indicator variables, one for each programme

Practical hints | 1

You may choose the data from region you prefer

All data files have the same structure

- *****testdata_noNaN.csv* have no missing values, while the other files have missing values

It may be an efficient procedure to develop the code & run it on the smallest data set

- *Small_testdata_noNaN.csv*
- Once everything is running, apply your programme to your region of choice

unicef | for every child

Practical hints | 2

There are two possible strategies to solve the take-home

- Compute the statistics with your software / package of choice
  - All statistical packages should be able to do this (perhaps with the exception of the standardized difference)
- Use the *mcf* module in Python for the descriptive statistics
  - All statistics discussed are implemented in the *mcf* Python module
  - If you will participate in the Astana hands-on (Thursday), then this is the efficient strategy
    - pip install *mcf* (from PyPI)
    - Follow advice about installation on *mcf* homepage (Modified Causal Forests — mcf 0.6.0 documentation (mcfpy.github.io))

1 | Introduction

2 | Information on the data provided

3 | Selectivity in the data

4 | Comprehensive CML estimators

5 | Common support

6 | Do-until-workshop-Astana task

7 | Conclusions & outlook

Take-away

A suitable descriptive analysis is an important 1st step in any causal analysis

- Check if data is ok (variables, population)
- Analyse common support

unicef | for every child

Next: The Astana workshop – see you there!

AND if you are not familiar with Python (& will participate in the hands-on sessions in Astana), attend the Python courses recommended by Alibi

Michael Lechner

Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland