Data-driven sensitivity analysis for Matching estimators

Giovanni Cerulli [1]

[1]IRCrES-CNR, Research Institute on Sustainable Economic Growth

London Stata Conference 2018
Cass Business School
September 6-7

Summary

- Motivation and objective
- Current approaches
- The LOCO approach
- Stata implementation via `sensimatch`
- Application
- Conclusion

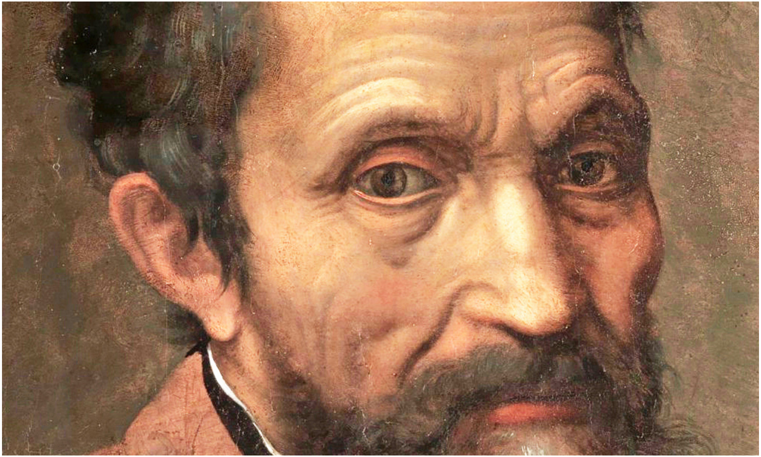
Motivation and objective

- Under "**unobservable selection**" Matching is an inconsistent estimator of the ATET
- Unobersevables are **context–dependent** (genuine and/or contingent unobservables)
- Alternative methods: instrumental–variables (IV), selection models (SM), and quasi-natural approaches (regression discontinuity design, RD), Diff–in–diffs
- Costly alternatives require **extra information** and assumptions, rarely available, not accessible, often unreliable
- **Sensitivity analysis** helps to detect whether Matching is robust to unobservable selection

Motivation and objective

This paper:

- proposes a (novel) sensitivity analysis for **unobservable selection** in Matching estimation based on a "leave–one–covariate–out" (**LOCO**) approach
- rooted in the **Machine Learning** literature
- based on a bootstrap over different **subsets** of covariates
- simulates **estimation scenarios** and compares them with the baseline Matching estimated by the analyst
- introduces `sensimatch`, a Stata routine I developed to run this method
- provides an instructional application on real data

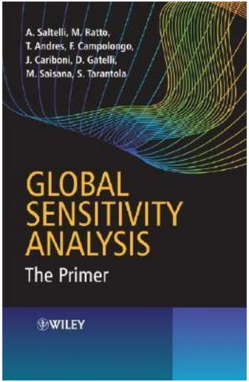


Io intendo scultura, quella che si fa per forza di **levare***: quella che si fa per via di* **porre***, è simile alla pittura*

(I mean sculpture, the one that one does by force of **removing***: what one does by* **posing***, is similar to painting)*

Michelangelo Buonarroti
"Letter to Sir Benedetto Varchi"
Florence, XVI Century

Sensitivity analysis: the study of how the *uncertainty* in the output of a model or system can be explained by different sources of uncertainty in its inputs



Sensitivty approaches in the Matching literature

Two Matching sensitivity tests for the possible presence of *unobservable selection*:

- The Rosenbaum (1987) test $\Longrightarrow$ based on the Wilcoxon's signed rank statistic
- The Ichino, Mealli, and Nannicini (IMN, 2008) test $\Longrightarrow$ based simulating the (possible) presence of unobeservable

Rosenbaum approach

- Assume perfect randomization (as restored after Matching)
- Define $\Gamma =$ "PS ratio between treated and untreatred" $\Rightarrow$ same odds under randomization
- Perturbate randomization by increasing $\Gamma \Rightarrow$ larger departure from randomization
- Look at what Γ the effect (ATET) is no longer significant (result overturning)
- A high level of critical Γ is a signal of Matching robustness

IMN approach

- Consider the baseline Matching estimates
- Define d and s as two probability ratios increasing with unobservable selection: 1. d: UCs effect on the outcome; 2. s: UCs effect on the treatment
- As soon as both d and s increase, ATET goes to zero
- Tabulate increasing values of d and s until ATET is no longer significant.
- A high level of critical d and s is a signal of Matching robustness

The logic of LOCO

- Previous methods follow a **posing** logic $\Rightarrow$ what happens when one perturbates the baseline model by adding up UCs

- LOCO follows a different but **specular** logic: "if the baseline model results are poorly (strongly) sensitive to adding up UCs, it is likely to be poorly (strongly) sensitive to removing them"

- We can obtain a specular result by **removing**, instead of **posing**

The LOCO algorithm

1. Start from running a Matching model using $\mathbf{x}=\{x_1, x_2, \ldots, x_K\}$ observable confounders, thus estimating one single ATET, and take this as the baseline estimate.
2. Starting from the K observables, select a subset size S with $S = 1, 2, \ldots, j, \ldots, M$, and $M < K$.
3. Draw H times at random and without replacement a set of covariates of size S from the original set of observables $\mathbf{x}$.
4. Run H Matching models of size S thus obtaining a number of H ATET point estimates, standard errors, and confidence intervals.
5. For each size S, average the obtained estimates over H , and check whether the results are sensibly changed by reducing S from $K - 1$ to 1.

The Stata module `sensimatch`

Title

sensimatch – Data-driven sensitivity analysis to assess Matching robustness to unobservable selection

Syntax

sensimatch *outcome* *treatment* [*varlist*] ,
sims(#) **mod**(*modeltype*) **seed**(#) **fac**(*varlist_f*)
vce(*vcetype*) **graph_options**(*options*)

modeltype

`reg`: Ordinary Least Squares

`match`: Nearest–neighbour propensity–score Matching

Application on real data

- **Dataset**: National Longitudinal Survey of Mature and Young Women (NLSW) in 1988
- **Objective**: Detecting the effect of "unionization" on hourly "wage" on 2,246 American women
- **Confounders**: *age*: age of the woman; *race*: race of the woman (white, black, other); *married*: married vs. non–married; *never_married*: whether or not never married; *grade*: grade obtained at school final exam; *south*: whether of not the woman comes from the South; *smsa*: whether she lives in SMSA; *c_city*: whether of not she lives in central city; *collgrad*: whether she is college graduated; *hours*: usual hours worked; *ttl_exp*: total work experience; *tenure*: job tenure in years; *industry*: type of industry; occupation: type of occupation.

Baseline propensity–score Matching results - psmatch2

```
****************************************************************
use nlsw88 , clear
****************************************************************
global y "wage"
global w "union"
global xvars age race married never_married ///
grade south smsa c_city collgrad hours ttl_exp tenure
global factors "industry occupation"
****************************************************************
xi: psmatch2  $w $xvars i.industry i.occupation , out($y) common

-----------------------------------------------------
          |     T      C        Diff      S.E.   T-stat
----------+------------------------------------------
DIM       |   8.67   7.25       1.44       .22     6.44
ATET      |   8.67   7.65       1.02       .37     2.76
----------+------------------------------------------
```

Rosenbaum sensitivity analysis - rbounds - #1

Using rbounds

```
. xi: psmatch2 $w $xvars i.ind i.occ , out($y) common

. gen delta = $y - _wage if _treated==1 & _support==1

. rbounds delta , gamma(1 (0.01) 2)
```

Rosenbaum sensitivity analysis - rbounds - #2

Gamma	sig+	sig-	t-hat+	t-hat-	CI+	CI-
1	2.6e-06	2.6e-06	1.08293	1.08293	.619968	1.53784
1.01	4.0e-06	1.7e-06	1.05878	1.10306	.595817	1.55797
1.02	6.1e-06	1.1e-06	1.03772	1.12319	.575685	1.58212
1.03	9.2e-06	6.9e-07	1.0145	1.14331	.556793	1.60628
1.04	.000014	4.4e-07	.994364	1.16345	.539451	1.62641
1.05	.00002	2.8e-07	.974235	1.1876	.515301	1.64654
1.06	.000029	1.8e-07	.954105	1.2037	.495169	1.66667
1.07	.000042	1.1e-07	.933976	1.22474	.47504	1.6868
1.08	.000059	6.9e-08	.913847	1.24798	.458934	1.70692
1.09	.000083	4.3e-08	.893721	1.26811	.434783	1.72705
1.1	.000116	2.7e-08	.873592	1.28422	.414655	1.74641
1.11	.000159	1.7e-08	.857484	1.30435	.394527	1.76731
1.12	.000218	1.0e-08	.837229	1.32448	.378421	1.78342
1.13	.000294	6.4e-09	.817228	1.34213	.358293	1.80354
1.14	.000394	3.9e-09	.797103	1.36071	.334139	1.81965
1.15	.000523	2.4e-09	.776974	1.38083	.314009	1.83978

....

Rosenbaum sensitivity analysis - rbounds - #3

1.35	.033501	7.9e-14	.438807	1.72593	-.036234	2.18196
1.36	.038743	4.6e-14	.421621	1.73913	-.052334	2.19659
1.37	.044587	2.7e-14	.406602	1.75523	-.068438	2.21417
1.38	.051068	1.6e-14	.3905	1.77523	-.08454	2.23027
1.39	.058221	9.0e-15	.378419	1.78744	-.100643	2.24235
1.4	.066076	5.2e-15	.362316	1.79952	-.116748	2.25845
1.41	.074661	3.0e-15	.342191	1.81562	-.132852	2.27455
1.42	.083999	1.8e-15	.326085	1.83172	-.152974	2.29054
1.43	.094111	1.0e-15	.309982	1.84523	-.165056	2.30274
1.44	.105012	5.6e-16	.293881	1.8599	-.17992	2.31884

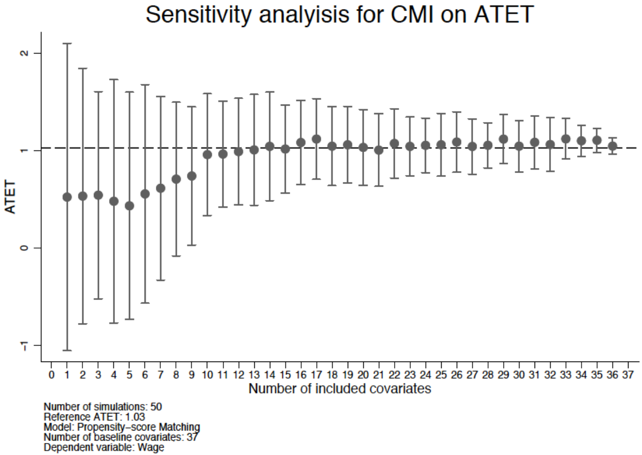
Unlikely circumstance ⇒ Matching **robust** to unobservable selection

LOCO sensitivity analysis - sensimatch - #1

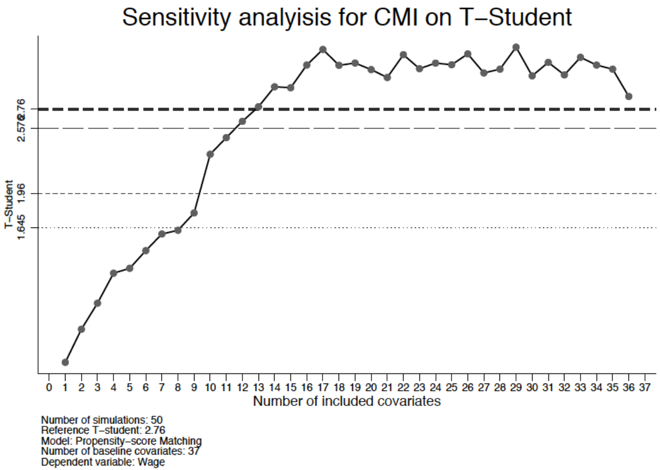
Using sensimatch

```
sensimatch $y $w $xvars , mod(match) sims(50)  ///
vce(robust) fac($factors) seed(1010)
```

LOCO sensitivity analysis - sensimatch - #2



LOCO sensitivity analysis - sensimatch - #3



LOCO sensitivity analysis - sensimatch - #4

As a possible **measure of sensitivity** to *unobservable selection* one can consider, for instance, "the ratio between the number of *not removed covariates* leading to lose α–significance and the number of the *baseline covariates*":

Sensitivity index

$$\rho_\alpha = \frac{S_{critical,\alpha}}{K}$$

As long as ρ_α increases, Matching sensitivity to *unobservable selection* increases accordingly.

LOCO sensitivity analysis - sensimatch - #5

In our previous example we have that:

$$\rho_1 = \frac{12}{37} = 0.33$$

$$\rho_1 = \frac{9}{37} = 0.24$$

$$\rho_1 = \frac{7}{37} = 0.18$$

One can pre-fix a given **threshold** for the accepted level of uncertainty as, for example, a ρ not larger than 90%. A value of ρ larger than 90 may signal a *severe* sensitivity of Matching to unobservable selection.

Conclusion

- The LOCO approach seems to lead to results **consistent** with those from the Rosenbaum approach
- It has the adavantage to be totally **data–driven** $\Longrightarrow$ it is **model–free**
- It can be **generalized** to whatever causal parameter and methods (for instance the IPW)
- It has the disadvantage to be **computationally intensive** and thus slower to provide results

Many thanks !!!

See you next year for the London Stata Conference 2019 !