Treatment Evaluation and Matching Methods

Applied Econometrics for Researchers, PhD

Vera Rocha, CBS-SI, vr.si@cbs.dk

16th November 2022

Agenda

1. Treatment effects: what and why?
2. Key assumptions
3. Matching & Propensity Score Estimators
4. PSM in research and in Stata (examples)

Key readings:

– Caliendo, M., Kopeinig, S. (2008), ”Some practical guidance for the implementation of propensity score matching”, *Journal of Economic Surveys*, 22(1), 31-72.

– **Cameron & Trivedi**, Microeconometrics: Methods and Applications, **chapter 25**.

  - Alternatively, Chapter 5 on Matching and Subclassification; Causal Inference, The Mixed Tape, by S. Cunningham

– Kaiser, U., Malchow-Møller, N. (2011), ”Is self-employment really a bad experience? The effects of previous self-employment experience on subsequent wage-employment wages”, *Journal of Business Venturing*, 26(5), 572-588. **(applied example)**

Treatment Evaluation

- **Objective:** To measure the **impact** of *interventions* (or *choices*) on outcomes of interest
- Examples (binary treatments)

Origin in medical trials; increasingly popular in economic policy, labor studies, management & strategy

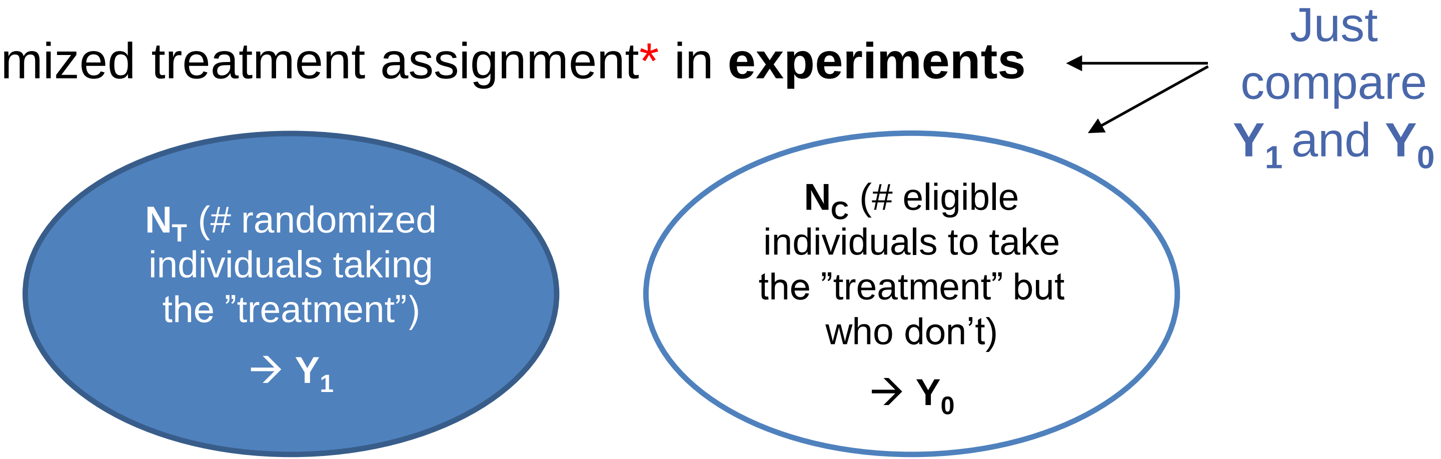
Treatment	**Outcome of interest**
Medical treatment (e.g., new drug)	Health status (e.g., life expectancy)
Enrollment in labor training program	Productivity; Wages; Labor participation
Occupational choice (e.g., self-employment)	Lifetime earnings
Regulatory changes (e.g., tax or entry barriers reductions); subsidies	Individual/firm decisions

The Evaluation Problem

- We want to measure the **response to the *treatment*** relative to **some benchmark** (often *no treatment*)
- BUT no individual is simultaneously observed in both states
- We lack a **counterfactual**: how would the outcome of an average untreated individual change if s/he received the treatment?
- ALSO: Data often do not come from randomized experiments, but from (non-randomized) **observational studies**
    - Probable interferences in the causal connection between the treatment and the outcome

Treatment Effects Framework

Randomized treatment assignment* in **experiments**



Just compare **Y_1** and **Y_0**

N_T (# randomized individuals taking the ”treatment”)

→ **Y_1**

N_C (# eligible individuals to take the ”treatment” but who don’t)

→ **Y_0**

*the assignment to the treatment ignores the possible impact of the treatment on the outcome (Y)

When using **observational data**, there is **no random assignment** to the treatment
⬄ Individuals might choose to be ”treated” or other reasons might not make it random!

Key parameters of interest: ATE & ATET

$$\Delta = \mathbf{Y_1} - \mathbf{Y_0}$$

Δ is not directly observable because no individual can be observed in both states

$$\mathbf{ATE} = \mathrm{E}[\Delta] = \mathrm{E}[\mathrm{Y}_1 - \mathrm{Y}_0]$$

$$\widehat{\mathrm{ATE}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} [\Delta_{\mathrm{i}}]$$

ATE is relevant when the treatment has universal applicability (not relevant for many policy studies; it includes the effect on persons for whom the treatment was never intended)

$$\mathbf{ATET} = \mathrm{E}[\Delta \mid \mathrm{T} = 1]$$

$$\widehat{\mathrm{ATET}} = \frac{1}{\mathrm{N_T}} \sum_{\mathrm{i}=1}^{\mathrm{N_T}} [\Delta_{\mathrm{i}} \mid \mathrm{T} = 1]$$

ATET is relevant to evaluate the effects on those for whom the treatment is actually intended (e.g., diabetes patients; unemployed individuals)

Identification

$$\mathbf{ATET} = \mathrm{E}[\mathrm{Y}_1 - \mathrm{Y}_0\,|\mathrm{T} = 1]$$

$$= \mathrm{E}[\mathrm{Y}_1\,|\mathrm{T} = 1] - \mathrm{E}[\mathrm{Y}_0\,|\mathrm{T} = 1]$$

Not observed!

Can be further decomposed into:

$$= \mathrm{E}[\mathrm{Y}_1\,|\mathrm{T} = 1] - \mathrm{E}[\mathrm{Y}_0\,|\mathrm{T} = 0] + [\mathrm{E}[\mathrm{Y}_0\,|\mathrm{T} = 0] - \mathrm{E}[\mathrm{Y}_0\,|\mathrm{T} = 1]]$$

Observed

Unobserved difference in outcomes for nontreated, had they had the treatment (ideally = 0)

In experiments: We would use $\mathrm{E}[\mathrm{Y}_0\,|\mathrm{T} = 0]$.

In observational data: Not a good idea. What determines the treatment is also likely to determine Y.

So we need a method (e.g., matching) to generate a *comparison group*.

King ~~Prince~~ Charles

Male
Born in 1948
Raised in the UK
Married Twice
Lives in a castle
Wealthy and Famous

Ozzy Osbourne

Male
Born in 1948
Raised in the UK
Married Twice
Lives in a castle
Wealthy and Famous

Key Assumptions

*If **validated**, T is exogeneous and matching methods are suitable. If **violated**, endogeneity is present – **later lectures!***

Conditional Independence Assumption

(also referred to as *unconfoundedness, ignorability or selection on observables* assumption)

$$\mathrm{Y}_0, \mathrm{Y}_1 \perp T \,|\, \boldsymbol{X}$$

Conditional on **X**, the outcomes are independent of the treatment assignment

⬄ **Random assignment to treatment**

If it holds, systematic differences in outcomes for persons with the same X can be attributed to T

(strong assumption, requires good data)

Note: If we are interested in ATET only, it is enough that: $\mathrm{Y}_0 \perp T \,|\, \boldsymbol{X}$

Key Assumptions

*If **violated**, we could have individuals with **X** vectors who are all treated, and those with a different **X** would all be untreated.*

Overlap Assumption

(also referred to as *matching* or *common support* assumption)

$$0 < \Pr\left(\mathrm{T} = 1 \mid \mathbf{X}\right) < 1$$

For each value of **X**, there are **both treated and untreated cases** ⇔ for each treated individual there is another **comparable** (i.e., with similar **X**) untreated individual

<u>In other words:</u> Persons with the same X values have a positive probability of being both treated and non-treated; some randomness is needed that guarantees that persons with identical characteristics can be observed in both states.

The concept of Propensity Score

When participation into the treatment is not random, but depends on a vector of variables **X** (e.g., age, gender), the conditional probability of treatment participation is given by the **Propensity Score** p(**x**):

$$\text{PS:}\ \mathrm{p}(\mathbf{x}) = \Pr\,(\mathrm{T} = 1 \mid \mathbf{X} = \mathbf{x})$$

Estimated by logit/probit

Balancing Condition: $\mathrm{T} \perp \mathbf{x} \mid \mathbf{p}(\mathbf{x})$

For individuals with the same propensity score, the assignment to the treatment is random and should look identical in terms of the **X** vector

Testable!

Matching methods: underlying logic

- If **assignment to T** directly depends on **(only) observed characteristics** of individuals (e.g., age, gender, socio-economic status), we can mimic an experimental setting by generating a **control group**, i.e. sample of **comparable individuals** (with comparable characteristics) who did not take the treatment. (selection on observables)
- But if there are **unobserved factors** that partly determine both T and Y (e.g., individual innate ability), matching methods are not enough to accurately measure ATET (selection on unobservables)
    - We will deal with this issue in later lectures
    - **For today: selection on observables**

Matching is persuasive and attractive if:

- We can control for a rich set of X variables

Treated and control subjects as similar as possible

- There are many potential control units in the data

- ATET is the parameter of interest

Matching methods

Exact matching

- Practicable when the vector of covariates is discrete and the sample contains many observations at each distinct value of x
- Impractical when there are many (continuous) variables to match.
- Possibility: **`cem`** in Stata
- Data-hungry!

Propensity Score Matching

- Rather than matching on X, it **matches on a single metric:** the propensity score p(x)
- Control group = individuals whose p(x) is *sufficiently close* to treated individuals
- Not so hungry in terms of data (as exact matching)
- **`teffects psmatch`** (or **`psmatch2`**) in Stata

When implementing matching, consider:

1) Whether to match *with* or *without* replacement

- *With*: a control individual can be used as a match multiple times (i.e., for multiple treated individuals)

- If *with*: higher matching quality on average (reduced bias), but higher variance

- *Without:* a control individual is matched to no more than one treated individual (more restrictive)

- If *without*: smaller comparison set; matches may not be so close in terms of p(x) → increased bias (lower matching quality), but reduced variance

- TRADE-OFF BETWEEN BIAS AND VARIANCE!

When implementing matching, consider:

2) Number of control units to use in the comparison set

- One single (the closest) match – lower bias, increased variance
- More than one match (”oversampling”) – lower variance, increased bias
  - Some may be poor matches. Possible solution: set a neighborhood for p(x)

3) Choice of matching method NEXT →

- Depends on the data (how rich in terms of X variables and size of comparison group?)
- Depends also on choices made in 1) and 2)

Matching Algorithms (1/2)

Nearest Neighbor (NN)

- For each treated individual choose the match(es) where the **difference in p(x) is smaller**
- Usually **with replacement** (so each untreated individual can be used multiple times as a match)
- **Oversampling** – possible to use more than one NN
- Note that **some matches may be poor** (even though they are the nearest, their PS may be far away)

→ *alternative*

Radius & Caliper Matching

- Establish a **neighborhood for the PS**; all matches falling within that **tolerance level (radius/caliper)** are used as matches
- **Bad (distant) matches are avoided**
- Possible to use one or more NN
- But if tolerance is too small, some units may not get matches

Matching Algorithms (2/2)

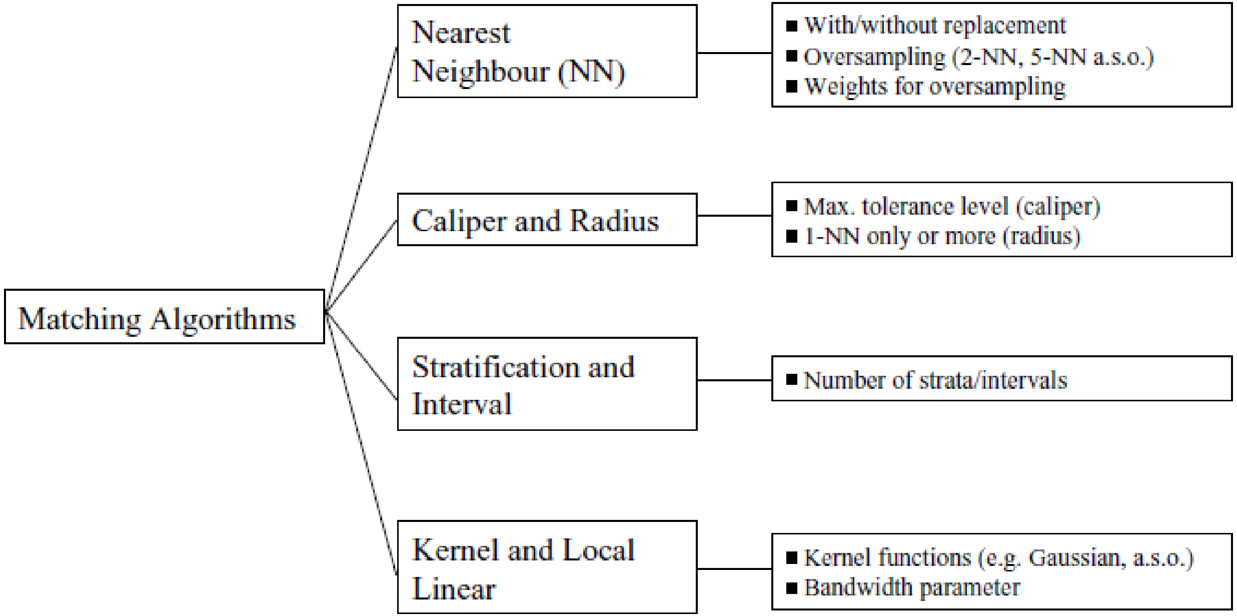
Kernel Matching

- **All treated** units are **matched with a weighted average** of **all control units**
- Weights are **inversely proportional to the Propensity Score distance** between treated and untreated units

Stratification or Interval Matching

- Divides the range of variation of the PS in **intervals** (default = 5)
- Within each interval, **T and C units have, on average, the same PS**
- ATET computed within each interval; global ATET = weighted average, depending on the distribution of treated units across the ”blocks”
- How many intervals? **Balancing condition should be verified within each interval**

How to select a specific matching algorithm?



Matching Algorithms

Nearest Neighbour (NN)
- With/without replacement
- Oversampling (2-NN, 5-NN a.s.o.)
- Weights for oversampling

Caliper and Radius
- Max. tolerance level (caliper)
- 1-NN only or more (radius)

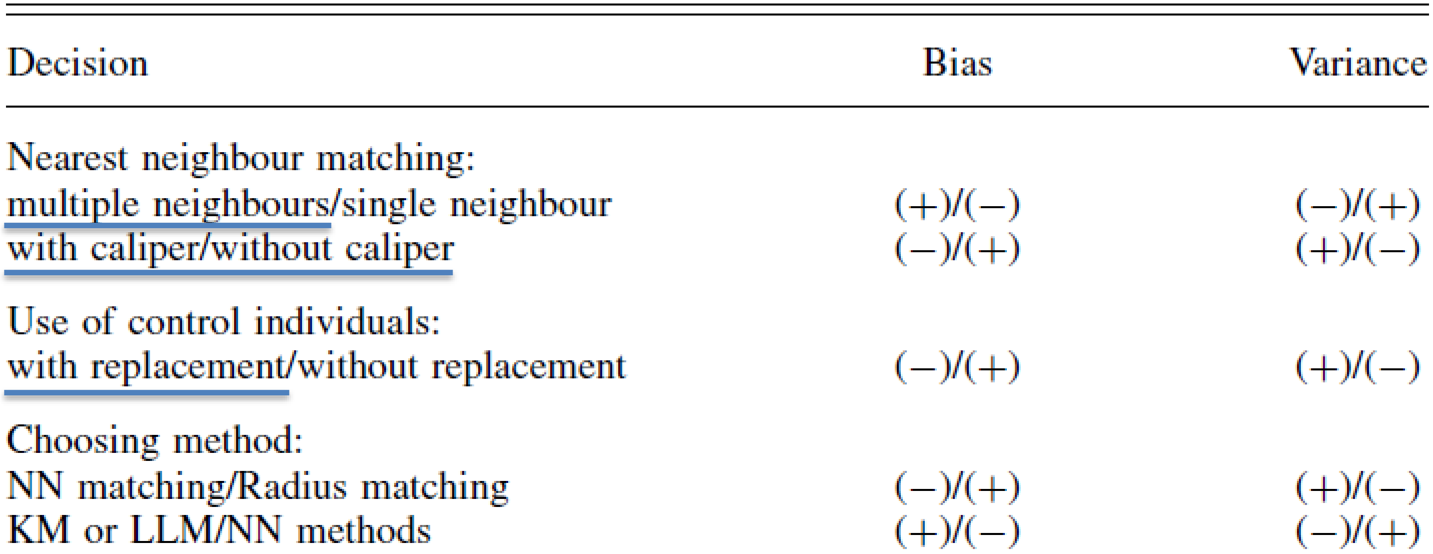
Stratification and Interval
- Number of strata/intervals

Kernel and Local Linear
- Kernel functions (e.g. Gaussian, a.s.o.)
- Bandwidth parameter

Source: Caliendo and Kopeinig (2008)

Trade-offs in terms of bias and efficiency

No winner for all situations! Depends on the data (*recall slides 13 & 16*).
In **smaller samples**, the choice might make a difference; in **larger samples**, all PSM methods should yield asymptotically similar results.



Decision	Bias	Variance
Nearest neighbour matching:		
multiple neighbours/single neighbour	(+)/(−)	(−)/(+)
with caliper/without caliper	(−)/(+)	(+)/(−)
Use of control individuals:		
with replacement/without replacement	(−)/(+)	(+)/(−)
Choosing method:		
NN matching/Radius matching	(−)/(+)	(+)/(−)
KM or LLM/NN methods	(+)/(−)	(−)/(+)

`teffects psmatch`

Source: Caliendo and Kopeinig (2008)

Important to check after matching:

Overlap/Common Support Assumption

- $0 < \Pr\left(T = 1 \mid \mathbf{X}\right) < 1$
- **Visual analysis** of the region of overlap/common support – e.g.:
- Compare min and max of PS in T and C groups
- Compare **density distribution of the PS in both groups**
- If there is too much mass around 0 or 1, the overlap assumption may be violated ⇔ matching estimator not satisfactory

Balancing Condition/Matching Quality

- $T \perp \mathbf{x} \mid \mathbf{p}(\mathbf{x})$
- If matching is properly done, **T and C groups should have no significant differences in terms of X variables**
- E.g., **t-test** in covariate means; **`tebalance`** after `teffects psmatch`
- If matching quality is not enough, **improve the estimation of the PS** (e.g., adding more variables, non-linear terms on some covariates, interaction terms)

Research example

Is self-employment really a bad experience? The effects of previous self-employment on subsequent wage-employment wages

Kaiser & Malchow-Møller (2011), JBV

The research setting

Treatment vs Control groups

Treatment: past self-employment (SE) experience (binary)

>>> then several different types of treatments

Outcome: subsequent earnings in wage employment (WE) – log (hourly wage) in 1996

Data: Danish men observed between 1990 and 1996; full-time wage-employed in both 1990 and 1996

PSM methods to estimate **ATET**

Ideal experiment: a fraction of individuals would be **randomly allocated to a (short) spell of SE** before being returned to WE, while the remaining individuals would be kept in WE during the whole period.

Observational data: having had SE experience is a **choice**; individuals with and without SE experience are likely to differ in a variety of characteristics

PSM to find **”clones”** (in terms of **observable characteristics**) for each individual with SE experience

Applying PSM to estimate the ATET

*”…given a set of observable characteristics, **x**, – **which is not affected by treatment** – potential outcomes are independent of the assignment to treatment”* ⇔ **conditional independence or unconfoundedness assumption** (recall slide 9)

*”we cannot formally test if [this] assumption is satisfied. We do **formally test whether T and C observations no longer differ significantly wrt. observable characteristics after matching**”* ⇔ **balancing condition** (recall slides 11 & 21)

1-to-1 matching infeasible: why?

PSM instead: Probability of **treatment (spell in SE)** estimated with a **probit** model

Wide set of **X that affect both T assignment and Y** (e.g., tenure, age, sector, education, family background, regional conditions, employer characteristics, initial wage in 1990).

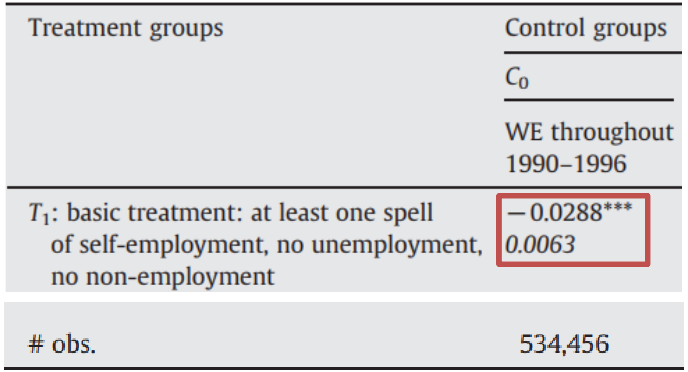
Several matching algorithms tested; **preference for nearest neighbor (single) with replacement** *”since it reduces estimation bias at the cost of higher variance”*

Applying PSM to estimate the (general) ATET

Before matching: former SE individuals **earn on average** (not controlling for X) **4% more** than consecutively employed individuals.

They also **differ in other observed characteristics**: e.g., shorter tenure, higher education levels, employed in smaller establishments.

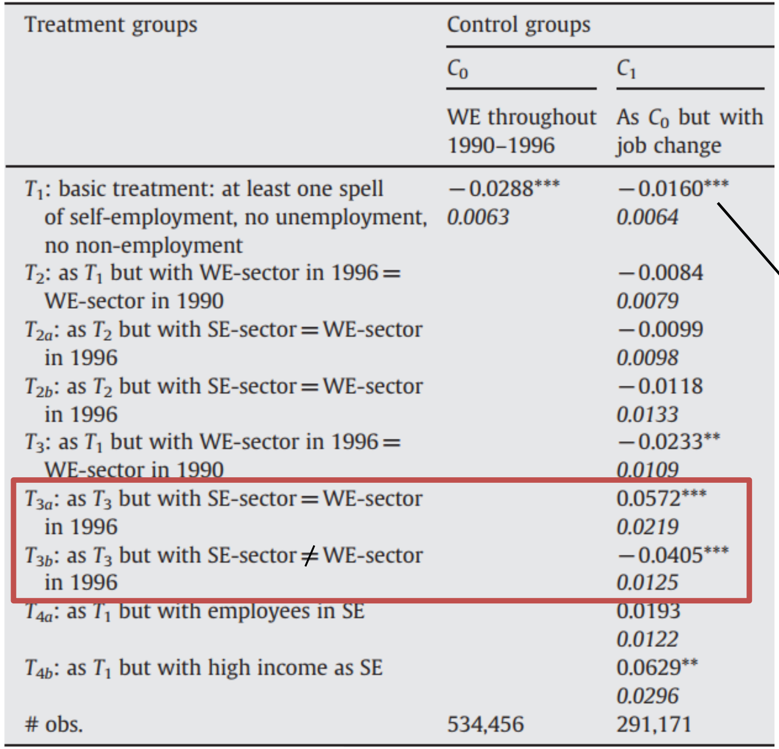
Differences (in **X**) no longer significant **after matching: balancing property satisfied**



Treatment groups	Control groups
	C_0
	WE throughout 1990–1996
T_1: basic treatment: at least one spell of self-employment, no unemployment, no non-employment	−0.0288*** *0.0063*
# obs.	534,456

ATET: a spell of SE between 1990 and 1996 goes along with a **reduction in hourly wages in subsequent wage employment of 2.9%** (vs. OLS coefficient: 0.0013***)

Still possible to study more specific treatments



Treatment groups	Control groups	
	C_0	C_1
	WE throughout 1990–1996	As C_0 but with job change
T_1: basic treatment: at least one spell of self-employment, no unemployment, no non-employment	−0.0288*** *0.0063*	−0.0160*** *0.0064*
T_2: as T_1 but with WE-sector in 1996 = WE-sector in 1990		−0.0084 *0.0079*
T_{2a}: as T_2 but with SE-sector = WE-sector in 1996		−0.0099 *0.0098*
T_{2b}: as T_2 but with SE-sector = WE-sector in 1996		−0.0118 *0.0133*
T_3: as T_1 but with WE-sector in 1996 = WE-sector in 1990		−0.0233** *0.0109*
T_{3a}: as T_3 but with SE-sector = WE-sector in 1996		0.0572*** *0.0219*
T_{3b}: as T_3 but with SE-sector ≠ WE-sector in 1996		−0.0405*** *0.0125*
T_{4a}: as T_1 but with employees in SE		0.0193 *0.0122*
T_{4b}: as T_1 but with high income as SE		0.0629** *0.0296*
# obs.	534,456	291,171

More refined control group: individuals who also experienced a job change in WE

Basic Treatment: wage reduction of 1.6%

Specific Treatment of SE in the same (different) sector of WE: wage increase of 5.7% (wage decrease of 4.1%) *(OLS finds a 4% wage premium in both)*

Propensity Score Matching in Stata

- For this course, give priority to `teffects psmatch`
- ***Note:*** `teffects psmatch` only allows matching with replacement and uses all ”good neighbours” as matches
- An alternative could be the user-written command `psmatch2` (check syntax and description here). It allows several other matching algorithms (as discussed before).
- However, be aware that the standard errors obtained with `psmatch2` are not correctly estimated. `teffects psmatch` is therefore preferred, or should be used in combination with `psmatch2.` A comparison between the two can be found here.

NOTE: For R, check chapter 5. Matching and subclassification, especially section 5.3.6. Nearest Neighbor Matching

teffects psmatch: Quick tour

```
teffects psmatch (y) (t x1 x2), atet
```

>> Estimates the **ATET of t on y by PSM**, using a logit model for t on x1 and x2. If probit is preferred, add ”probit” option. If ATE is needed, drop the ATET option.

```
teffects psmatch (y) (t x1 x2), nn(3) gen(match)
```

>> At least 3 matches should be used per observation (default is 1)

>> **gen(match)** creates a new variable containing the ID of the nearest match(es)

```
predict ps0 ps1, ps
```

>> predicts propensity score (probability of being in T vs C groups)

```
predict y0 y1, po
```

>> predicts potential outcomes

```
teffects overlap
```

>> produces graph to analyze overlap assumption

(…)

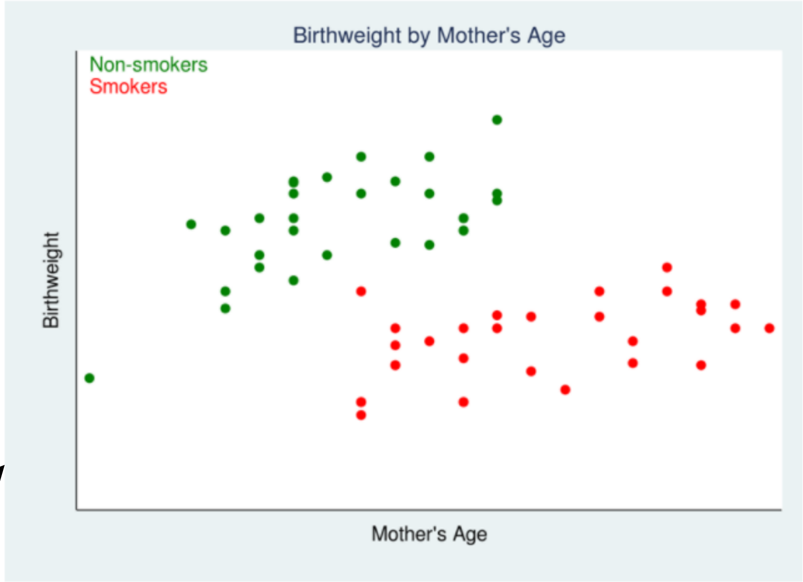
```
help teffects psmatch
help teffects_postestimation
```

Stata Example

Effect of maternal smoking during pregnancy (**treatment**) on baby's weight at birth (**outcome**)

Why can't we estimate this effect by comparing the birth weights of babies of smoking vs. non-smoking mothers?

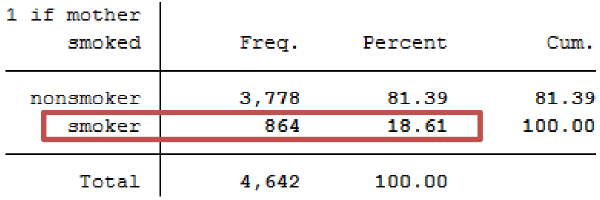
Mothers are **not randomly assigned to "smoking status"; it is their choice.** Other patterns?



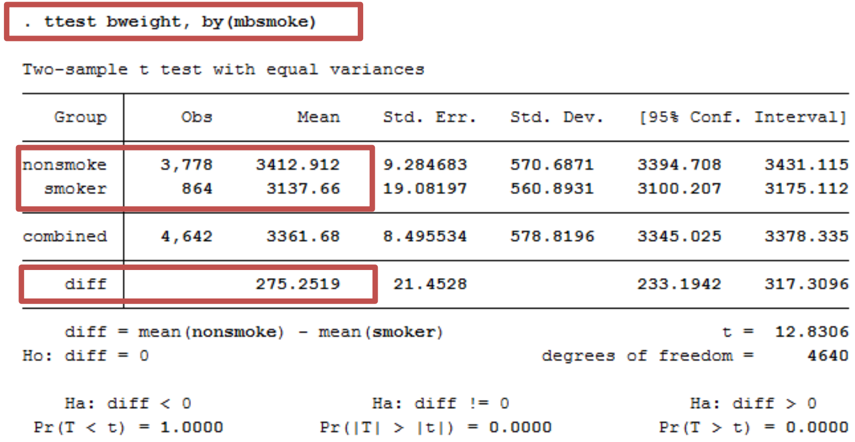
(selected data points)

Before treatment evaluation: descriptives

- How many women are in the ”treatment group” (i.e., smoked during pregnancy)?
- Observed difference in babies’ birthweight of smoking vs. non-smoking mothers?
- **Other observed differences between smoking vs. non-smoking mothers?** e.g.:
    - Marital status?
    - Education level?
    - First pregnancy?



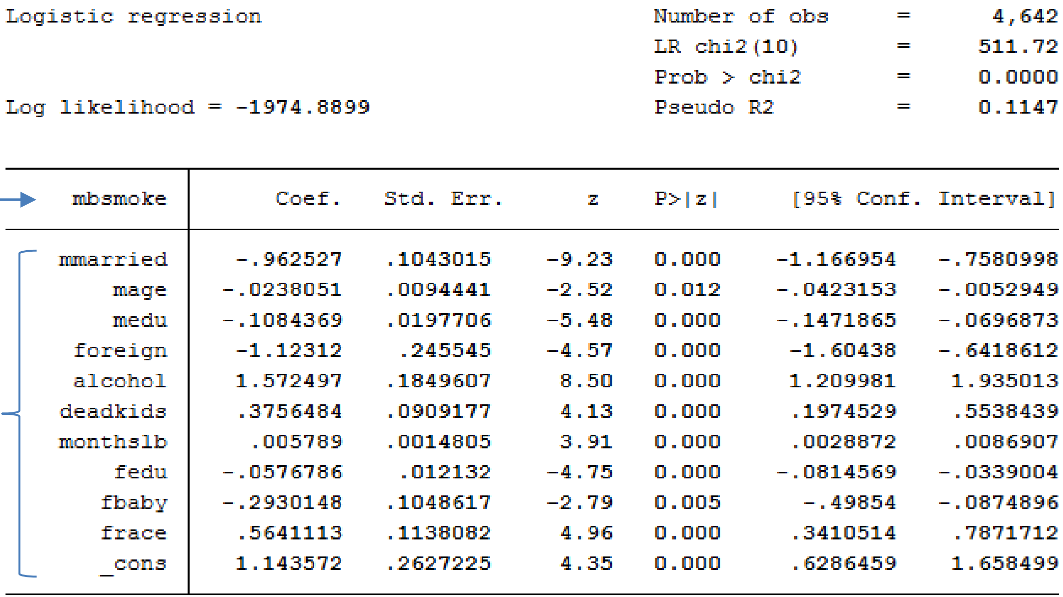
```
1 if mother |
     smoked |      Freq.     Percent        Cum.
------------+-----------------------------------
  nonsmoker |      3,778       81.39       81.39
     smoker |        864       18.61      100.00
------------+-----------------------------------
      Total |      4,642      100.00
```



```
. ttest bweight, by(mbsmoke)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
nonsmoke |   3,778    3412.912    9.284683    570.6871    3394.708    3431.115
  smoker |     864     3137.66    19.08197    560.8931    3100.207    3175.112
---------+--------------------------------------------------------------------
combined |   4,642     3361.68    8.495534    578.8196    3345.025    3378.335
---------+--------------------------------------------------------------------
    diff |            275.2519     21.4528                233.1942    317.3096
------------------------------------------------------------------------------
    diff = mean(nonsmoke) - mean(smoker)                          t =  12.8306
Ho: diff = 0                                     degrees of freedom =     4640

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Critical to select which variables predict the Propensity Score →

What explains the prob(treatment)?



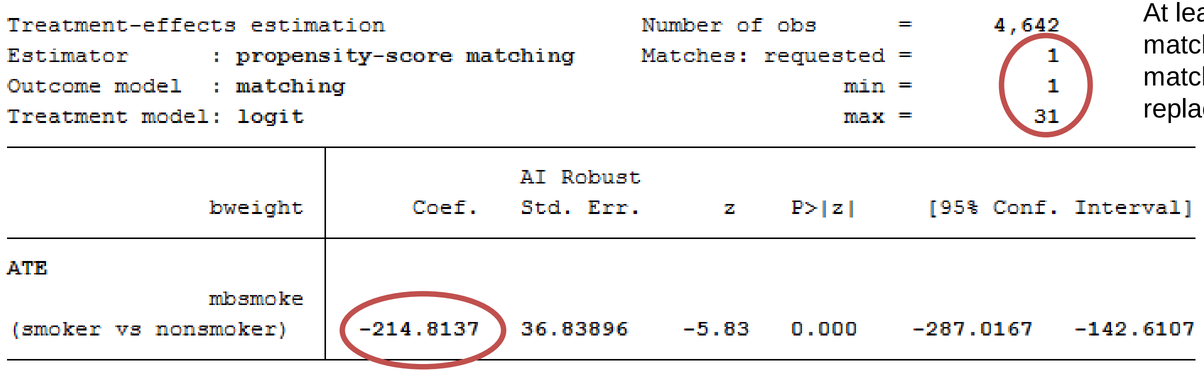
```
Logistic regression                             Number of obs     =      4,642
                                                LR chi2(10)       =     511.72
                                                Prob > chi2       =     0.0000
Log likelihood = -1974.8899                     Pseudo R2         =     0.1147

------------------------------------------------------------------------------
     mbsmoke |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mmarried |   -.962527   .1043015    -9.23   0.000    -1.166954   -.7580998
        mage |  -.0238051   .0094441    -2.52   0.012    -.0423153   -.0052949
        medu |  -.1084369   .0197706    -5.48   0.000    -.1471865   -.0696873
     foreign |   -1.12312    .245545    -4.57   0.000     -1.60438   -.6418612
     alcohol |   1.572497   .1849607     8.50   0.000     1.209981    1.935013
    deadkids |   .3756484   .0909177     4.13   0.000     .1974529    .5538439
    monthslb |    .005789   .0014805     3.91   0.000     .0028872    .0086907
        fedu |  -.0576786    .012132    -4.75   0.000    -.0814569   -.0339004
       fbaby |  -.2930148   .1048617    -2.79   0.005      -.49854   -.0874896
       frace |   .5641113   .1138082     4.96   0.000     .3410514    .7871712
       _cons |   1.143572   .2627225     4.35   0.000     .6286459    1.658499
------------------------------------------------------------------------------
```

treatment →

Use pairwise correlations between X and the treatment variable & t-tests to identify observed differences between groups

Estimating the ATE

```
teffects psmatch (bweight) (mbsmoke mmarried mage medu foreign
alcohol deadkids monthslb fedu fbaby frace)
```



```
Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                     min =          1
Treatment model: logit                                        max =         31
------------------------------------------------------------------------------
                         |              AI Robust
                 bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+----------------------------------------------------
ATE                      |
                 mbsmoke |
(smoker vs nonsmoker)    |  -214.8137   36.83896    -5.83   0.000    -287.0167   -142.6107
------------------------------------------------------------------------------
```

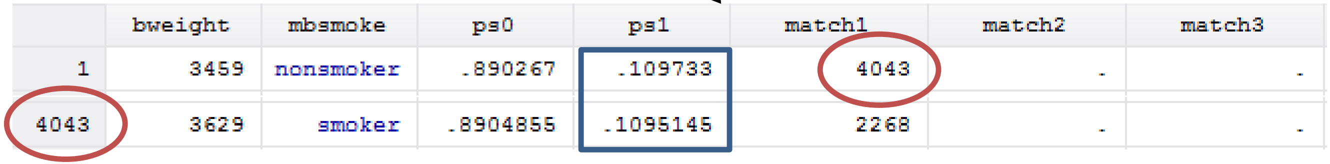
At least one (the nearest) match, up to 31 close matches – matching with replacement

The average birthweight **if all mothers were to smoke** would be **215 grams less** than the average that would occur if none of the mothers had smoked.

(ATE relevant here: treatment not ”intended”/targeted to a group)

Stata behind the scenes…

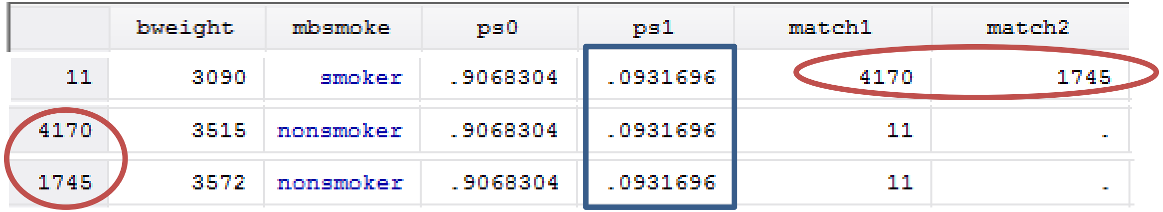
`predict ps0 ps1, ps`



	bweight	mbsmoke	ps0	ps1	match1	match2	match3
1	3459	nonsmoker	.890267	.109733	4043	.	.
4043	3629	smoker	.8904855	.1095145	2268	.	.

Individual #1, a non-smoker, is matched with only one smoker (individual #4043), who has the closest propensity score.

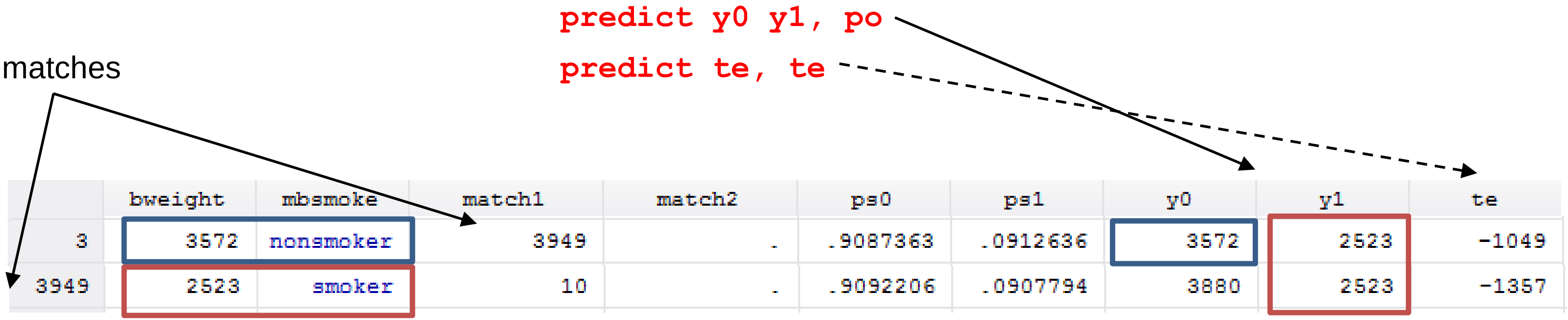
This individual is then matched to #2268, who has an even closer PS (0.109531).



	bweight	mbsmoke	ps0	ps1	match1	match2
11	3090	smoker	.9068304	.0931696	4170	1745
4170	3515	nonsmoker	.9068304	.0931696	11	.
1745	3572	nonsmoker	.9068304	.0931696	11	.

Individual #11 (treated) finds two good matches (untreated) with the same PS

Stata behind the scenes…



matches

```
predict y0 y1, po
predict te, te
```

	bweight	mbsmoke	match1	match2	ps0	ps1	y0	y1	te
3	3572	nonsmoker	3949	.	.9087363	.0912636	3572	2523	-1049
3949	2523	smoker	10	.	.9092206	.0907794	3880	2523	-1357

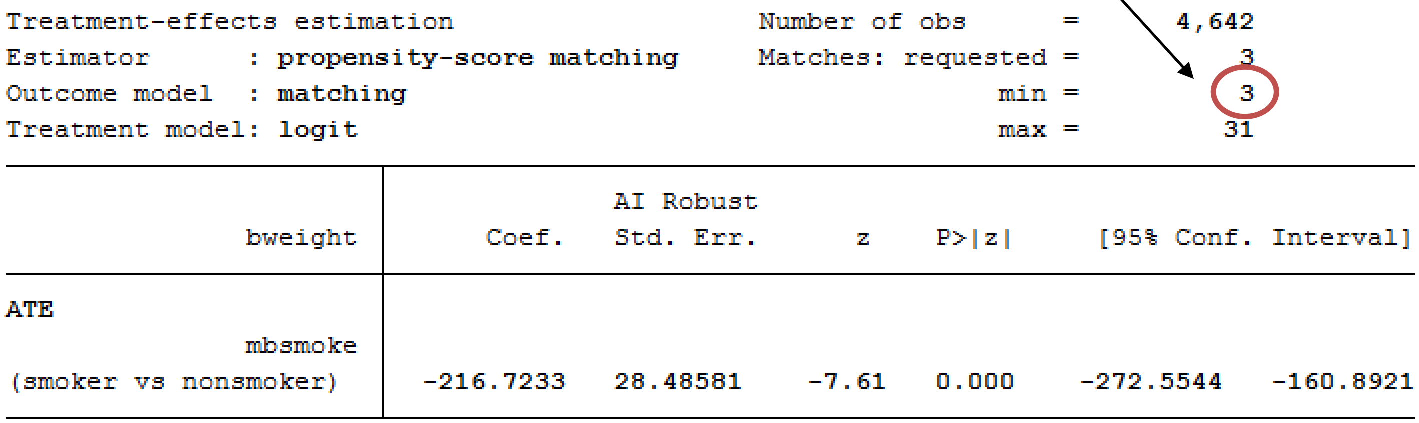
If ”nonsmoker”: y0 = observed bweight and y1 = bweight of the matched ”smoker”

Treatment effect = y1-y0

ATE = average of ”te”
ATET = average of ”te” for ”smokers”

Variations: Minimum number of matches

```
teffects psmatch (bweight) (mbsmoke mmarried mage medu foreign
alcohol deadkids monthslb fedu fbaby frace), nn(3)
```



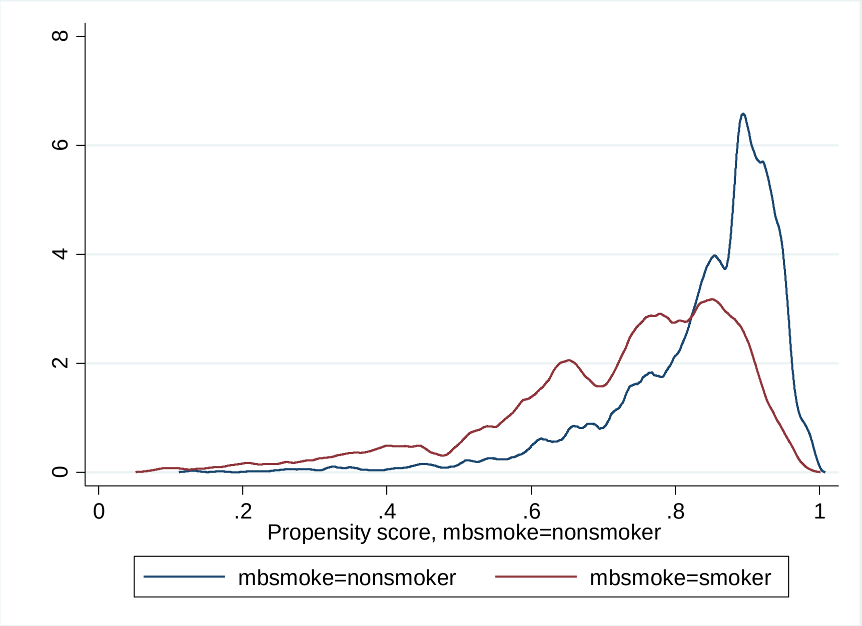
```
Treatment-effects estimation                    Number of obs      =      4,642
Estimator      : propensity-score matching      Matches: requested =          3
Outcome model  : matching                                      min =          3
Treatment model: logit                                         max =         31
-----------------------------------------------------------------------------------
                       |              AI Robust
               bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+-----------------------------------------------------------
ATE                    |
               mbsmoke |
 (smoker vs nonsmoker) |  -216.7233   28.48581    -7.61   0.000    -272.5544   -160.8921
-----------------------------------------------------------------------------------
```

Remember the trade-off between bias and variance!

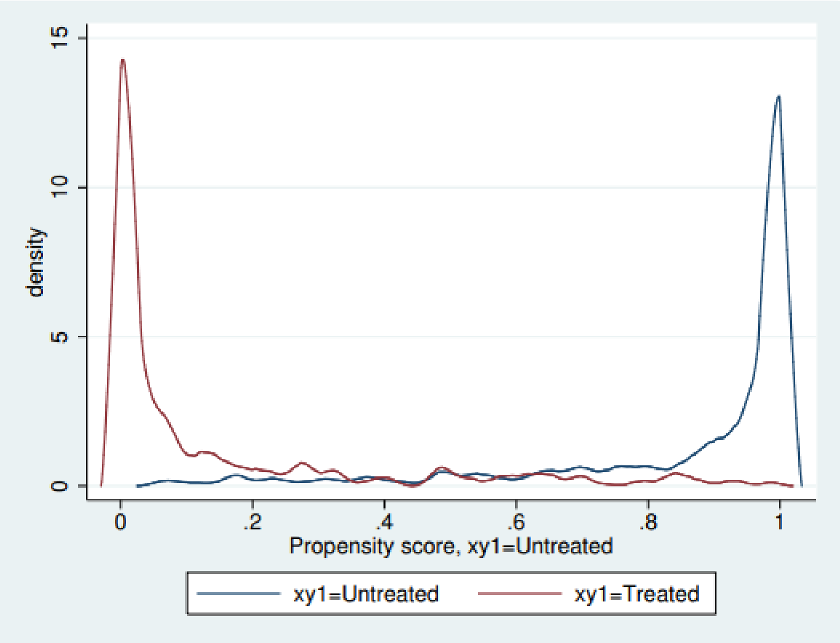
Overlap assumption

- Estimated density of the predicted probabilities
- Ideally not too much mass around 0 or 1
- The two estimated densities have most of their respective masses in regions in which they overlap each other
- **No evidence that the overlap assumption is violated.**

```
teffects overlap
```

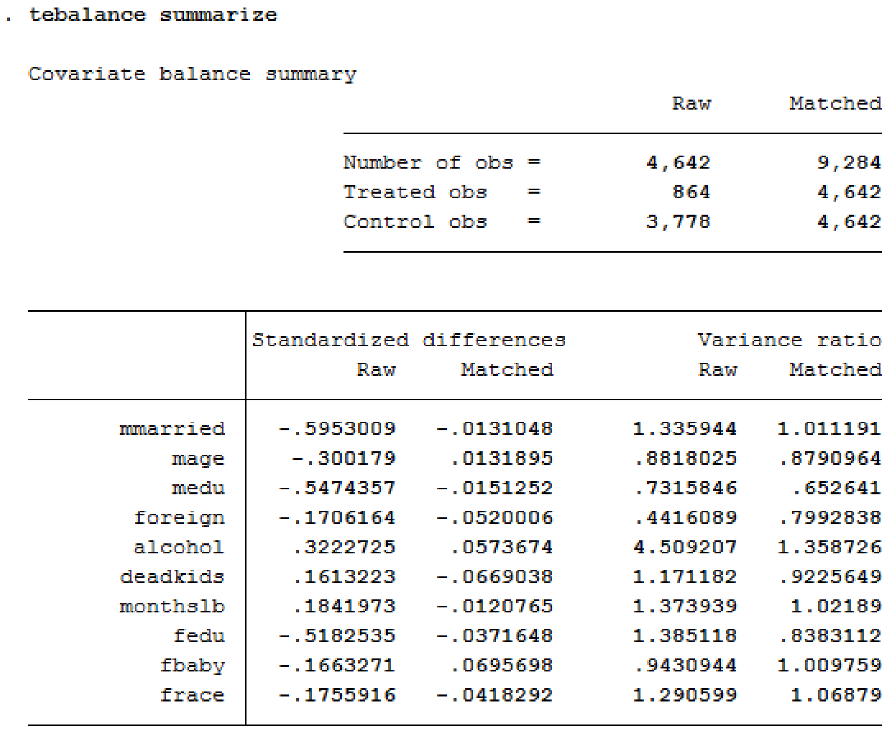


Violating Overlap Assumption: Example



Checking balancing condition (1/2)

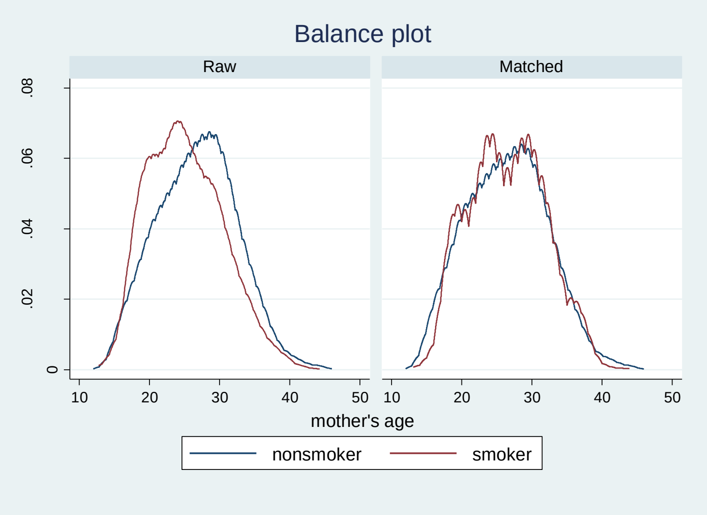
- A covariate is said to be **balanced** when its distribution does not vary over treatment levels.
- A perfectly balanced covariate would have a **standardized difference of 0 and a variance ratio of 1.**
- **Improved level of balance for all variables**, though for some it could be better (e.g., mother age/education).
- To **try to achieve better balance**, we could specify a richer model for the PS (e.g., with interactions between some covariates).



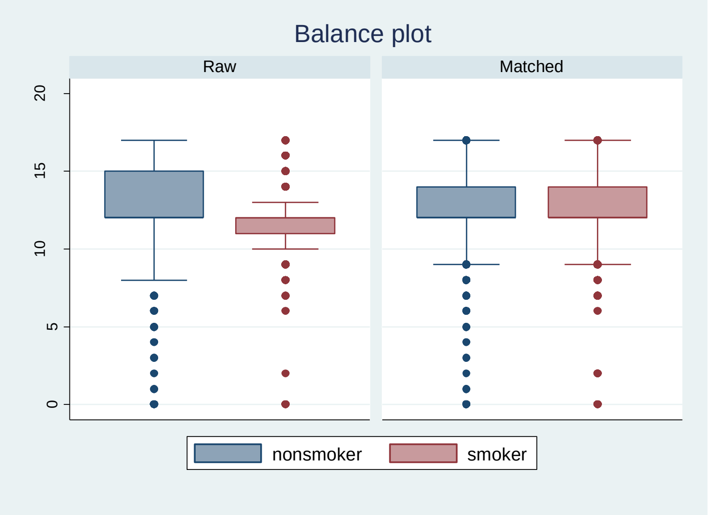
```
. tebalance summarize

  Covariate balance summary
                                               Raw      Matched
                      -----------------------------------------
                      Number of obs =        4,642        9,284
                      Treated obs   =          864        4,642
                      Control obs   =        3,778        4,642
                      -----------------------------------------

  ----------------------------------------------------------------
                 |Standardized differences          Variance ratio
                 |        Raw        Matched           Raw   Matched
  ---------------+------------------------------------------------
        mmarried |  -.5953009   -.0131048      1.335944   1.011191
            mage |   -.300179    .0131895      .8818025   .8790964
            medu |  -.5474357   -.0151252      .7315846    .652641
         foreign |  -.1706164   -.0520006      .4416089   .7992838
         alcohol |   .3222725    .0573674      4.509207   1.358726
        deadkids |   .1613223   -.0669038      1.171182   .9225649
        monthslb |   .1841973   -.0120765      1.373939    1.02189
            fedu |  -.5182535   -.0371648      1.385118   .8383112
           fbaby |  -.1663271    .0695698      .9430944   1.009759
           frace |  -.1755916   -.0418292      1.290599    1.06879
  ----------------------------------------------------------------
```

Checking balancing condition (2/2)

```
tebalance density mage
```



```
tebalance box medu
```

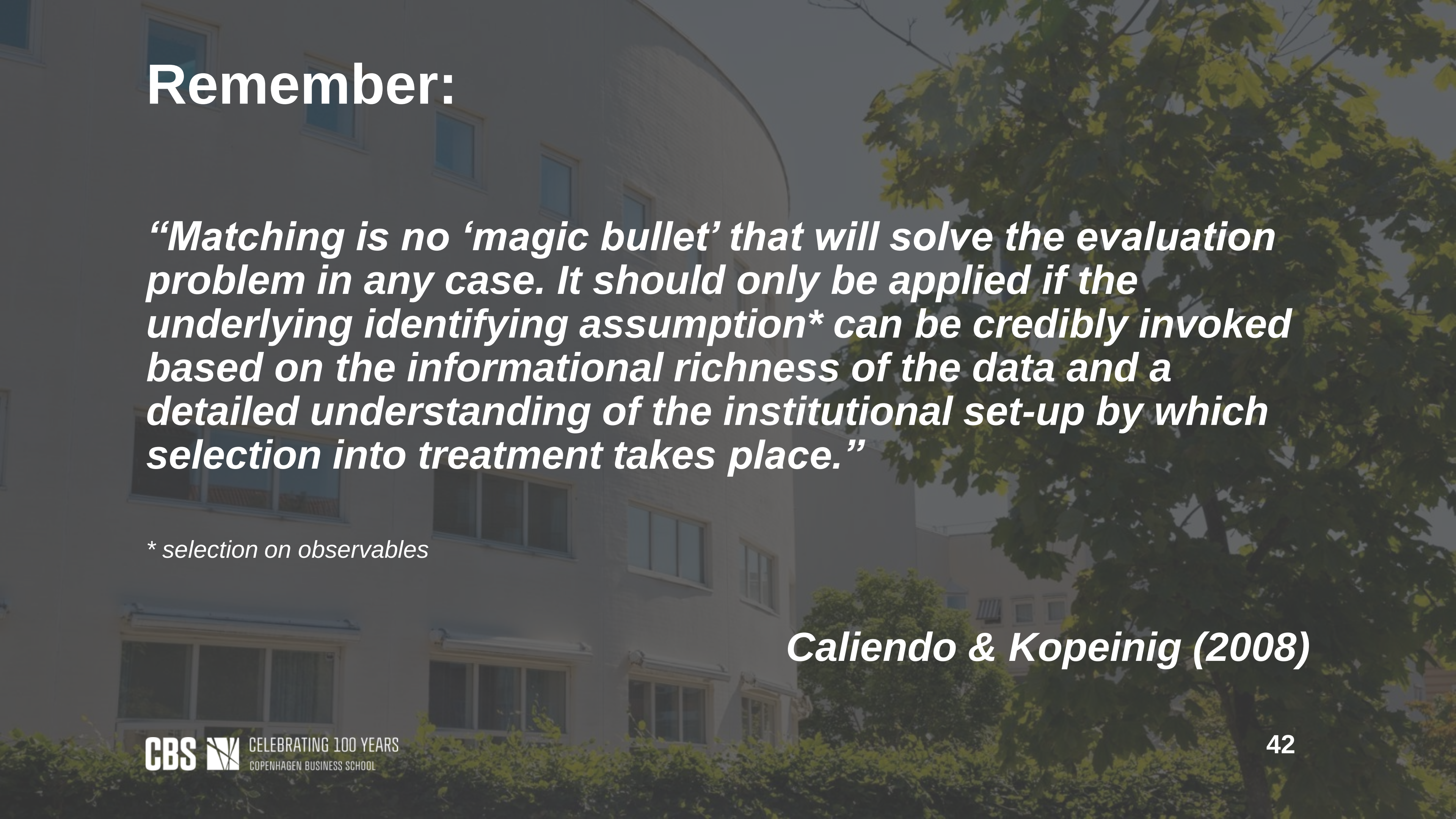


Wrap-up of today and next sessions

	Heckman models (11/09)	Matching Models (PSM) (11/16)	Instrumental Variables (11/23)
When	Y is missing in some cases (for a non-random reason)	X is a binary intervention/choice	…
Problem	The missings in Y are driven by a "selection process"	T & C groups are very different	…
Stata commands	*heckman, (twostep)*	*teffects psmatch, tebalance, teffects overlap*	…
Key tests	Significance of the IMR or of the *rho*	Balancing and overlapping conditions	…
Attention!	Need for valid exclusion restrictions; selection bias important when *IMR/rho* significant and *X* predicts selection (1st stage)	T & C only matched on observable characteristics. If unobservables matter, PSM does not provide causal effects → IV	…
First stage	Probit predicting selection into the sample (Y ≠ missing)	Probit predicting probability of being treated (X)	…

Your roadmap when implementing PSM

1. **Model choice** (logit/probit) and **variable choice** (guided by theory and prior evidence; X should not be influenced by the treatment!) Remember “Conditional Independence Assumption”
2. **Matching algorithm**: NN? Caliper? Number of NN? With or without replacement? (some Stata commands may limit your choices)
3. Check **overlap assumption** (visually + min/max of PS)
4. Check **balancing condition** (table of mean differences and variance ratio before and after matching; plots)
5. Satisfactory results? If not, **iterate and improve** PS model estimation (e.g. adding interaction terms, adding or removing variables) and follow the list again until you can “trust” your **ATET**



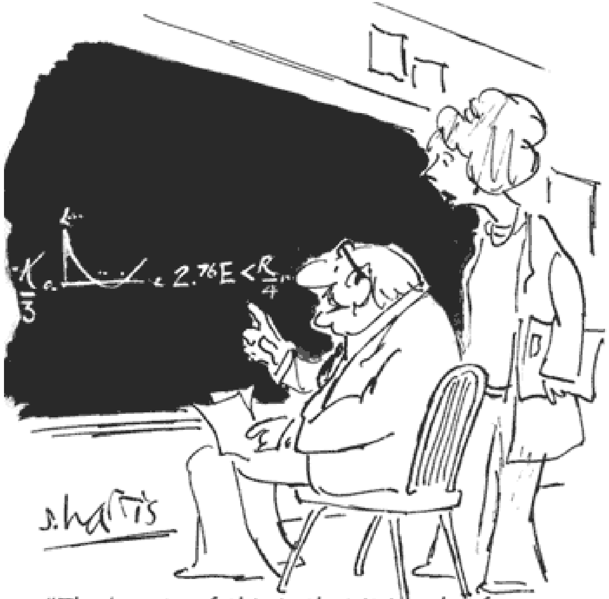
Remember:

“Matching is no ‘magic bullet’ that will solve the evaluation problem in any case. It should only be applied if the underlying identifying assumption can be credibly invoked based on the informational richness of the data and a detailed understanding of the institutional set-up by which selection into treatment takes place.”*

** selection on observables*

Caliendo & Kopeinig (2008)

Hopefully you disagree ☺



"The beauty of this is that it is only of theoretical importance, and there is no way it can be of any practical use whatsoever."