upf. Universitat Pompeu Fabra Barcelona

RECSM
Research and Expertise Centre for Survey Methodology

Propensity score matching with clustered data in Stata

Bruno Arpino
Pompeu Fabra University
bruno.arpino@upf.edu
https://sites.google.com/site/brunoarpino

11th Spanish Stata Conference 2018
Barcelona, 24 de octubre de 2018

Outline

- Brief intro to Propensity Score Matching (PSM) for estimation of causal effects in observational studies

- PSM with **clustered** (multilevel, hierarchical) data

- PSM in Stata
  - Available routines
  - How to implement PSM with clustered data

Do-file and dataset to replicate the analyses in these slides can be found at: https://sites.google.com/site/brunoarpino/software

Motivating case study (1/3)

- Goal: estimating the causal effect of doing **homeworks** on **mathematical proficiency**
- We use a subset of the **National Education Longitudinal Study of 1988** (NELS-88), a nationally representative, longitudinal study of 8th graders in 1988 in the US
- Our data is a subsample of the original full NELS-88 dataset provided by Kraft and de Leeuw (1998)

Motivating case study (2/3)

- **Treatment**: T = 1 for students that spend at least 1 hour doing math homeworks per week; 0 otherwise
- **Outcome**: Y, is the score on a math test
- The dataset contains 260 students from 10 schools and several potential **confounders** on both students (X) and schools (Z)

Motivating case study (3/3)

- **Selection mechanism**: what are the factors influencing time spent doing homework (that may also influence math proficiency)? For the sake of illustration we only consider:
- **Individual-level**: *ses* (a standardised continuous measure of family socio-economic status), *male* (1 = male; 0 = female) and *white* (1 = white; 0 = other race)
- **School-level**: *public* (1 = public schools; 0 = private)

Potential outcome framework

- Let T be the binary treatment indicator:

  = 1 at least 1 hour doing math homeworks per week;

  = 0 otherwise

- Let Y(1) and Y(0) denote the potential outcomes, i.e. math score we would observe if students were assigned to the treatment or control group, respectively

- Causal estimand of interest: ATT = E[Y(1) - Y(0) | T = 1]

- Y(0) is always unobserved for treated students (T = 1)

Propensity score (PS) methods

- **Identifying assumptions**:
  - $Y(1), Y(0) \perp T \mid X, Z$ (unconfoundedness)
  - $0 < P(T=1 \mid X, Z) < 1$ (overlap)

- PS: $e(X) \equiv \Pr\{T = 1 \mid X, Z\} = E\{T \mid X, Z\}$.
- Rosenbaum and Rubin (1983):
  - the propensity score is a balancing score, i.e., $X, Z \perp T \mid e(X, Z)$
  - if unconfoundedness holds, then $Y(1), Y(0) \perp T \mid e(X, Z)$

- These results justify matching / stratification / weighting on e(X, Z) instead than on (X, Z)

PSM as a two-step procedure

- **Design phase:** match similar treated and control individuals to make them as similar as possible in terms of (X, Z)

- **Outcome phase:** estimate causal effects on the matched data

- It reduces model dependence (extrapolation; Drake, 1993)
- It increses objective causal inference (Rubin, 2008)
- Matching as a data pre-processing (Ho et al., 2007)

Clustered data structures

- Very common in many fields (patients into hospitals, individuals into geographical areas, students into schools)
- Few methodological and applied works exist in the case of clustered data
- In clustered data bias can arise from omitted individual and/or cluster-level confounders
- **How should we apply PS methods to these data?**
- **How can we exploit knowledge on clusters' memberships?**

Existing studies with clustered data

- Arpino and Mealli (2011)
  - Show the benefit of using random or fixed effects models for the estimation of the propensity score to reduce the bias due to unmeasured cluster-level variables in PS matching (PSM)
  - Focus on high number of small clusters

- Thoemmes and West (2011) and Li et al (2013) considered stratification and re-weighting using PS, respectively

Arpino and Cannas (2016)

- **Unbalanced** data structure with both big and small clusters
- We compare different approaches:

Approaches	**PS model**	**Matching**
Naïve (NV)	Single-level logit	Pooled
Within (W)	Single-level logit	Within-cluster
Preferential (PW)	Single-level logit	“**Preferential**” within-cluster
Random-effects (RE)	Random-effect logit	Pooled
Fixed-effects (RE)	Fixed-effect logit	Pooled

- R package: CMatching

«Naïve» approach

- It ignores the clustered structure in both PS estimation:

$$\operatorname{logit}(e_{ij}) = \alpha_0 + X_{ij}\beta \qquad (1)$$

- and matching

$$A_{rj} = \{kj' \in I_0 : \hat{e}_{kj'} = \min_{kj' \in I_0} \left| \hat{e}_{rj} - \hat{e}_{kj'} \right| < 0.25\hat{\sigma}_e\} \qquad (2)$$

- We use **one-to-one nearest neighbor** matching within a **caliper** of 0.25 standard deviations of the estimated PS (with replacement)

Estimating the ATT

- The matched dataset is built as the subset of treated and control units that have been matched:

$$M = \{rj : A_{rj} \neq \emptyset\} \cup \left\{ \bigcup_{rj} A_{rj} \right\} \quad (3)$$

- and the ATT is estimated on this set using:

$$A\hat{T}T = \frac{1}{card(M)} \left\{ \sum_{rj \in I_1 \cap M} \left(Y_{rj} - \sum_{kj'} Y_{kj'} w(rj, kj') \right) \right\} \quad (4)$$

«Within» approach

- Uses the same PS model than method A (2) but adjusts for clustering in the implementation of the matching that is forced to be within-cluster:

$$A_{rj} = \{kj' \in I_0 : \hat{e}_{kj'} = \min_{kj' \in I_0} \left| \hat{e}_{rj} - \hat{e}_{kj'} \right| < 0.25\hat{\sigma}_e ;\ \ j = j'\} \quad (5)$$

- Automatically guarantees that all cluster-level variables are perfectly balanced. But, balance of individual-level variables could be worse than with the “Naïve” approach. Also the no. of unmatched units will be higher.

«Preferential» approach

- Tries to combine the benefits of the previous two approaches (“Naïve” and “Within”).

- Starts by searching control units within-cluster (according to (5)). If none is found, control units are searched in other clusters (according to (2)).

- It is expected to improve the balancing of cluster-level variables with respect to the “Naïve” approach and reduces the loss of units compared to the “Within” approach.

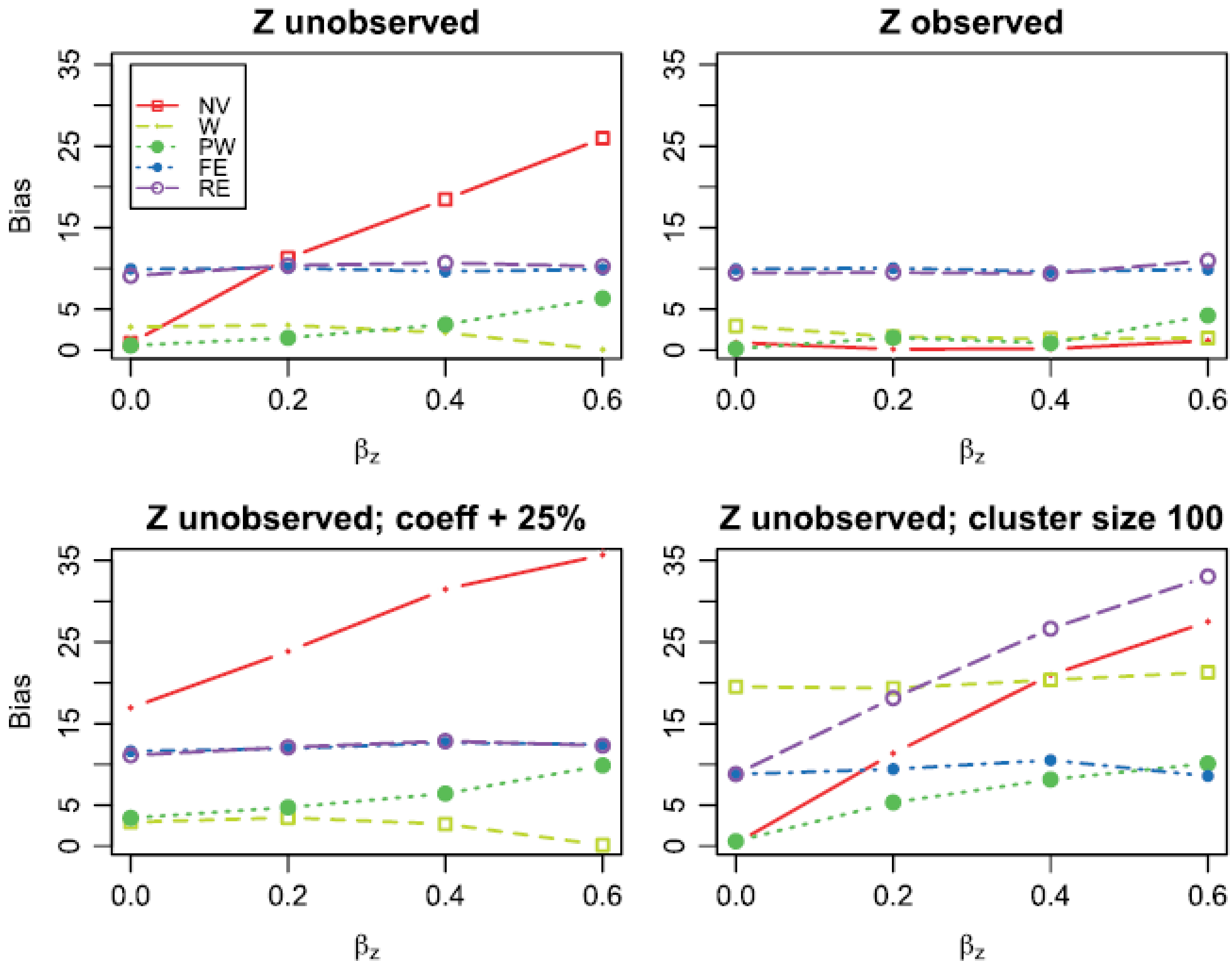
«Random-effects» and «Fixed-effects» approaches

- They keep clustering into account in the estimation of the propensity score:

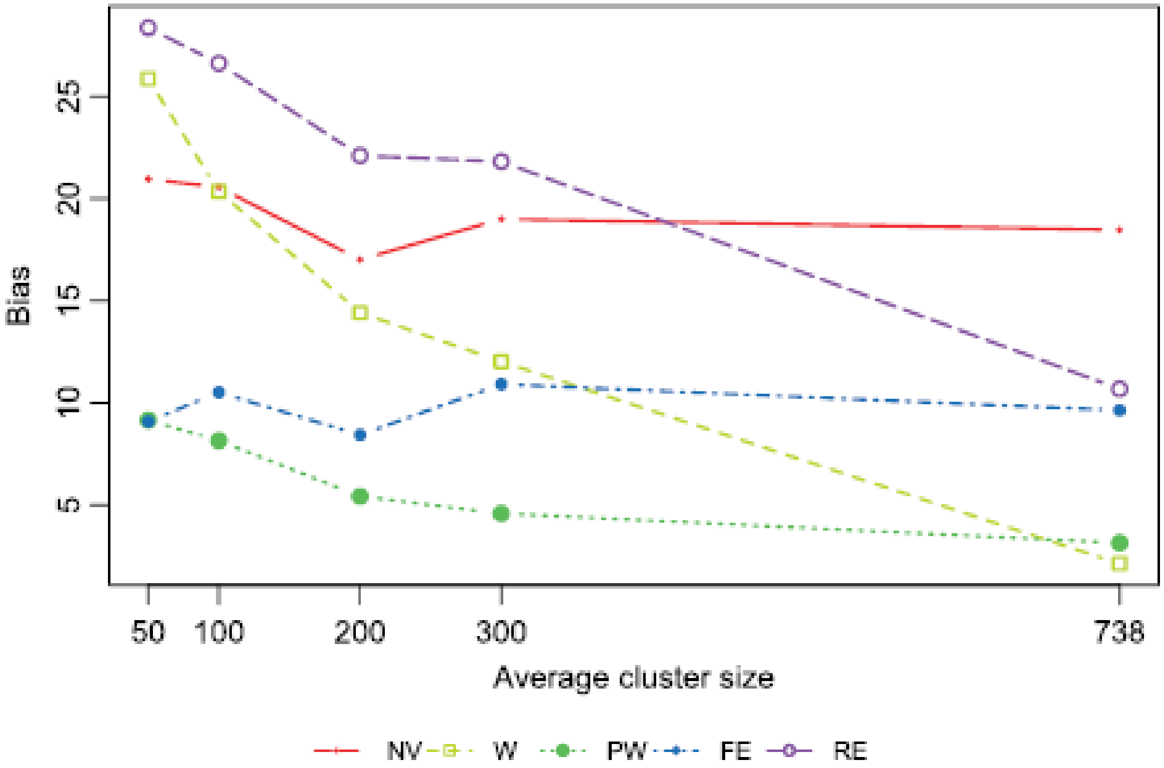
$$\text{logit}\,(e_{ij}) = \alpha_j + X_{ij}\beta \qquad (6)$$

by estimating cluster-specific random or fixed intercepts, respectively (Arpino and Mealli, 2011).

Simulation results (1/2)



Simulation results (2/2)



Note: βz, overall sampe size, etc. are kept fixed. Z is unobserved.

Implementing matching in Stata

- **psmatch2** (Leuven and Sianesi 2003)
  - PSM and covariate matching
  - severalalgorithms (nn and caliper matching (with and w/o replacement), kernel, radius, local linear matching
  - common support plots (psgraph) and covariate **imbalance** testing (pstest)
  - standard errors obtained using bootstrap methods or variance approximation
- **nnmatch** (Abadie, Drukker, Herr, and Imbens 2004)
  - nearest neighbour matching with different distance metrics (replacement allowed)
  - allows **exact matching** (or as close as possible) on a subset of variables
  - allows for bias correction
  - sample or population variance, with or w/o assuming a constant treatment effects

Implementing matching in Stata

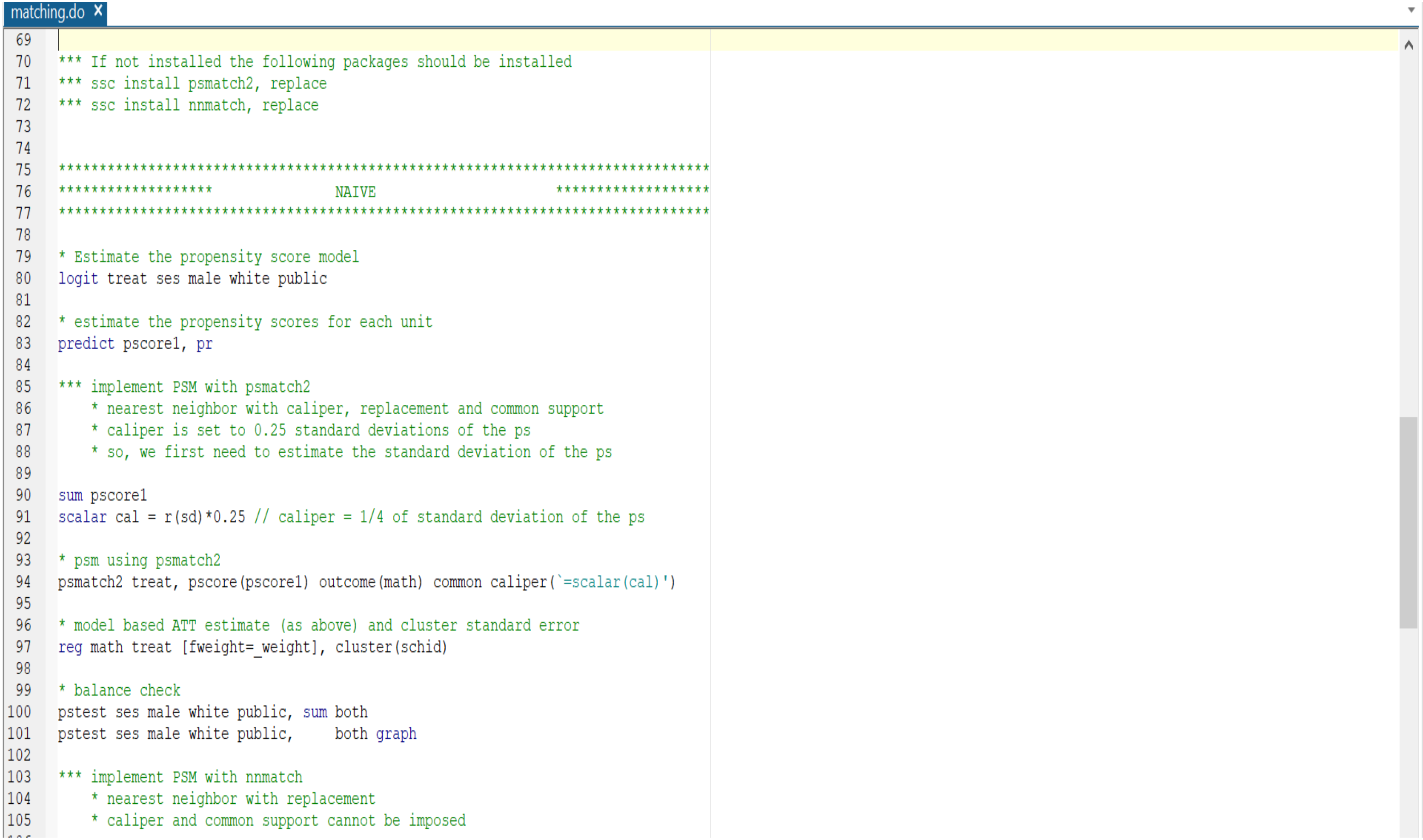
- **teffects** (built-in)
  - PSM (some of the features of psmatch2). It does not offer balance checks
  - covariate matching (nnmatch)
  - it calculates **standard errors** that take into account that propensity scores are estimated. Theoretical results for clustered data are not yet available
- **kmatch** (Jann, 2017)
  - PSM and covariate matching (nn, kernel, ridge)
  - several options for optimal bandwidth selection; exact matching; bias adjustment
  - tools for common support and balance diagnostics
- **cem** (Iacus, King and Porro 2008)
  - coarsened exact matching
- **There is no command designed specifically for clustered data**

PSM in Stata with clustered data

Approaches	**PS model**	**Matching**
Naïve (NV)	Single-level logit (**logit**)	Pooled (**psmatch2**; nnmatch)
Within (W)	Single-level logit (**logit**)	Within-cluster (cycle on **psmatch2**; nnmatch with exact option)
Preferential (PW)	Single-level logit (**logit**)	"**Preferential**" within-cluster (ad hoc procedure based on **psmatch2** or nnmatch)
Random-effects (RE)	Random-effect logit (e.g., **xtmelogit**)	Pooled **(psmatch2**; nnmatch**)**
Fixed-effects (RE)	Fixed-effect logit (e.g., **logit + clusters' dummies**)	Pooled **(psmatch2**; nnmatch**)**

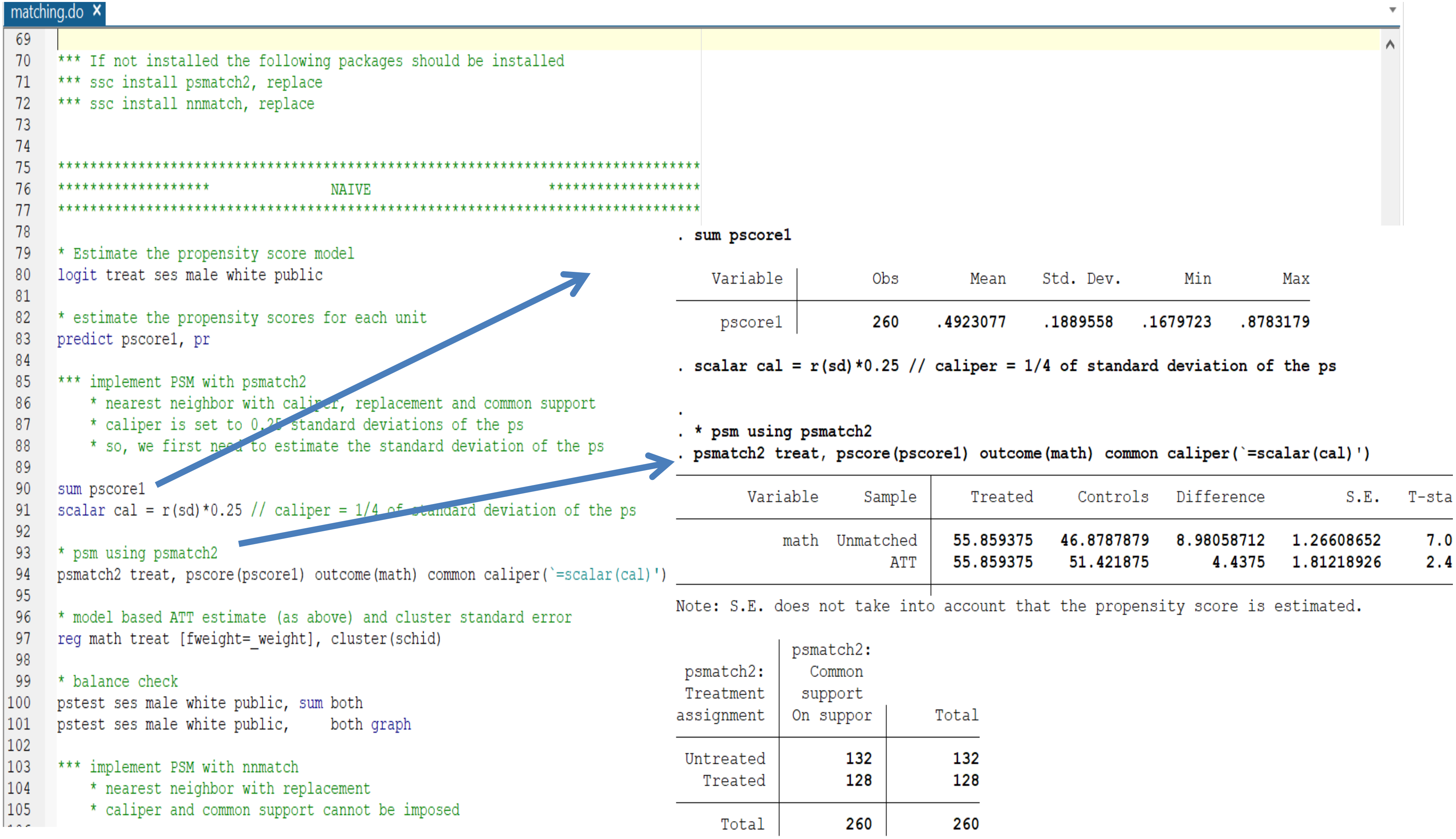
Outcome analysis should account for clustering (robust se)

Case study: naive PSM



```

*** If not installed the following packages should be installed
*** ssc install psmatch2, replace
*** ssc install nnmatch, replace


********************************************************************************
*******************               NAIVE                      *******************
********************************************************************************

* Estimate the propensity score model
logit treat ses male white public

* estimate the propensity scores for each unit
predict pscore1, pr

*** implement PSM with psmatch2
    * nearest neighbor with caliper, replacement and common support
    * caliper is set to 0.25 standard deviations of the ps
    * so, we first need to estimate the standard deviation of the ps

sum pscore1
scalar cal = r(sd)*0.25 // caliper = 1/4 of standard deviation of the ps

* psm using psmatch2
psmatch2 treat, pscore(pscore1) outcome(math) common caliper(`=scalar(cal)')

* model based ATT estimate (as above) and cluster standard error
reg math treat [fweight=_weight], cluster(schid)

* balance check
pstest ses male white public, sum both
pstest ses male white public,     both graph

*** implement PSM with nnmatch
    * nearest neighbor with replacement
    * caliper and common support cannot be imposed
```

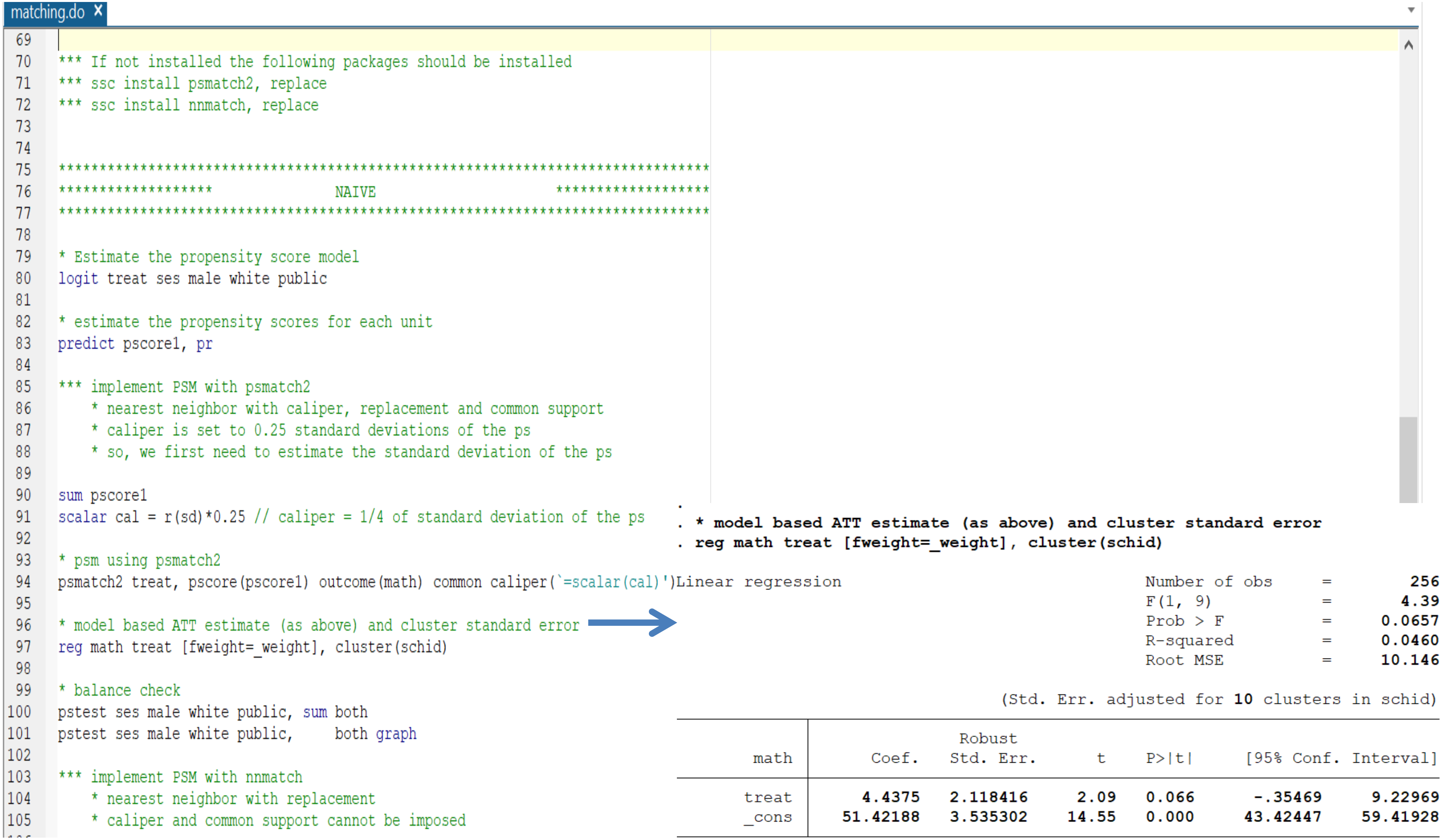
Case study: naive PSM



```
*** If not installed the following packages should be installed
*** ssc install psmatch2, replace
*** ssc install nnmatch, replace


********************************************************************************
*******************                  NAIVE                   *******************
********************************************************************************

* Estimate the propensity score model
logit treat ses male white public

* estimate the propensity scores for each unit
predict pscore1, pr

*** implement PSM with psmatch2
    * nearest neighbor with caliper, replacement and common support
    * caliper is set to 0.25 standard deviations of the ps
    * so, we first need to estimate the standard deviation of the ps

sum pscore1
scalar cal = r(sd)*0.25 // caliper = 1/4 of standard deviation of the ps

* psm using psmatch2
psmatch2 treat, pscore(pscore1) outcome(math) common caliper(`=scalar(cal)')

* model based ATT estimate (as above) and cluster standard error
reg math treat [fweight=_weight], cluster(schid)

* balance check
pstest ses male white public, sum both
pstest ses male white public,     both graph

*** implement PSM with nnmatch
    * nearest neighbor with replacement
    * caliper and common support cannot be imposed
```

```
. sum pscore1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     pscore1 |        260    .4923077    .1889558   .1679723   .8783179

. scalar cal = r(sd)*0.25 // caliper = 1/4 of standard deviation of the ps

.
. * psm using psmatch2
. psmatch2 treat, pscore(pscore1) outcome(math) common caliper(`=scalar(cal)')
----------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-sta
----------------------------+-----------------------------------------------------------
            math  Unmatched |  55.859375   46.8787879   8.98058712   1.26608652     7.0
                        ATT |  55.859375    51.421875       4.4375   1.81218926     2.4
----------------------------+-----------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |       132 |       132
   Treated |       128 |       128
-----------+-----------+----------
     Total |       260 |       260
```

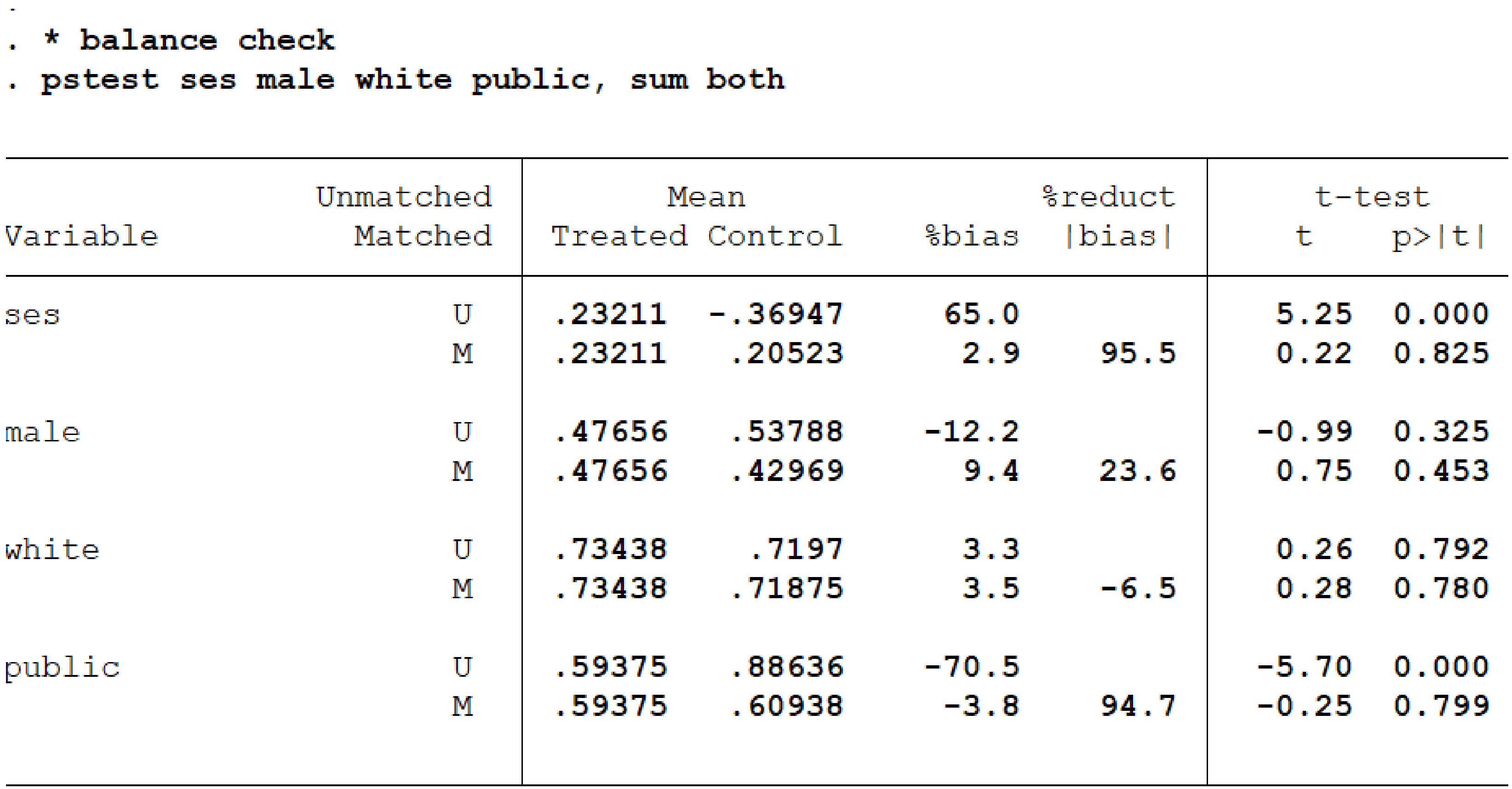
Case study: naive PSM



```
*** If not installed the following packages should be installed
*** ssc install psmatch2, replace
*** ssc install nnmatch, replace


*******************************************************************************
*******************                    NAIVE                    ******************
*******************************************************************************

* Estimate the propensity score model
logit treat ses male white public

* estimate the propensity scores for each unit
predict pscore1, pr

*** implement PSM with psmatch2
    * nearest neighbor with caliper, replacement and common support
    * caliper is set to 0.25 standard deviations of the ps
    * so, we first need to estimate the standard deviation of the ps

sum pscore1
scalar cal = r(sd)*0.25 // caliper = 1/4 of standard deviation of the ps

* psm using psmatch2
psmatch2 treat, pscore(pscore1) outcome(math) common caliper(`=scalar(cal)')

* model based ATT estimate (as above) and cluster standard error
reg math treat [fweight=_weight], cluster(schid)

* balance check
pstest ses male white public, sum both
pstest ses male white public,     both graph

*** implement PSM with nnmatch
    * nearest neighbor with replacement
    * caliper and common support cannot be imposed
```

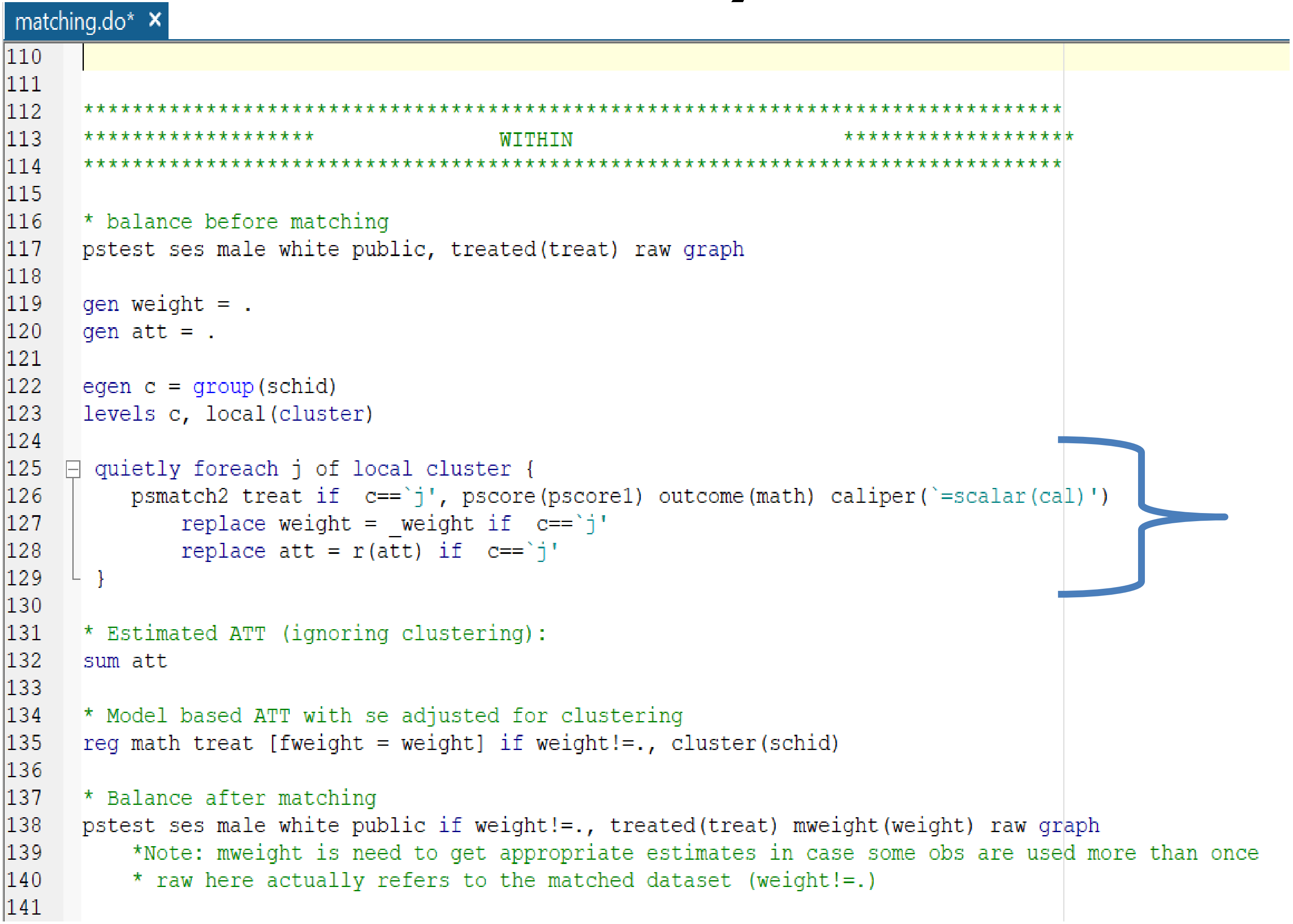
```
.
. * model based ATT estimate (as above) and cluster standard error
. reg math treat [fweight=_weight], cluster(schid)

Linear regression                               Number of obs     =        256
                                                F(1, 9)           =       4.39
                                                Prob > F          =     0.0657
                                                R-squared         =     0.0460
                                                Root MSE          =     10.146

                                   (Std. Err. adjusted for 10 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |     4.4375   2.118416     2.09   0.066     -.35469     9.22969
       _cons |   51.42188   3.535302    14.55   0.000    43.42447    59.41928
------------------------------------------------------------------------------
```

Case study: naive PSM



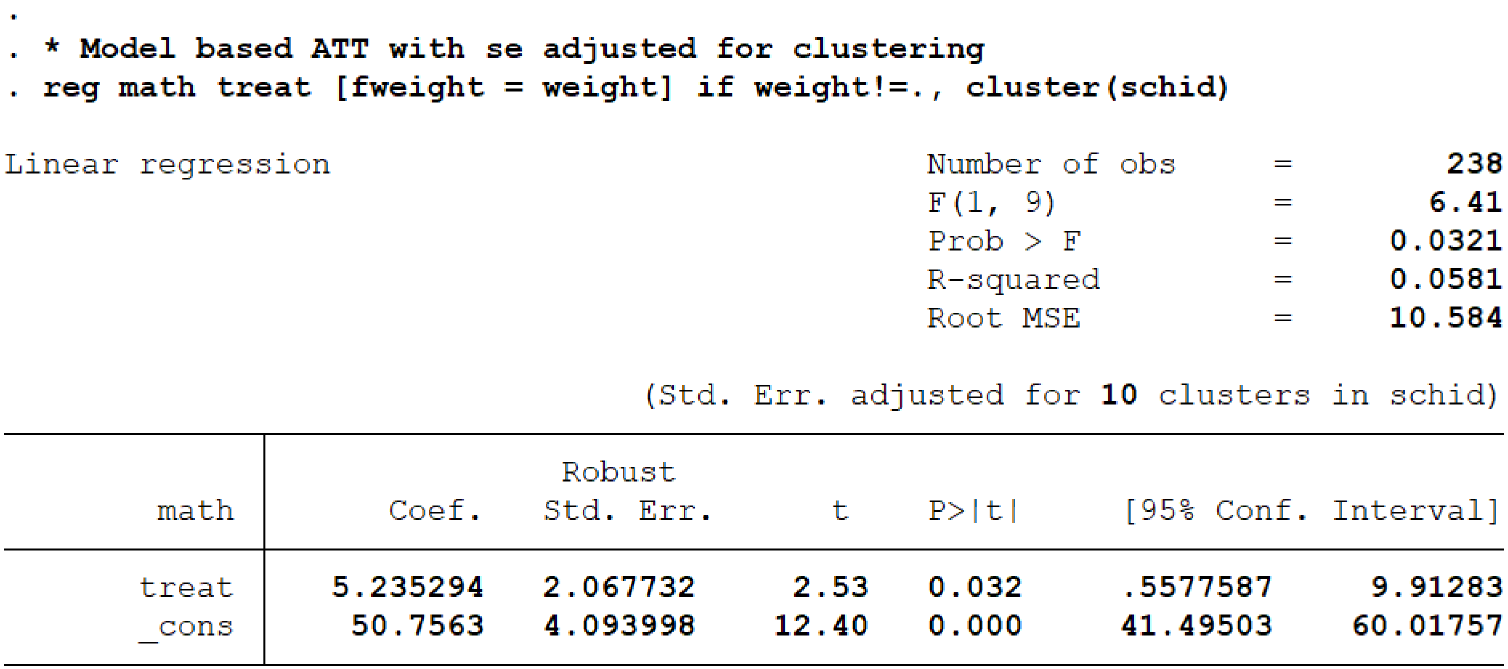
```
.
. * balance check
. pstest ses male white public, sum both

----------------------------------------------------------------------------
                        |       Mean               %reduct |     t-test
Variable        Matched | Treated Control    %bias  |bias| |    t    p>|t|
------------------------+----------------------------------+----------------
ses                   U | .23211  -.36947     65.0         |   5.25  0.000
                      M | .23211   .20523      2.9    95.5 |   0.22  0.825
                        |                                  |
male                  U | .47656   .53788    -12.2         |  -0.99  0.325
                      M | .47656   .42969      9.4    23.6 |   0.75  0.453
                        |                                  |
white                 U | .73438    .7197      3.3         |   0.26  0.792
                      M | .73438   .71875      3.5    -6.5 |   0.28  0.780
                        |                                  |
public                U | .59375   .88636    -70.5         |  -5.70  0.000
                      M | .59375   .60938     -3.8    94.7 |  -0.25  0.799
                        |                                  |
----------------------------------------------------------------------------
```

Case study: within PSM



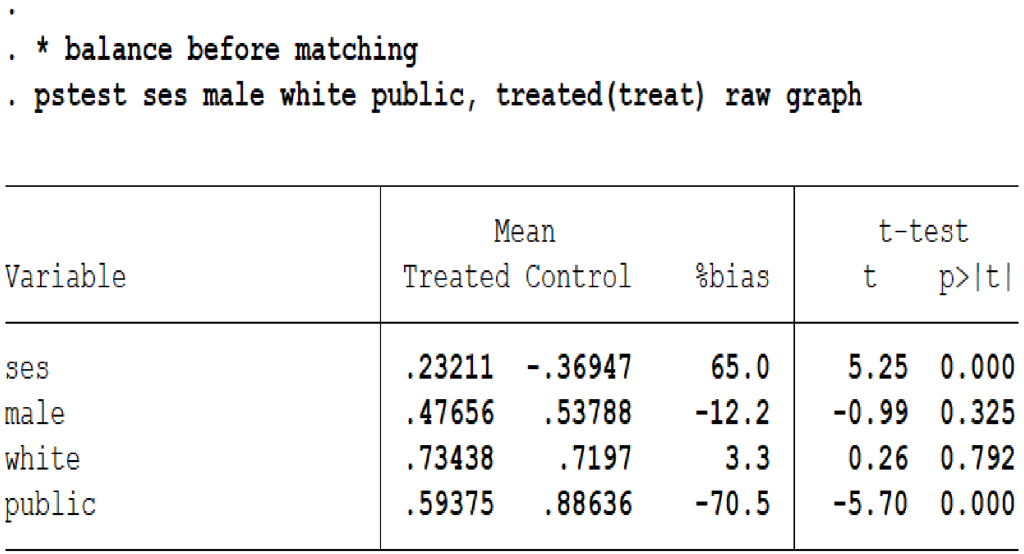
```
********************************************************************************
*******************               WITHIN                      *******************
********************************************************************************

* balance before matching
pstest ses male white public, treated(treat) raw graph

gen weight = .
gen att = .

egen c = group(schid)
levels c, local(cluster)

 quietly foreach j of local cluster {
    psmatch2 treat if  c==`j', pscore(pscore1) outcome(math) caliper(`=scalar(cal)')
        replace weight = _weight if  c==`j'
        replace att = r(att) if  c==`j'
 }

* Estimated ATT (ignoring clustering):
sum att

* Model based ATT with se adjusted for clustering
reg math treat [fweight = weight] if weight!=., cluster(schid)

* Balance after matching
pstest ses male white public if weight!=., treated(treat) mweight(weight) raw graph
    *Note: mweight is need to get appropriate estimates in case some obs are used more than once
    * raw here actually refers to the matched dataset (weight!=.)

```

Case study: within PSM

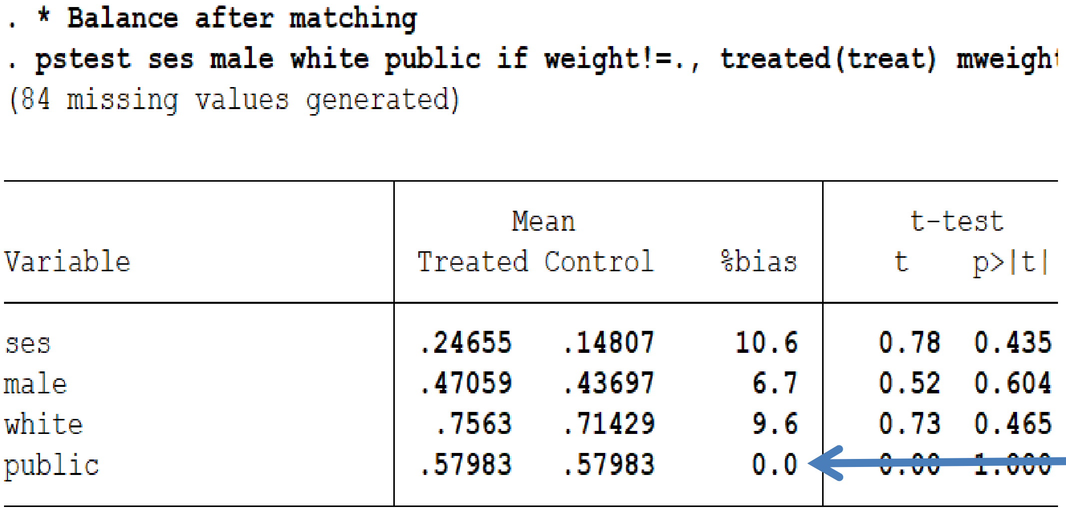


```
.
. * Model based ATT with se adjusted for clustering
. reg math treat [fweight = weight] if weight!=., cluster(schid)

Linear regression                               Number of obs     =        238
                                                F(1, 9)           =       6.41
                                                Prob > F          =     0.0321
                                                R-squared         =     0.0581
                                                Root MSE          =     10.584

                                  (Std. Err. adjusted for 10 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   5.235294   2.067732     2.53   0.032     .5577587    9.91283
       _cons |    50.7563   4.093998    12.40   0.000     41.49503   60.01757
------------------------------------------------------------------------------
```

Case study: within PSM



```
.
. * balance before matching
. pstest ses male white public, treated(treat) raw graph

-----------------------------------------------------------------
                     |       Mean                |     t-test
Variable             | Treated Control    %bias  |    t    p>|t|
---------------------+---------------------------+---------------
ses                  | .23211  -.36947     65.0  |   5.25  0.000
male                 | .47656   .53788    -12.2  |  -0.99  0.325
white                | .73438    .7197      3.3  |   0.26  0.792
public               | .59375   .88636    -70.5  |  -5.70  0.000
-----------------------------------------------------------------
```

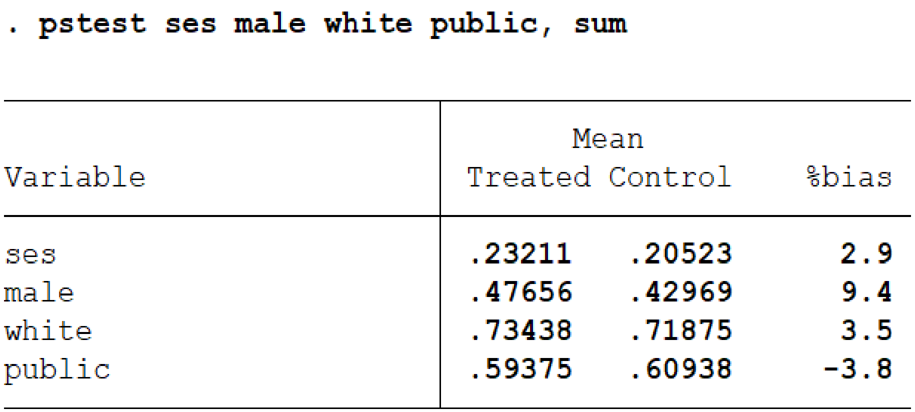


```
. * Balance after matching
. pstest ses male white public if weight!=., treated(treat) mweigh
(84 missing values generated)

------------------------------------------------------------------
                     |       Mean                 |     t-test
Variable             | Treated Control     %bias  |    t    p>|t|
---------------------+----------------------------+---------------
ses                  | .24655   .14807      10.6  |   0.78  0.435
male                 | .47059   .43697       6.7  |   0.52  0.604
white                |  .7563   .71429       9.6  |   0.73  0.465
public               | .57983   .57983       0.0  |   0.00  1.000
------------------------------------------------------------------
```

key advantage of within matching! (%bias = 0)

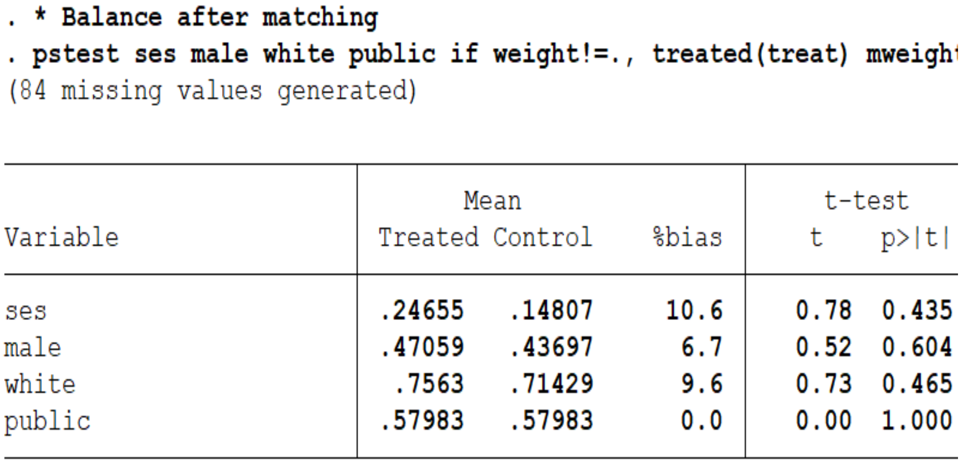
Comparing balance: naive vs within PSM

naive



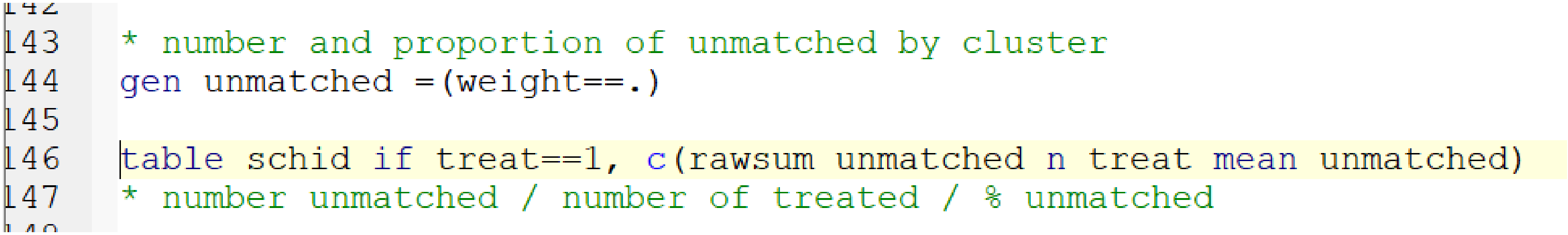
```
. pstest ses male white public, sum

----------------------------------------------------
                          |      Mean
Variable                  | Treated Control   %bias
--------------------------+-------------------------
ses                       |  .23211   .20523    2.9
male                      |  .47656   .42969    9.4
white                     |  .73438   .71875    3.5
public                    |  .59375   .60938   -3.8
----------------------------------------------------
```

within



```
. * Balance after matching
. pstest ses male white public if weight!=., treated(treat) mweigh
(84 missing values generated)

------------------------------------------------------------------
                         |      Mean               |    t-test
Variable                 | Treated Control   %bias |    t    p>|t|
-------------------------+-------------------------+--------------
ses                      |  .24655   .14807   10.6 |  0.78  0.435
male                     |  .47059   .43697    6.7 |  0.52  0.604
white                    |   .7563   .71429    9.6 |  0.73  0.465
public                   |  .57983   .57983    0.0 |  0.00  1.000
------------------------------------------------------------------
```

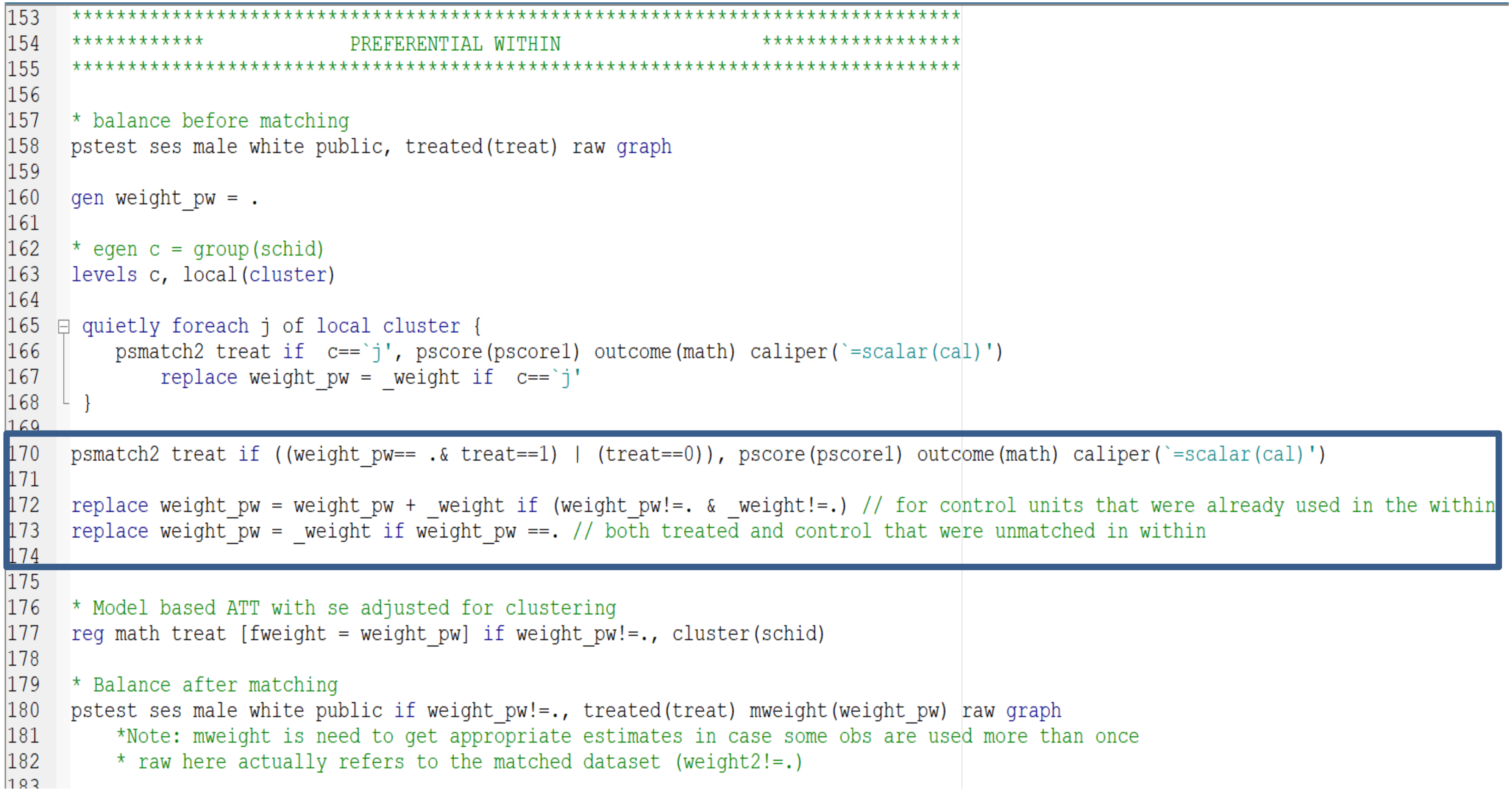
Case study: within PSM



```
* number and proportion of unmatched by cluster
gen unmatched =(weight==.)

table schid if treat==1, c(rawsum unmatched n treat mean unmatched)
* number unmatched / number of treated / % unmatched
```

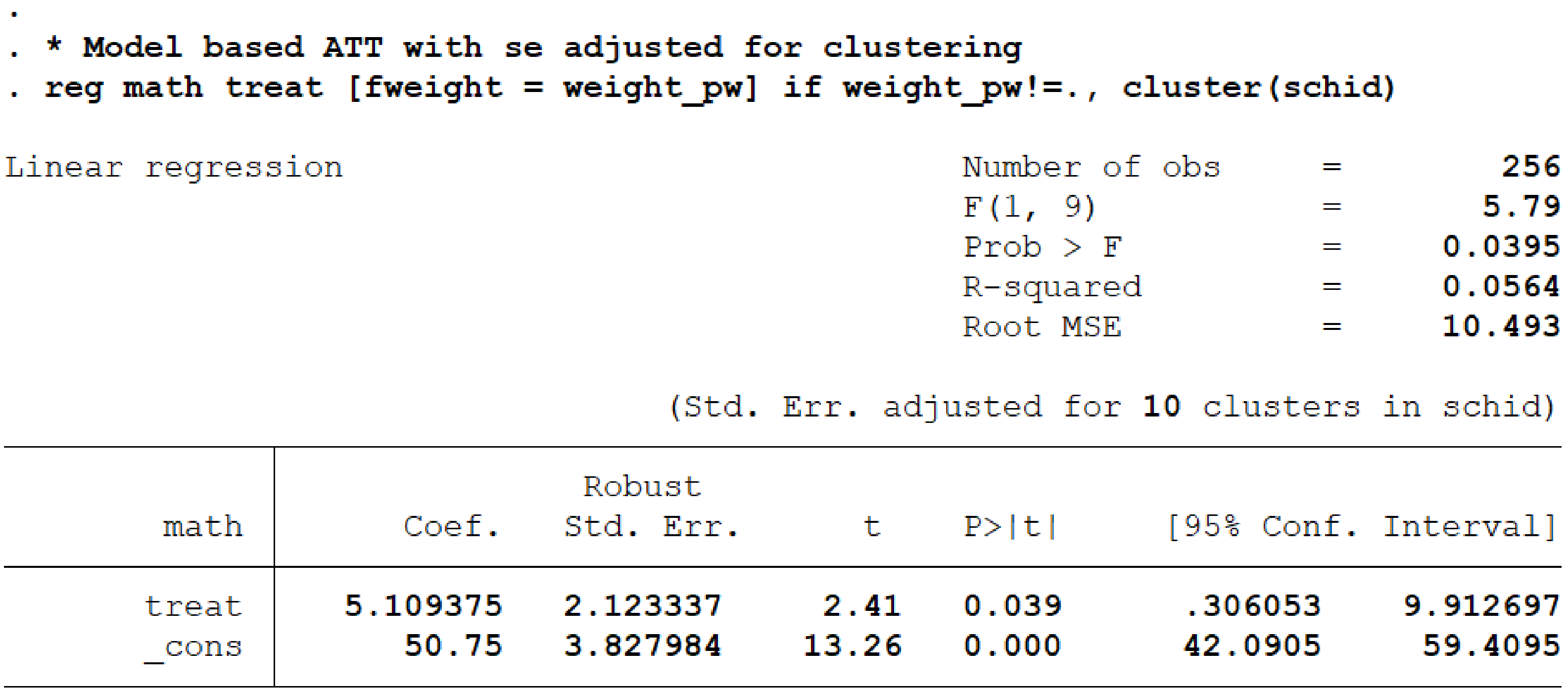
```
. table schid if treat==1, c(rawsum unmatched n treat mean unmatched)

----------------------------------------------------------------
School ID | rawsum(unmatc~d)          N(treat)    mean(unmatc~d)
----------+-----------------------------------------------------
     7472 |                0                 6                 0
     7829 |                2                14          .1428571
     7930 |                0                12                 0
    24725 |                0                 7                 0
    25456 |                0                 5                 0
    25642 |                0                 4                 0
    62821 |                2                52          .0384615
    68448 |                5                13          .3846154
    68493 |                0                 6                 0
    72292 |                0                 9                 0
----------------------------------------------------------------
```

Case study: preferential within PSM



```
********************************************************************************
************             PREFERENTIAL WITHIN                  ******************
********************************************************************************

* balance before matching
pstest ses male white public, treated(treat) raw graph

gen weight_pw = .

* egen c = group(schid)
levels c, local(cluster)

quietly foreach j of local cluster {
    psmatch2 treat if  c==`j', pscore(pscore1) outcome(math) caliper(`=scalar(cal)')
        replace weight_pw = _weight if  c==`j'
}

psmatch2 treat if ((weight_pw== .& treat==1) | (treat==0)), pscore(pscore1) outcome(math) caliper(`=scalar(cal)')

replace weight_pw = weight_pw + _weight if (weight_pw!=. & _weight!=.) // for control units that were already used in the within
replace weight_pw = _weight if weight_pw ==. // both treated and control that were unmatched in within


* Model based ATT with se adjusted for clustering
reg math treat [fweight = weight_pw] if weight_pw!=., cluster(schid)

* Balance after matching
pstest ses male white public if weight_pw!=., treated(treat) mweight(weight_pw) raw graph
    *Note: mweight is need to get appropriate estimates in case some obs are used more than once
    * raw here actually refers to the matched dataset (weight2!=.)
```

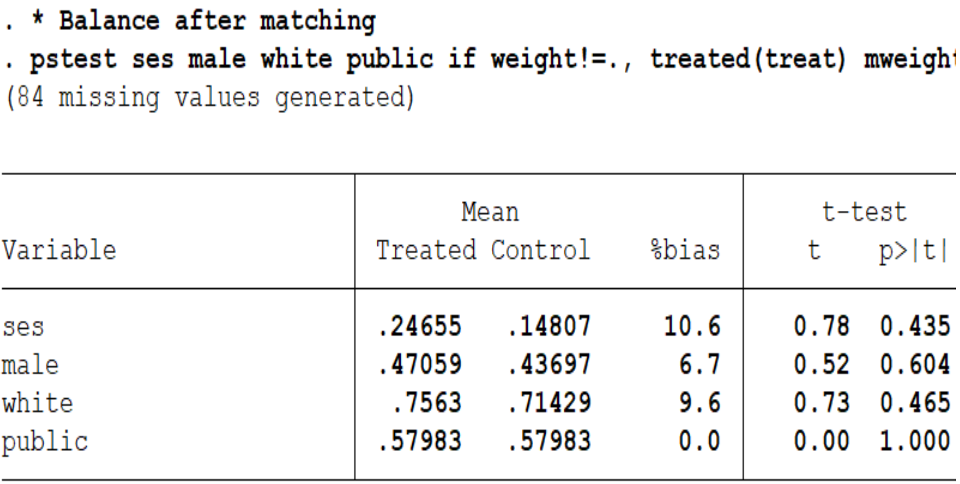
Case study: preferential within PSM



```
.
. * Model based ATT with se adjusted for clustering
. reg math treat [fweight = weight_pw] if weight_pw!=., cluster(schid)

Linear regression                               Number of obs     =        256
                                                F(1, 9)           =       5.79
                                                Prob > F          =     0.0395
                                                R-squared         =     0.0564
                                                Root MSE          =     10.493

                                   (Std. Err. adjusted for 10 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   5.109375   2.123337     2.41   0.039      .306053    9.912697
       _cons |      50.75   3.827984    13.26   0.000      42.0905     59.4095
------------------------------------------------------------------------------
```

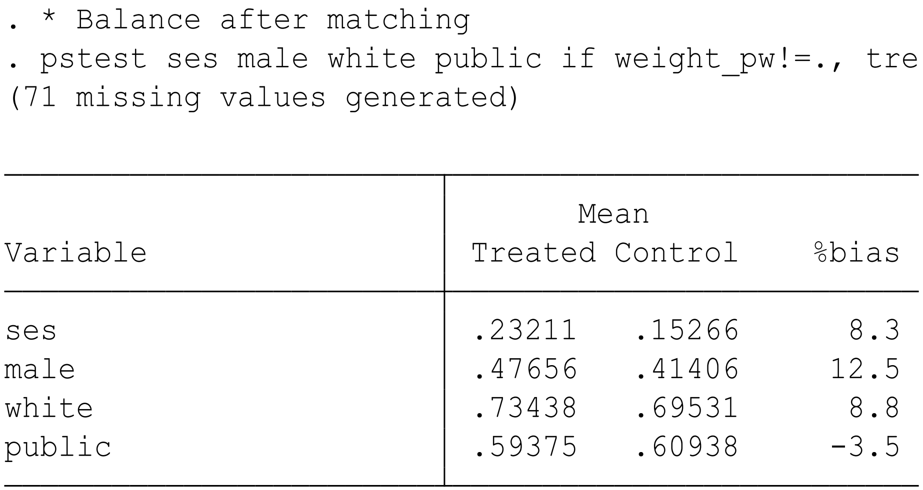
Comparing balance: within vs preferential within PSM

within



```
. * Balance after matching
. pstest ses male white public if weight!=., treated(treat) mweigh
(84 missing values generated)

----------------------------------------------------------------
                  |       Mean               |     t-test
Variable          | Treated Control    %bias |    t    p>|t|
------------------+--------------------------+------------------
ses               | .24655   .14807     10.6 |  0.78  0.435
male              | .47059   .43697      6.7 |  0.52  0.604
white             |  .7563   .71429      9.6 |  0.73  0.465
public            | .57983   .57983      0.0 |  0.00  1.000
----------------------------------------------------------------
```

preferential



```
. * Balance after matching
. pstest ses male white public if weight_pw!=., tre
(71 missing values generated)

--------------------------------------------------
                        |       Mean
Variable                | Treated   Control  %bias
------------------------+-------------------------
ses                     | .23211    .15266     8.3
male                    | .47656    .41406    12.5
white                   | .73438    .69531     8.8
public                  | .59375    .60938    -3.5
--------------------------------------------------
```

Case study: preferential within PSM

```
. table schid if treat==1, c(rawsum unmatched_pw n treat mean unmatched_pw)

------------------------------------------------------------------
School ID | rawsum(unmatc~w)          N(treat)    mean(unmatc~w)
----------+-------------------------------------------------------
     7472 |                0                 6                 0
     7829 |                0                14                 0
     7930 |                0                12                 0
    24725 |                0                 7                 0
    25456 |                0                 5                 0
    25642 |                0                 4                 0
    62821 |                0                52                 0
    68448 |                0                13                 0
    68493 |                0                 6                 0
    72292 |                0                 9                 0
------------------------------------------------------------------
```

Concluding remarks

- In the presence of clustered data several approaches can be followed to implement PSM
- “Within” matching works well with big clusters
- “Preferential” within matching is an attractive alternative when all or some clusters are small
- Available routines in Stata can be adapted to clustered data
- Future developments:
  - Standard errors accounting for estimation of PS (as in teffects psmatch2)
  - Within-cluster balance

(Some) References

- Abadie, A., D. M. Drukker, J. L. Herr, and G. W. Imbens. 2004. Implementing matching estimators for average treatment effects in Stata. Stata Journal 4: 290–311.
- Arpino B. and Cannas M. (2016) Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score, *Statistics in Medicine*, 35(12), 2074–2091.
- Arpino B. and Mealli F. (2011) The specification of the propensity score in multilevel studies*, Computational Statistics and Data Analysis*, 55, 1770-1780.
- Imbens, G. W. (2014). Matching methods in practice: Three examples (NBER Working Paper No. 19959). Cambridge, MA: National Bureau of Economic Research
- Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.
- Li F, Zaslavsky AM, Landrum MB. (2013) Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19): 3373–3387.
- Thoemmes FJ, West SG. (2011) The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46(3): 514–543.

Bruno Arpino
Pompeu Fabra University
bruno.arpino@upf.edu
https://sites.google.com/site/brunoarpino