Propensity Score Matching (PSM)

DENG KUANGKUANG

Ideal empirical setting in economics

- Economics borrow methodology from physics
- Ideally, we could conduct experiment in a fully controlled setting
    - We have a groups of observations
    - Allocate the obs into a treatment group and a control group **randomly**
    - Give the treatment only to the treatment group
    - Make sure other things are the same for both groups
    - Observe the different responses of the treatment and the control
- However, we have to deal with real data, not experiments.
- Causal inference is difficult with real data

About methodology

Recommended reading:

- Karl popper, *Conjectures and refutations: the growth of scientific knowledge*

PSM

- Propensity score matching (倾向评分匹配): firstly proposed by Rosenbaum and Rubin (1983)

- Scenario:
  - **Non-experimental settings**: non-randomized observational studies like in most cases of economic research
  - But you want to **infer causal effects** from a treatment group and a control group

  - **Problem**: units in the comparison group are not perfectly comparable to the treatment units
  - **Objective**: to solve the <u>sample selection</u> bias
  - **Approach**: (1) select a subsample of the control group that is comparable to the treatment group; (2) estimate treatment effects with the treatment group and the selected control group
  - HOW?

Motivation anecdote

- Two heart surgeons walk into a room
  - Surgeon A: Man, I just finished my 100th heart surgery!
  - Surgeon B: Oh yeah, I finished my 100th heart surgery last week. I bet I'm a better surgeon than you. How many of your patients died within 3 months of surgery? Only 10 of my patients died."
  - Surgeon A smugly responds: Only 5 of mine died, so I must be the better surgeon.
  - Surgeon B: My patients were probably older and had a higher risk than your patients.

Estimands

The causal estimands of interest are usually average treatment effects on the whole population or on subpopulations.

- Average treatment effect: ATE = E[Y (1) - Y (0)] is useful to evaluate what is the expected effect on the outcome if individuals in the population were randomly assigned to treatment.
  - ATE might not be of relevance to policy makers because it includes the effect on persons for whom the program was never intended.
- the average treatment effect on the treated (ATT, 被干预样本的平均干预效应)

  $$ATT = E[Y(1) - Y(0) \mid W = 1]$$

  is useful to explicitly evaluate the effects on those for whom the treatment is actually intended.

From here on, we consider ATT, the parameter of interest in most evaluation studies.

ATT = $E[Y(1) - Y(0) \mid W = 1]$

- Note that $E[Y(0) \mid W = 1]$, i.e. the counterfactual mean for those being treated is not observed

=> choose a proper substitute for it in order to estimate ATT.

- Should we use the mean outcome of untreated individuals $E[Y(0) \mid W = 0]$?
- in observational studies, this is not a good idea. Because covariates which determine the treatment decision may also determine the outcome variable of interest, e.g. patients’ age, probability of doing heart surgery, probability of dying within 3 months.

⇩

The outcomes of individuals from the treatment and comparison groups would differ ***even in the absence of treatment*** leading to the so-called selection bias

Sources of selection bias

- non overlapping supports of X in the treated and comparison group(i.e., the presence of units in one group that cannot find suitable comparison in the other);
  - Oldest patient for Surgeon A is 70; that for Surgeon B is 50.

- unbalance in observed confounders between the groups of treated and control units (selection on observables);
  - Surgeon A treat more patients in their 70s; Surgeon B’s patients are younger on average

- unbalance in unobserved confounders between the groups of treated and control units (selection on unobservables);
  - E.g. Differential health habits of patients

Observational studies

How can we reduce the bias in estimating treatment effects?

- With an observational data set, we try to structure it so that we can conceptualize the data as having arisen from an underlying regular assignment mechanism.

- We need to adjust any difference in average outcomes for differences in pre-treatment characteristics (not being affected by the treatment)
  - Model-based imputation methods (e.g. regression models)
  - Matching methods
  - Methods based on propensity score
  - Stratification
  - Weighting
  - Mixed methods

Matching approach

- Matching techniques have origins in experimental work from the first half of the twentieth century (see e.g. Rubin (1974) or Lechner (1998)) and were advanced and developed in a set of papers by Rosenbaum and Rubin (1983a, 1984, 1985a, 1985b).

- To ensure that the matching estimators identify and consistently estimate the treatment effects of interest, we assume:

  - **unconfoundedness**: assignment to treatment is independent of the outcomes, conditional on the covariates

$$(Y(0); Y(1) \perp W | X)$$

- Given X, the assignment of treatment (W) is uncorrelated with the outcome.
- If a factor affects both the assignment of treatment and the outcome, we call it a confounding variable.

  - **overlap** or common support condition: the probability of assignment is bounded away from zero and one

$$0 < Pr(W = 1 \mid X) < 1$$

  - **Strong ignorability**: when both assumptions are satisfied, the treatment can be called strongly ignorable

Matching approach (cont'd)

The underlying identifying assumption is **unconfoundedness** (selection on observables or conditional independence).

- intuition: If the decision to take the treatment is purely random for individuals with similar values of the pre-treatment variables, then we could use the average outcome of some similar individuals who were not exposed to the treatment.
  - for each *i* , matching estimators impute the missing outcome by finding other individuals in the data whose covariates are similar but who were exposed to the other treatment.
  - in this way, differences in outcomes of this well selected and thus adequate control group of participants can be attributed to the treatment.

Overlap (cont'd)

$$0 < Pr(W = 1 \mid X) < 1$$

- The assignment mechanism can be interpreted as if, within subpopulations of units with the same value for the covariate, completely randomized experiment was carried out.
  - So, in the case where the oldest patient for Surgeon A is 70; that for Surgeon B is 50:
  - Patients over 50 are not randomly distributed to treatment and control groups

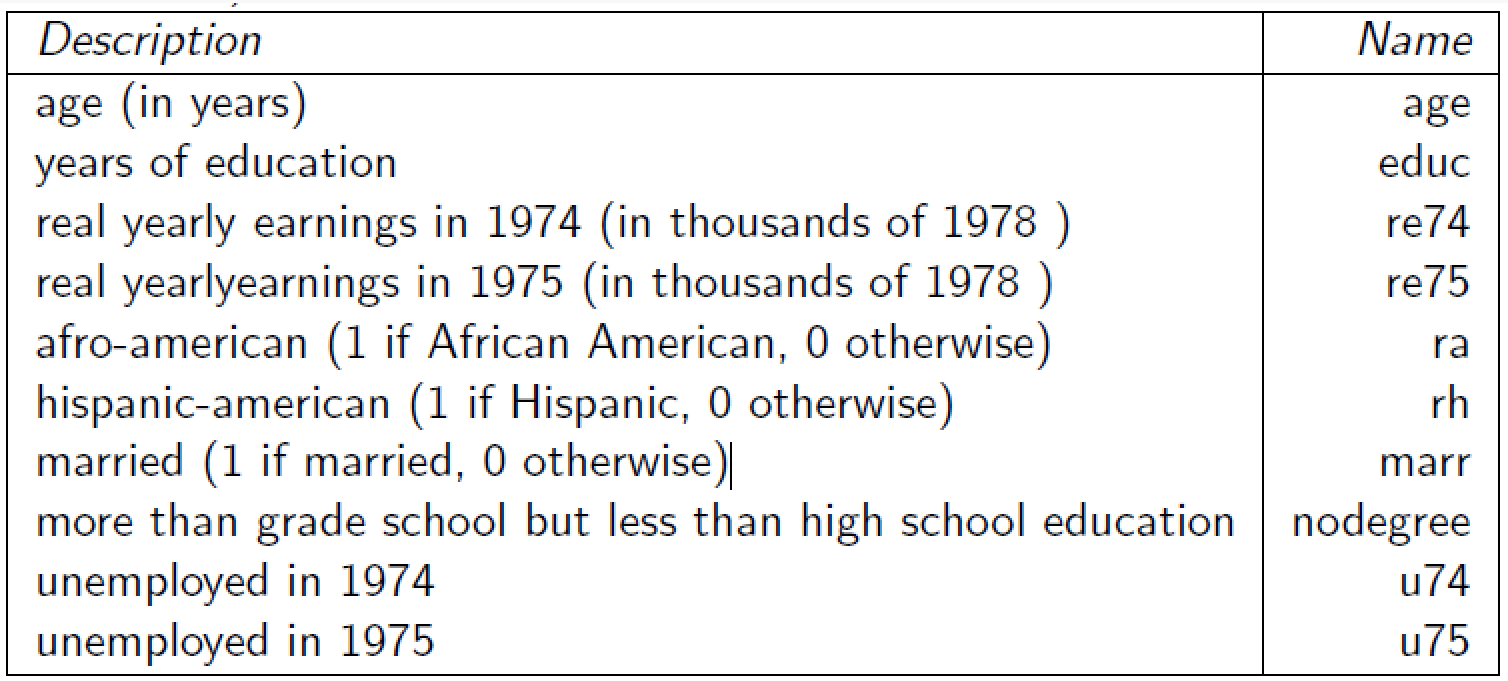
Effect of participation in a job training program on individuals earnings

Data used on Lalonde (1986)

- We are interested in a possible effect of participation in a job training program on individuals' earnings in 1978
- This dataset has been used by many authors (Abadie et al. 2004, Becker and Ichino, 2002, Dehejia and Wahba, 1999)
- We use a subset of the data constructed by Dehejia and Wahba (1999, see their paper for details), which can be downloaded here:
  - https://economics.mit.edu/faculty/angrist/data1/mhe/dehejia
  - http://users.nber.org/~rdehejia/nswdata.html
- Variables:
  - Treatment t: participation in the job training program
  - Outcome re78: 1978 earnings of the individual in the sample in terms of 1978 dollars
  - Observable covariates

Example: covariates

The data set includes information on pre-treatment (background; confounder) variables



Description	*Name*
age (in years)	age
years of education	educ
real yearly earnings in 1974 (in thousands of 1978)	re74
real yearlyearnings in 1975 (in thousands of 1978)	re75
afro-american (1 if African American, 0 otherwise)	ra
hispanic-american (1 if Hispanic, 0 otherwise)	rh
married (1 if married, 0 otherwise)	marr
more than grade school but less than high school education	nodegree
unemployed in 1974	u74
unemployed in 1975	u75

Regression-based estimation

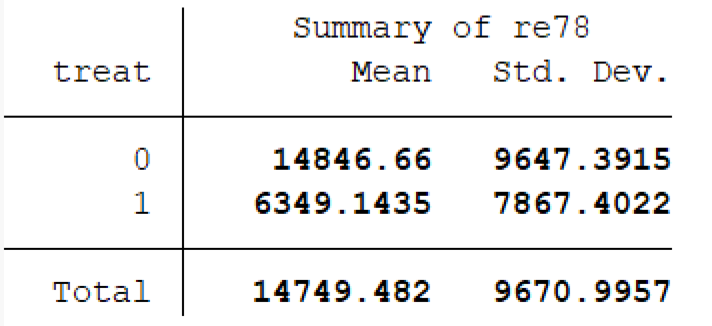
- We need to adjust any difference in average outcomes for differences in pre-treatment characteristics (not being affected by the treatment)
  - We can adjust via specification of a conditional model for the potential outcome => regression models
  - In a standard regression approach, unconfoundedness is implicitly assumed together with other functional or distributional assumptions

$$\widehat{Y_i^{obs}} = \alpha + \tau W_i + \beta X_i + \varepsilon_i$$

With the usual exogeneity assumption that $\varepsilon_i \perp W_i, X_i$

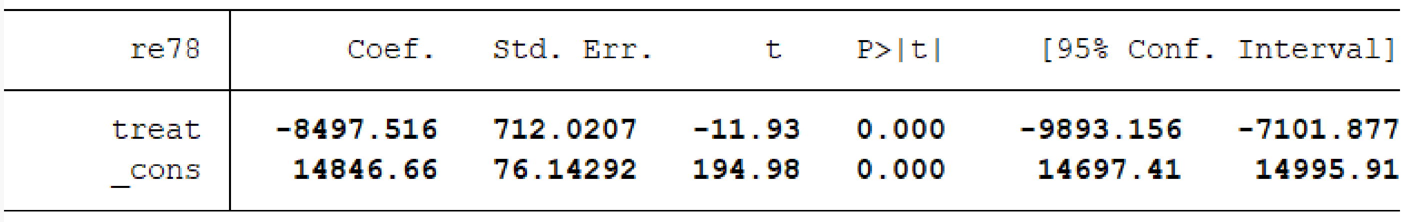
Regression-based estimation

tabulate treat, summarize(re78) means standard



```
            |   Summary of re78
      treat |        Mean   Std. Dev.
------------+------------------------
          0 |    14846.66   9647.3915
          1 |   6349.1435   7867.4022
------------+------------------------
      Total |   14749.482   9670.9957
```

reg re78 treat



```
------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -8497.516   712.0207   -11.93   0.000    -9893.156   -7101.877
       _cons |   14846.66   76.14292   194.98   0.000     14697.41    14995.91
------------------------------------------------------------------------------
```

- Should we conclude that the treatment is dangerous because the expected average earning for treated is lower than for control?
- Are the assumptions underlying the linear regression model plausible in this case?

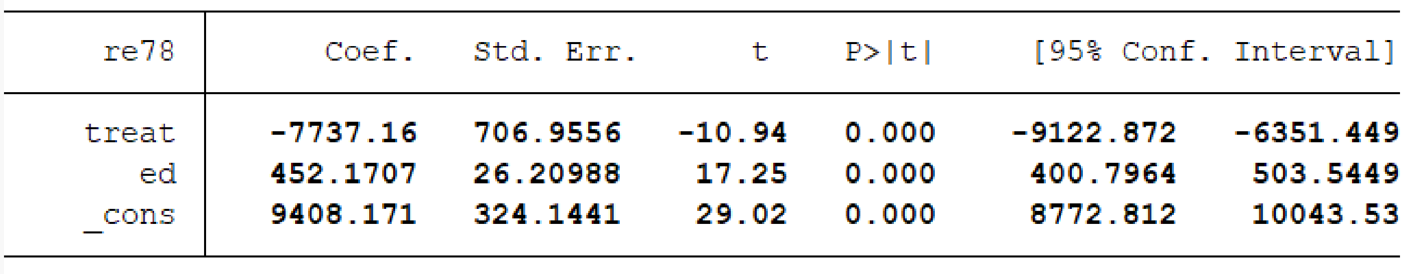
Multi-linear regression model

- Adjusting for confounding variables, we can estimate the conditional ATT

$$E[(Y(1) - Y(0)\ |X = x]$$

- Imagine that the only confounder is EDUCATION
  - Well educated people can earn more, but are less likely to participate in job training.

$$E(Y_i^{obs}|t, ed) = \alpha + \tau t_i + \beta ed_i$$



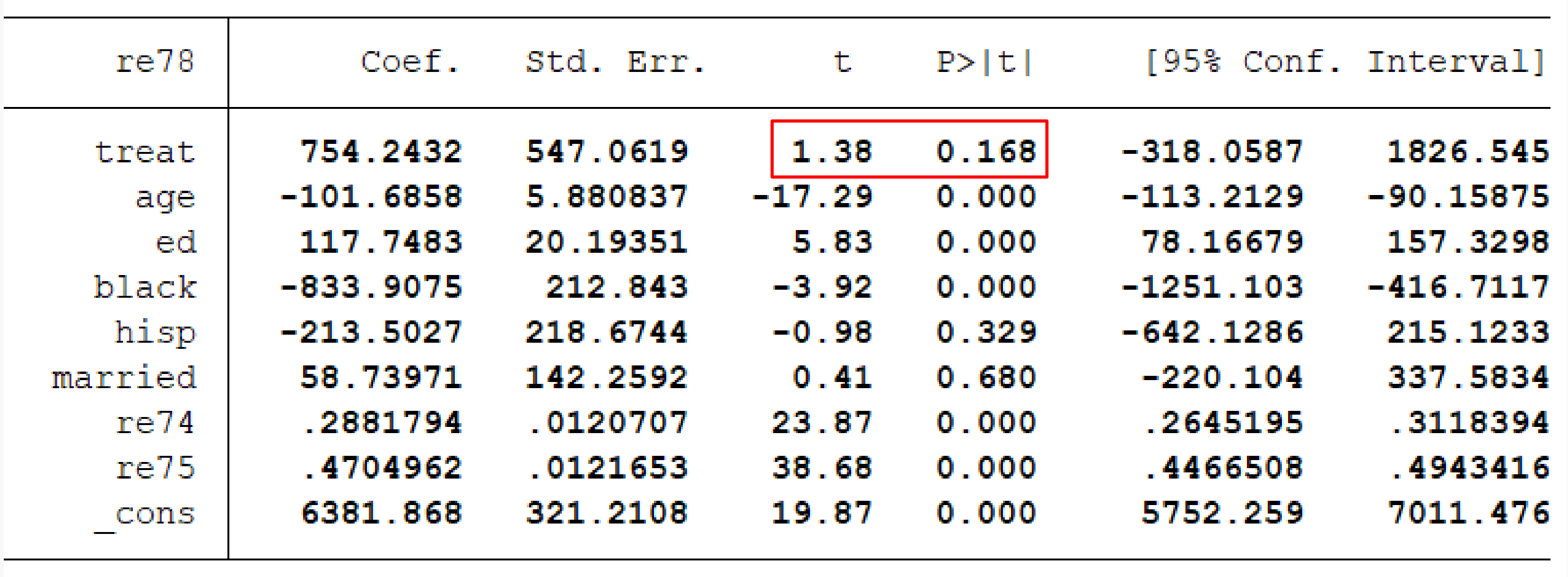
```
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   -7737.16   706.9556   -10.94   0.000    -9122.872   -6351.449
          ed |   452.1707   26.20988    17.25   0.000     400.7964    503.5449
       _cons |   9408.171   324.1441    29.02   0.000     8772.812    10043.53
------------------------------------------------------------------------------
```

Estimated ATT less negative!

Multi-linear regression model

Let’s include all pre-treatment variables available in the dataset

reg re78 treat age ed black hisp marr re74 re75



```
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   754.2432   547.0619     1.38   0.168    -318.0587    1826.545
         age |  -101.6858   5.880837   -17.29   0.000    -113.2129   -90.15875
          ed |   117.7483   20.19351     5.83   0.000     78.16679    157.3298
       black |  -833.9075    212.843    -3.92   0.000    -1251.103   -416.7117
        hisp |  -213.5027   218.6744    -0.98   0.329    -642.1286    215.1233
     married |   58.73971   142.2592     0.41   0.680     -220.104    337.5834
        re74 |   .2881794   .0120707    23.87   0.000     .2645195    .3118394
        re75 |   .4704962   .0121653    38.68   0.000     .4466508    .4943416
       _cons |   6381.868   321.2108    19.87   0.000     5752.259    7011.476
------------------------------------------------------------------------------
```

Estimated ATT positive, as expected, but statistically insignificant !

Regression: overlap problems

- To identify causal effects, unconfoundedness is not enough, to achieve ignorability, we need also overlap, i.e. $0 < p_i(x) < 1$ for each value $x \in X$
- Let us consider the following example:
  - We are interested in evaluating the effect of training on earnings
  - We can assume unconfoundedness of education
  - But for the data in hand, education has three values for the treated (1,2,3), but only two (1 and 3) for the control
  - This implies that the treated with X=2 cannot find good comparisons in the control (no overlap)
  - Regression analysis masks this fact and assumes that the estimated equation is good for everybody, even for those never observed!

Matching vs OLS

The main assumption underlying the matching approaches (unconfoundedness) is the same as OLS.

=> as OLS, the matching is as good as its X are!

Matching could be better than OLS:

- The additional **common support** condition focus on comparison of comparable subjects
- Matching is a **non-parametric** technique
  - It avoids potential misspecification
  - It allows for arbitrary heterogeneity in causal effects

- If OLS is correctly specified, it is more efficient than matching.

Balancing scores and propensity scores

Conditioning on all relevant covariates is limited in the case of a high dimensional vector X.

- Rosenbaum and Rubin (1983) suggest the use of so-called balancing scores b(X), i.e. functions of the relevant observed covariates X such that the conditional distribution of X given b(X) is independent of assignment into treatment.

$$X_i \perp W_i \mid b(X_i)$$

- Balancing scores are not unique
- One possible balancing score is the propensity score, i.e. the probability to be treated given observed characteristics X.

$$e(X) = \Pr(W = 1 \mid X = x) = E[W \mid X = x]$$

The role of propensity score

- Many of the procedures for estimating and assessing causal effects under unconfoundedness involve the propensity score.
  - If the balancing hypothesis

$$W \perp X \mid e(X)$$

  is satisfied, observations with the same propensity score must have the same distribution of observable (and unobservable) characteristics independently of treatment status

=> For a given propensity score, exposure to treatment is random and therefore treated and control units should be on average observationally identical

Pre-treatment variables choice

What variables should be included in the model for the PS?

In general, the choice of covariates to insert in the propensity score model should be based on

- Theory and previous empirical findings
- Formal (statistical) tests (e.g. Heckman et al. , 1998, Heckman and Smith, 1999 and Black and Smith, 2004)
- The model for the propensity scores does not need a behavioral interpretation.

Pre-treatment variables choice

- In the literature, some advice is available regarding the inclusion (or exclusion) of covariates in the propensity score model.
  - Only variables that influence **simultaneously** the treatment status and the outcome variable should be included (see e.g., Sianesi, 2004; Smith and Todd, 2005).
  - Only variables that are unaffected by treatment should be included in the model. To ensure this, variables should either be fixed over time or measured before participation.
  - If e(X) = 0 or e(X) = 1 for some values of X, then we cannot use matching conditional on those X values to estimate a treatment effect, because persons with such characteristics either always or never receive treatment. Hence, the common support condition (overlap) fails and matches cannot be performed.

Matching strategy and ATT estimation

The standard matching strategy is the following:

- pair each treated subject i with one or more comparable non-treated Subjects.
- associate to the outcome Y_i^{obs} a matched outcome $\widehat{Y}_i(0)$ given by the (weighted) outcomes of its neighbors in the comparison group

$$\widehat{Y}_i(0) = \sum_{j \in C(i)} w_{ij} Y_j^{obs}$$

$where$

- $C(i)$ is the set of neighbors with W=0 of the treated subject i
- w_{ij} is the weight of non-treated j (usually simply take average)

Propensity-score matching with STATA

The Stata command **psmatch2** (Leuven and Sianesi 2003) will perform PSM

- many matching methods are available: nearest neighbor (with or without within caliper, with or without replacement), k-nearest neighbors, radius, kernel, etc.

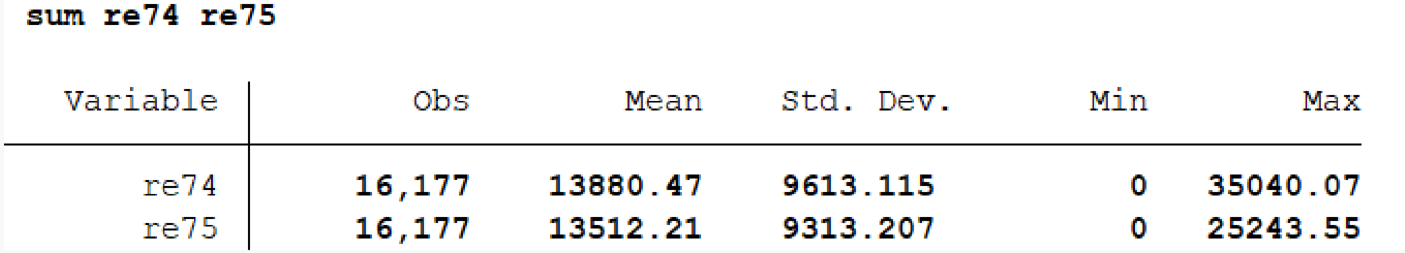
- it includes routines for common support graphing (**psgraph**) and covariate imbalance testing (**pstest**);

Nearest Neighbor Matching

- NN match treated and control units taking each treated unit and searching for the control unit with the closest propensity score; i.e., the Nearest Neighbor.
- Although it is not necessary, the method is usually *applied with replacement*, in the sense that a control unit can be a best match for more than one treated unit.
- Once each treated unit is matched with a control unit, the difference between the outcome of the treated units and the outcome of the matched control units is computed.
- The ATT of interest is then obtained by averaging these differences.
- All treated units find a match. However, it is obvious that some of these matches are fairly poor because for some treated units the nearest neighbor may have a very different propensity score, and, nevertheless, it would contribute to the estimation of the treatment effect independently of this difference.

Example: Real earning and unemployed subjects

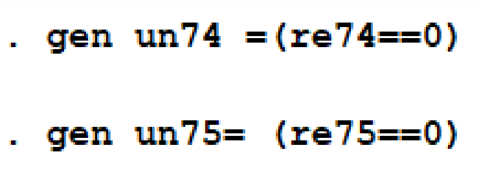
- In the distributions of real earnings before the treatment (re74 and re75) there are some 0: subjects with zero values were unemployed



```
sum re74 re75

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        re74 |     16,177    13880.47    9613.115          0   35040.07
        re75 |     16,177    13512.21    9313.207          0   25243.55
```

- The unemployed are likely to be the most interested in receiving the training.
- In order to balance the proportion of unemployed in the treatment and control groups, we created two dummy indicators for unemployment and use these new variables together with real earnings in the propensity score model



```
. gen un74 =(re74==0)

. gen un75= (re75==0)
```

Example: Real earning and unemployed subjects

- Assume unconfoundedness holds
- Estimate a logit model for the PS

```
logit treat age ed black hisp marr re74 re75 un74 un75
```

- Predict the pi (x) for each i

```
predict pscore, pr
```

- use **psmatch2** for matching: a simple NN matching without replacement; conditioning on the common support.
  - Since there are observations with identical propensity score values, the sort order of the data could affect matching results.
  - it is advisible to sort randomly the data before calling **psmatch2**.
- use **pstest** to test the balancing

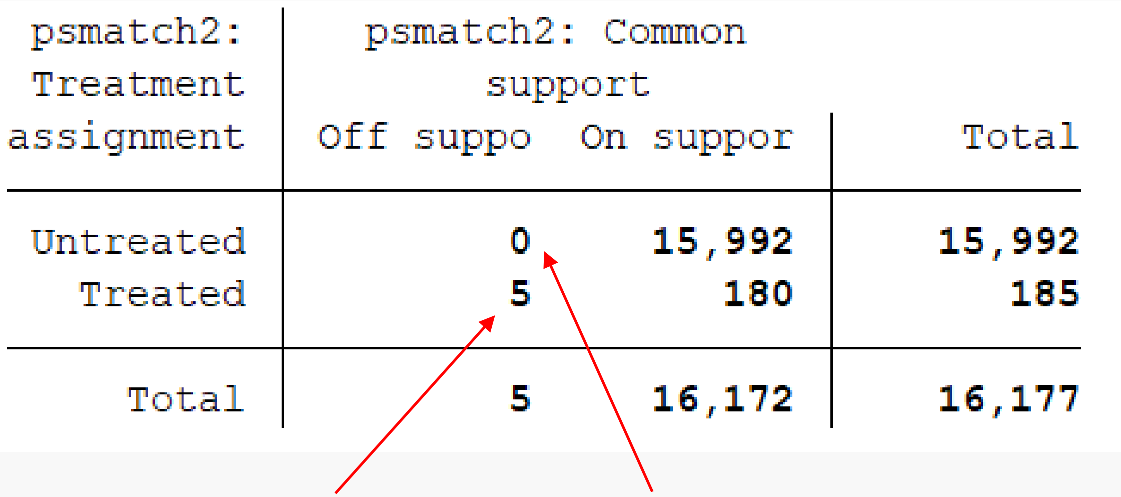
Example: psmatch2 output

```
psmatch2 treat, pscore(pscore) outcome(re78) common noreplacement
```

- the common option imposes a common support by dropping treatment observations whose pscore is higher than the maximum or less than the minimum pscore of the controls.
- Default matching method is single nearest-neighbour (without caliper).
- the noreplacement option perform 1-to-1 matching without replacement (available for NN PS matching only).

Example: psmatch2 output

Summary of units off and on support (here we discard 5 treated units).



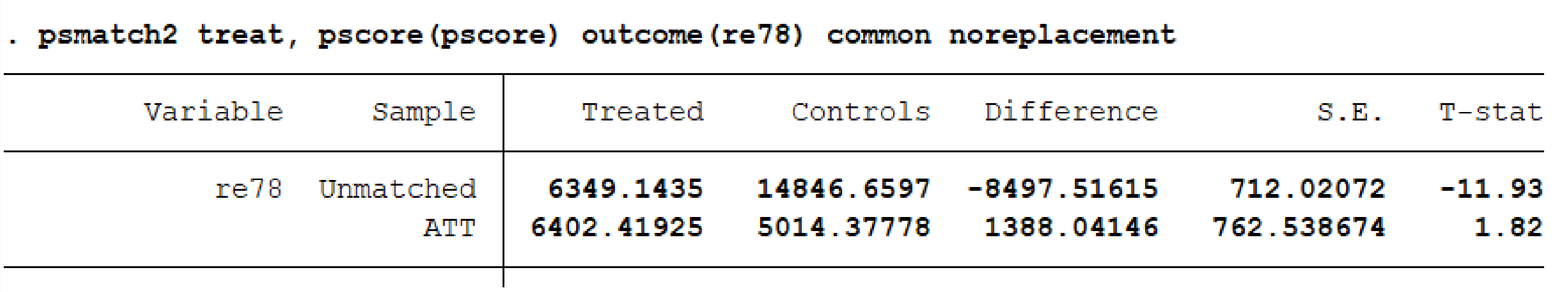
```
  psmatch2: |   psmatch2: Common
  Treatment |        support
 assignment | Off suppo  On suppor |     Total
------------+----------------------+----------
  Untreated |         0     15,992 |    15,992
    Treated |         5        180 |       185
------------+----------------------+----------
      Total |         5     16,172 |    16,177
```

What if there are many treated obs are off support?

What if there are many untreated obs are off support?

Example: psmatch2 output

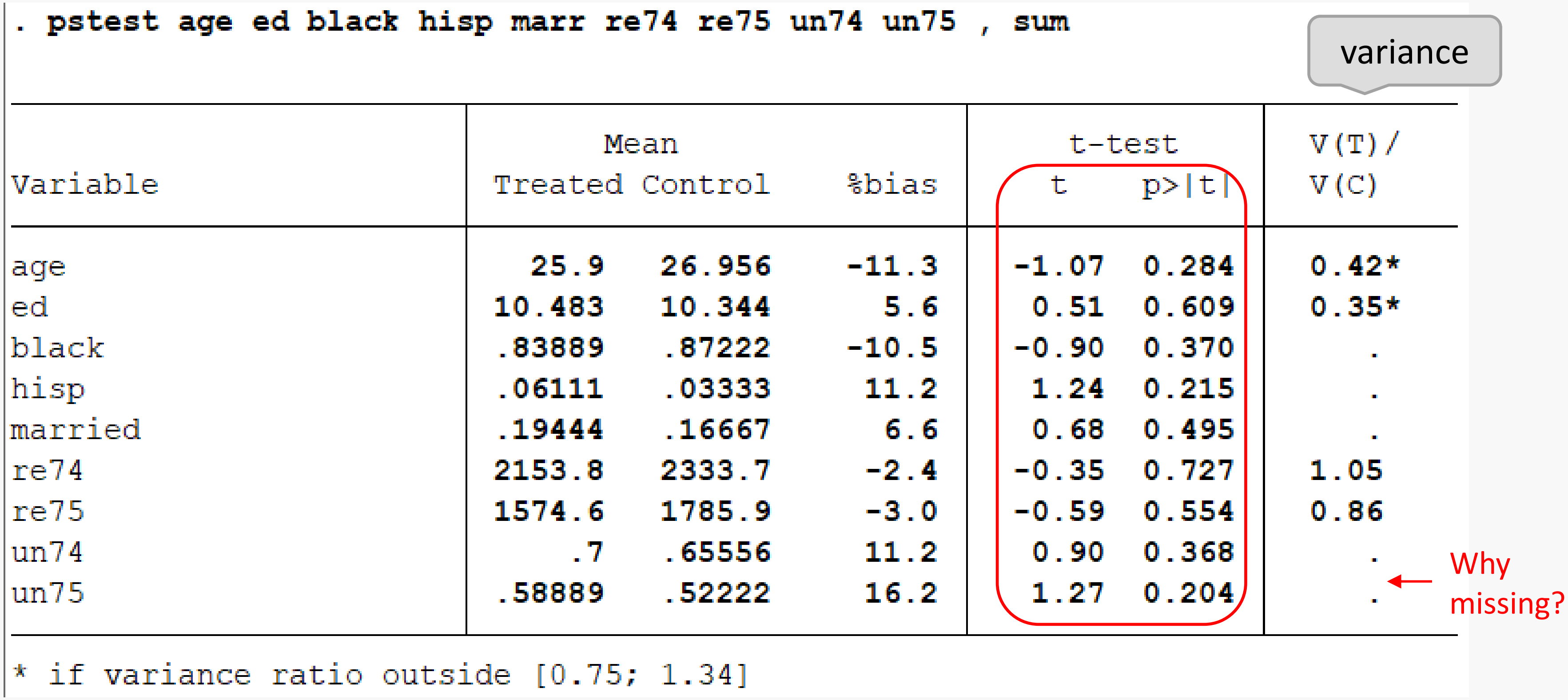
Estimated ATT



```
. psmatch2 treat, pscore(pscore) outcome(re78) common noreplacement
----------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
----------------------------+-----------------------------------------------------------
            re78  Unmatched |  6349.1435   14846.6597  -8497.51615    712.02072   -11.93
                        ATT | 6402.41925   5014.37778   1388.04146   762.538674     1.82
----------------------------+-----------------------------------------------------------
```

We need to check balancing before trusting the ATT estimation!

Balance checking



```
. pstest age ed black hisp marr re74 re75 un74 un75 , sum

----------------------------------------------------------------------
                        |       Mean               |     t-test    |  V(T)/
Variable                | Treated Control    %bias |    t    p>|t| |  V(C)
------------------------+--------------------------+---------------+--------
age                     |   25.9   26.956    -11.3 |  -1.07  0.284 |  0.42*
ed                      | 10.483   10.344      5.6 |   0.51  0.609 |  0.35*
black                   | .83889   .87222    -10.5 |  -0.90  0.370 |      .
hisp                    | .06111   .03333     11.2 |   1.24  0.215 |      .
married                 | .19444   .16667      6.6 |   0.68  0.495 |      .
re74                    | 2153.8   2333.7     -2.4 |  -0.35  0.727 |  1.05
re75                    | 1574.6   1785.9     -3.0 |  -0.59  0.554 |  0.86
un74                    |     .7   .65556     11.2 |   0.90  0.368 |      .
un75                    | .58889   .52222     16.2 |   1.27  0.204 |      .
----------------------------------------------------------------------
* if variance ratio outside [0.75; 1.34]
```

variance

Why missing?

- Very well balanced treated and control groups!

What if psmatch2 does not give balanced sample?

- We can change the propensity score model and re-do the matching
  - interaction terms
  - Higher order terms

- We can change the matching method
  - in the NN method, all treated units find a match. However, some of these matches are fairly poor because for some treated units the nearest neighbor may have a very different propensity score
  - caliper matching and radius matching (among others) offer a solution to this problem

Caliper matching

- NN matching (consider M=1): treated unit i is matched to the non-treated unit j such that

$$\|p_i - p_j\| = min_{k\epsilon W=0}\|p_i - p_k\|$$

- Caliper matching (Cochran and Rubin, 1973) is a variation of NN matching that attempts to avoid bad matches (i.e. p_j far from p_i) by imposing a tolerance on the maximum distance.
- That is, for a pre-specified $\delta > 0$ treated unit i is matched to the non-treated unit j if

$$\delta > \|p_i - p_j\| = min_{k\epsilon W=0}\|p_i - p_k\|$$

- If none of the non-treated units is within δ from treated unit i , i is excluded from the analysis (which is one way of imposing a common support condition).
- A drawback of Caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Radius matching

Each treated unit is matched **only** with the control units whose propensity score falls into a predefined neighborhood of the propensity score of the treated unit.

- all the control units with p_j falling within a radius r from p_i
-

$$\|p_i - p_j\| < r,$$

are matched to the treated unit i.

How to choose the radius?

- The smaller the radius ...
  - ... the better the quality of the matches.
  - ... the higher the possibility that some treated units are not matched because the neighborhood does not contain control units.

References

- Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 84(1), 151-161.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1), 31-72.