WHAT IS (PROPENSITY SCORE) MATCHING?

Barbara Sianesi

PEPA node of NCRM, Institute for Fiscal Studies

5th ESRC Research Methods Festival

Oxford, July 2012

(PS)MATCHING IS EXTREMELY POPULAR…

→ 270,000 entries by googling: propensity score matching

→ 13,000 downloads of –psmatch2–
501st of 1,100,000 items in the RePEc/IDEA database

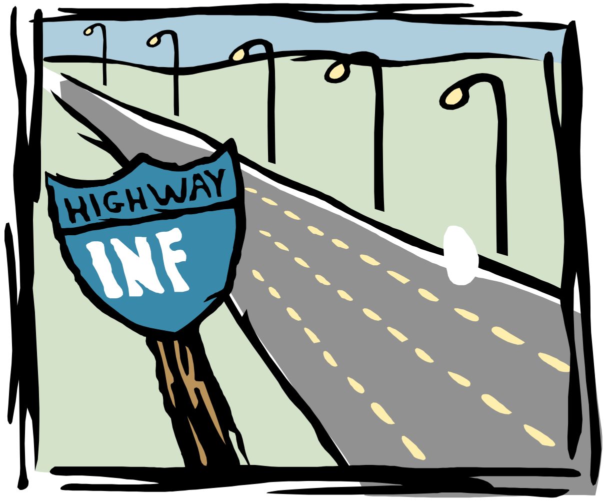
→ >1,500 support emails

- Europe, US, Canada, Central + South America, former SU, Australia, Asia, Africa and the Middle East
- epidemiology, sociology, economics, statistics, criminology, agricultural economics, health economics, transport economics, public health, nutrition, paediatrics, biostatistics, finance, urban planning, geography and geosciences

What is (PS)Matching?

(PS)Matching is a method/device
to make two groups look the same.

Roadmap



1. The counterfactual concept of causality
2. What is matching?
3. How do we use it?
4. Should we use it?

The Counterfactual Concept of Causality

The Evaluation Problem

to evaluate average *causal* effects of a ‘treatment’ on an outcome.

The Potential Outcome model

Y_1	Outcome under treatment
Y_0	Outcome without treatment
$Y_1 - Y_0$	Treatment effect
$D \in \{0, 1\}$	Treatment indicator
$Y = \begin{cases} Y_0 & if\ D = 0 \\ Y_1 & if\ D = 1 \end{cases}$	Observed outcome
X	Set of observed characteristics

The parameters of interest

- ATT $\equiv E(Y_1 - Y_0 \mid D=1) = E(Y \mid D=1) - E(Y_0 \mid D=1)$
- ATNT $\equiv E(Y_1 - Y_0 \mid D=0) = E(Y_1 \mid D=0) - E(Y \mid D=0)$
- ATE $\equiv E(Y_1 - Y_0) = \text{ATT} \cdot P(D=1) + \text{ATNT} \cdot P(D=0)$

The Fundamental Problem of Causal Inference

Need to invoke (untestable) assumptions to identify **average unobserved** ***counterfactuals*****.**

MATCHING METHODS – INTUITION (FOR ATT)

Ex post mimic a RCT by constructing a suitable comparison group by carefully matching treated and non-treated

→ selected comparison group is as similar as possible to the treatment group…in terms of their *observable* characteristics

MATCHING METHODS – ASSUMPTIONS

1. Identifying assumption: **Selection on Observables**
   (conditional independence CIA, exogeneity, ignorability, unconfoundedness)
   All the relevant differences between treated and non-treated are captured in *X*:

   ATT: $E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D=0)$
   ATNT: $E(Y_1 \mid X, D=1) = E(Y_1 \mid X, D=0)$
   ATE: both

2. To give it empirical content: **Common Support**
   We observe participants and non-participants with the same characteristics:

   ATT: $P(D=1 \mid X) < 1$
   ATNT: $0 < P(D=1 \mid X)$
   ATE: $0 < P(D=1 \mid X) <1$

⇒ can use the (observed) mean outcome of the non-treated to estimate the mean (counterfactual) outcome the treated would have had they not been treated.

OPERATIONALISING MATCHING METHODS

Curse of dimensionality

- impose linearity in the parameters (regression analysis)
- choose a distance metric

❖ **Mahalanobis metric**

$d(i,j) = (\boldsymbol{X}_i - \boldsymbol{X}_j)' \, V^{-1} \, (\boldsymbol{X}_i - \boldsymbol{X}_j)$

❖ **Propensity Score** $p(x) \equiv P(D{=}1 \,|\, X{=}x)$

Conditional treatment probability (given confounders X)

The propensity score is a balancing score, i.e.
$X \perp D \,|\, p(X)$

If CIA holds given $X \rightarrow$ CIA holds given $p(X)$

Overview of Matching Estimators

1. pair to each treated *i* some group of ‘comparable’ non-treated individuals
2. associate to the outcome y_i of treated *i*, a matched outcome $\hat{y}_i$ given by the (weighted) outcomes of his ‘neighbours’ in the comparison group:

$$\hat{y}_i = \sum_{j \in C^0(p_i)} w_{ij} y_j$$

- $C^0(p_i)$ = set of neighbours of treated *i* in the *D*=0 group
- w_{ij} = weight on non-treated *j* in forming a comparison with treated *i*, where $\sum_{j \in C^0(p_i)} w_{ij} = 1$

General form of the matching estimator for ATT (within S_{10}):

$$\hat{ATT} = \frac{1}{\#(D=1 \cap S_{10})} \sum_{i \in \{D_i = 1 \cap S_{10}\}} \{y_i - \hat{y}_i\}$$

= $E(Y \mid$ treated on $S_{10}) - E(Y \mid$ matched/reweighted non-treated)

TRADITIONAL MATCHING ESTIMATORS

One-to-one matching

– with or without replacement
– nearest neighbour or within caliper

SIMPLE SMOOTHED MATCHING ESTIMATORS

- *K*-nearest neighbours
  – with or without replacement
  – nearest neighbour or within caliper
- Radius matching

WEIGHTED SMOOTHED MATCHING ESTIMATORS

- Kernel-based matching
- Local linear regression-based matching

❑ bandwidth choice
❑ kernel choice

Checking matching quality

Check (and possibly improve on) balancing of observables

$$\boxed{D \perp X \mid \hat{p}(X)}$$

- for each variable
- overall measures

Inference

- naïve variance
- bootstrapping
- Abadie-Imbens heteroskedasticity-robust standard errors when matching on *X*
- Abadie-Imbens analytical asy std errors taking into account estimation of e(*X*) for PS nearest neighbour(s) matching with replacement

MATCHING VS OLS

- ***same* identifying assumption**

  If unobserved confounders, just as biased as OLS – internal validity

- **avoids any additional assumption**

  (1) **COMMON SUPPORT**

  Matching performed only over Sup_{10}, hence compares only comparable people
  Might recover a different causal impact: $\text{ATT}(\text{Sup}_{10}) \neq \text{ATT}(\text{Sup}_{1})$ – external validity

  (2) **NON-PARAMETRIC**

  Avoids potential misspecification of $E(Y_0 \mid X)$

  Allows for arbitrary X-heterogeneity in impacts $E(Y_1 - Y_0 \mid X)$

  ⇒ Matching focuses on **comparability** in terms of **observables**,
  i.e. on constructing a suitable comparison group by carefully matching treated and non-treated on X / reweighting the non-treated to realign their X

But: if OLS is correctly specified, OLS is more efficient.

BUT we don’t need matching to make OLS less parametric…

FULLY INTERACTED OLS -FILM-

$$Y = m_0(X_1, X_2) + \delta D + \delta_1(X_1 D) + \delta_2(X_2 D) + \delta_{12}(X_1 X_2 D) + e$$

$$\beta_{ATT} = \delta + \delta_1 \bar{X}_{1|D=1} + \delta_2 \bar{X}_{2|D=1} + \delta_{12}(\overline{X_1 X_2})_{|D=1}$$

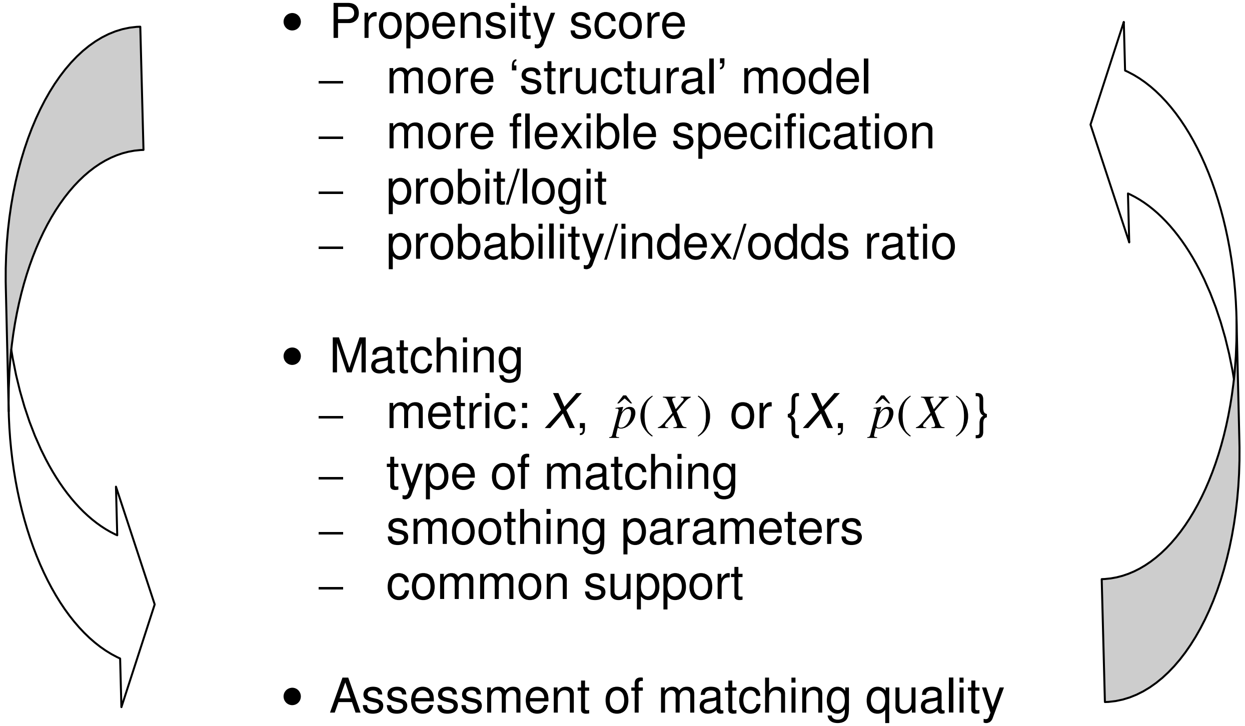
$$\beta_{ATNT} = \delta + \delta_1 \bar{X}_{1|D=0} + \delta_2 \bar{X}_{2|D=0} + \delta_{12}(\overline{X_1 X_2})_{|D=0}$$

$$\beta_{ATE} = \delta + \delta_1 \bar{X}_1 + \delta_2 \bar{X}_2 + \delta_{12}(\overline{X_1 X_2})$$

Can F-test for presence of heterogeneous effects.

STILL, matching (≠ OLS) highlights comparability of groups

Check matching quality



- Propensity score
  - more ‘structural’ model
  - more flexible specification
  - probit/logit
  - probability/index/odds ratio

- Matching
  - metric: X, $\hat{p}(X)$ or $\{X, \hat{p}(X)\}$
  - type of matching
  - smoothing parameters
  - common support

- Assessment of matching quality

CAN we get the two groups balanced (in terms of *X*)?
[Think back to RCT…]

STRENGTHS AND WEAKNESSES

☺ **Advantages** ☺

- controls for selection on observables and on observably heterogeneous impacts
- non-(or semi-) parametric:
  no specific form for outcome equation, decision process or either unobservable term
- Sup_{10}: compare only comparable people and help determining which results reliable
- flexible and easy

☹ **Disadvantages** ☹

- selection on observables: matching as good as its *X*'s
- restricting to Sup_{10} may change parameter being estimated → unable to identify ATT
- data hungry

EXAMPLE: IMPACT OF NSW

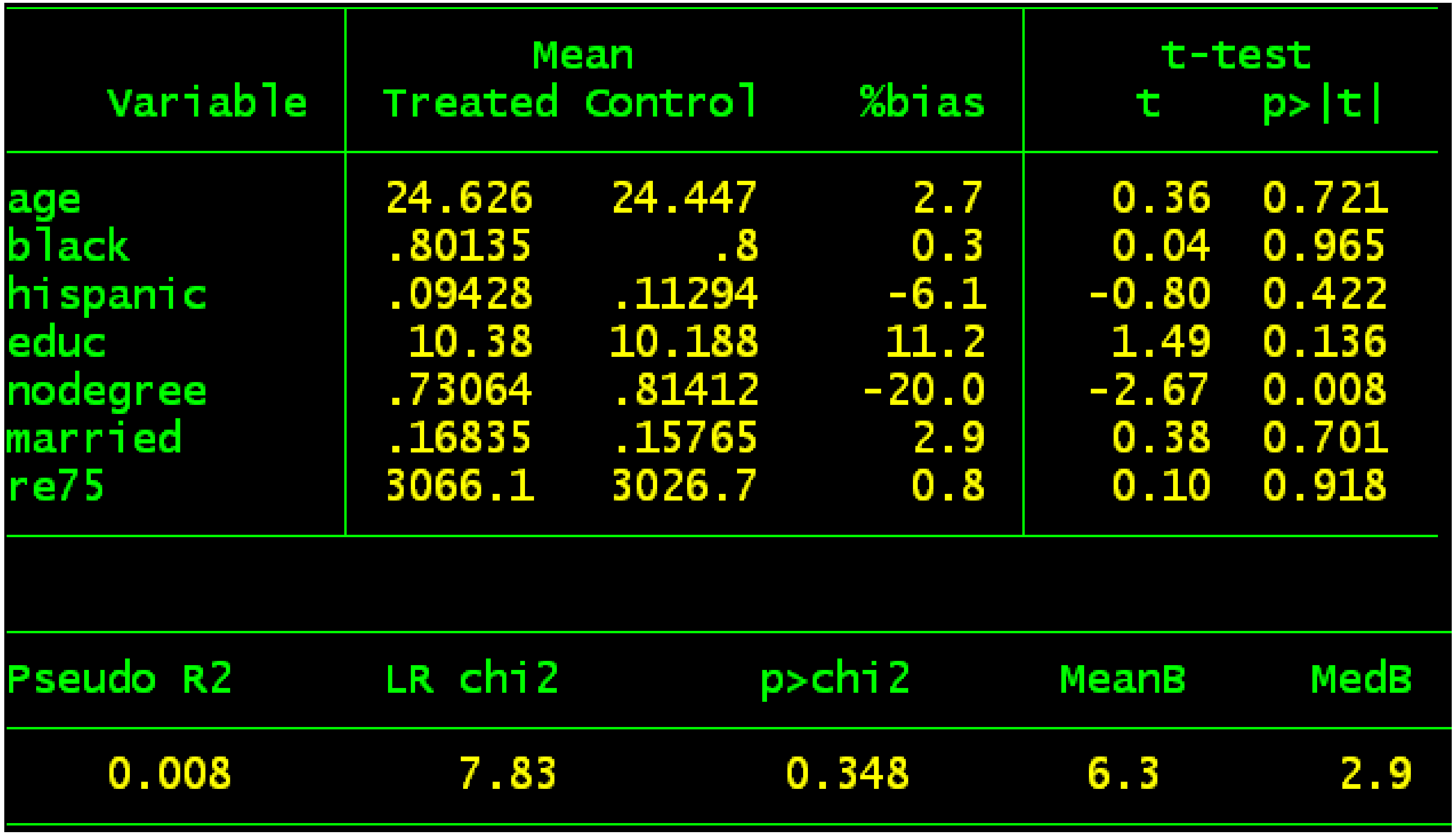
Very famous data in the evaluation literature, combining treatment and controls from a randomised evaluation of the NSW Demonstration with non-experimental individuals drawn from various sources.

NSW male treated (297) with
male comparisons drawn from the PSID (2,490)

Y = real earnings in 1978
X = age, ethnicity (black and hisp), education (years and <12 years), real pre-programme earnings in 1975

COMPARABILITY OF GROUPS

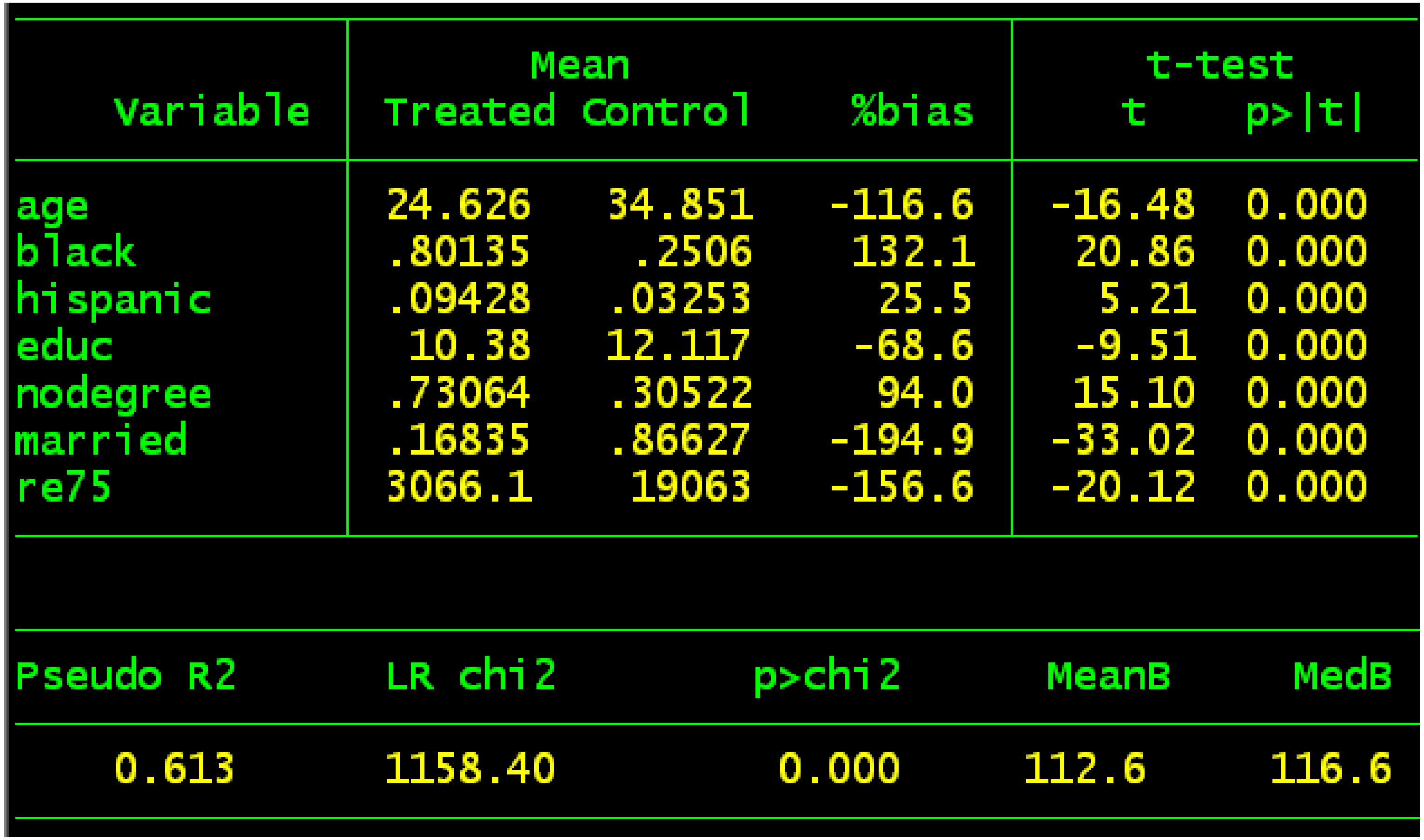
NSW trainees vs NSW control group



```
             |       Mean               |     t-test
    Variable | Treated  Control   %bias |    t    p>|t|
-------------+--------------------------+----------------
age          | 24.626    24.447     2.7 |   0.36  0.721
black        | .80135        .8     0.3 |   0.04  0.965
hispanic     | .09428    .11294    -6.1 |  -0.80  0.422
educ         |  10.38    10.188    11.2 |   1.49  0.136
nodegree     | .73064    .81412   -20.0 |  -2.67  0.008
married      | .16835    .15765     2.9 |   0.38  0.701
re75         | 3066.1    3026.7     0.8 |   0.10  0.918
-------------+--------------------------+----------------


------------------------------------------------------
Pseudo R2     LR chi2       p>chi2      MeanB     MedB
------------------------------------------------------
    0.008        7.83        0.348        6.3      2.9
------------------------------------------------------
```

True ATT (experimental estimator) = 886*

NSW trainees vs PSID comparison group



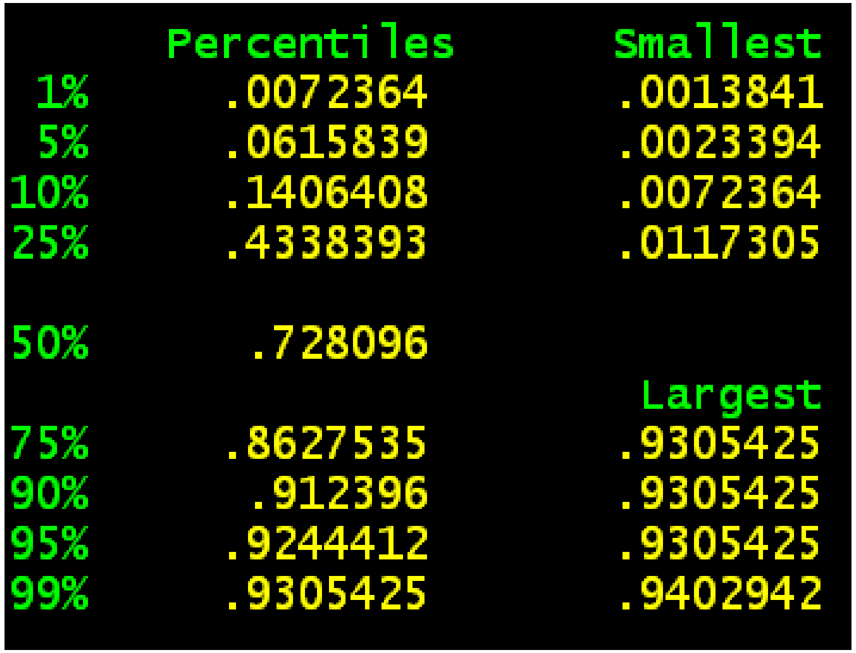
```
                 |       Mean               |     t-test
    Variable     | Treated  Control   %bias |    t    p>|t|
-----------------+--------------------------+----------------
age              |  24.626   34.851  -116.6 | -16.48  0.000
black            |  .80135    .2506   132.1 |  20.86  0.000
hispanic         |  .09428   .03253    25.5 |   5.21  0.000
educ             |   10.38   12.117   -68.6 |  -9.51  0.000
nodegree         |  .73064   .30522    94.0 |  15.10  0.000
married          |  .16835   .86627  -194.9 | -33.02  0.000
re75             |  3066.1    19063  -156.6 | -20.12  0.000
-----------------+--------------------------+----------------


-------------------------------------------------------------
Pseudo R2       LR chi2           p>chi2       MeanB       MedB
-------------------------------------------------------------
    0.613       1158.40            0.000      112.6      116.6
-------------------------------------------------------------
```

→ expect naïve comparison to be downward biased

Naïve estimator = -15,578***

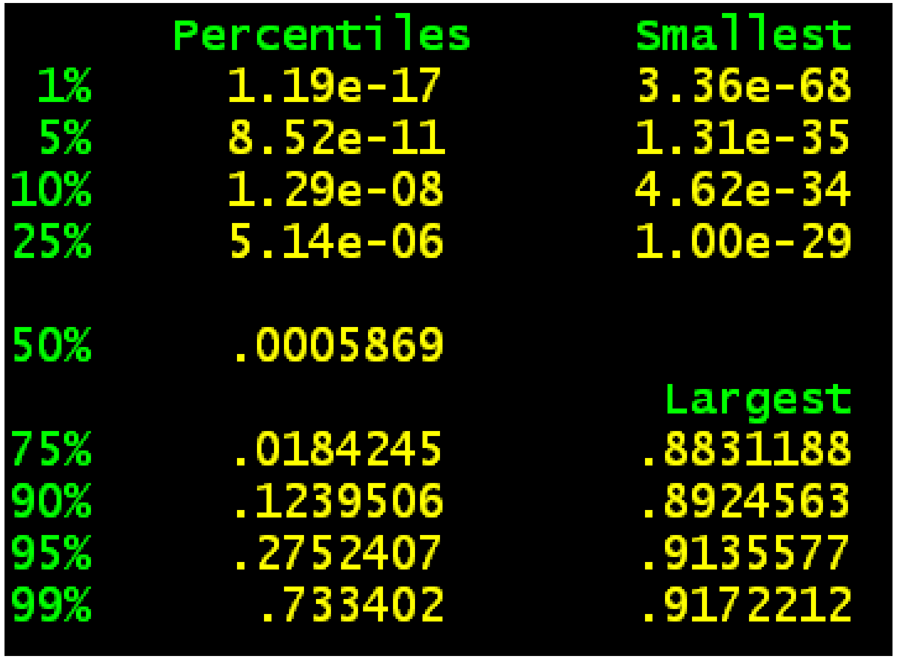
Distribution of $\hat{p}(X)$

NSW treated



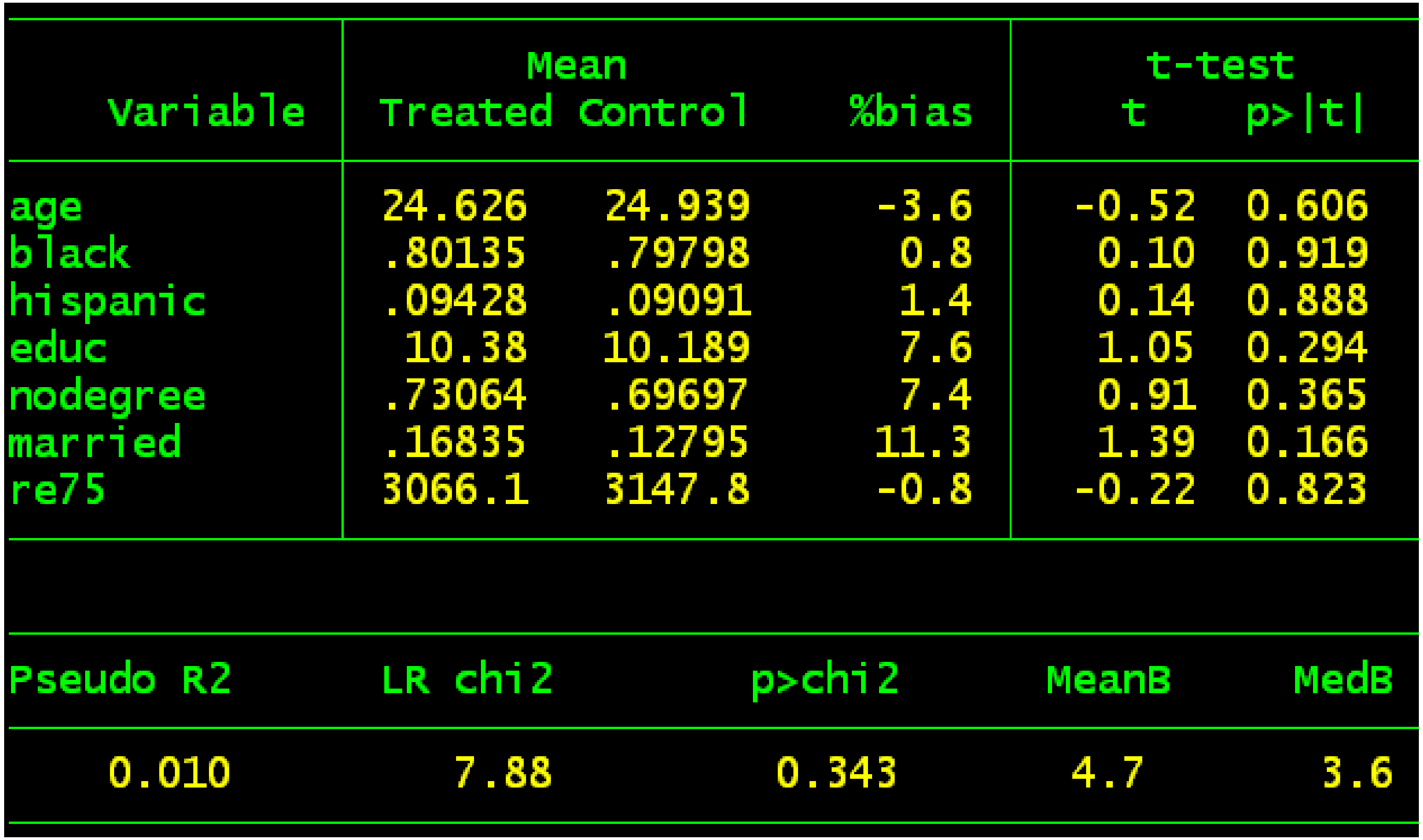
```
        Percentiles      Smallest
 1%      .0072364        .0013841
 5%      .0615839        .0023394
10%      .1406408        .0072364
25%      .4338393        .0117305

50%       .728096
                          Largest
75%      .8627535        .9305425
90%       .912396        .9305425
95%      .9244412        .9305425
99%      .9305425        .9402942
```

PSID comparisons



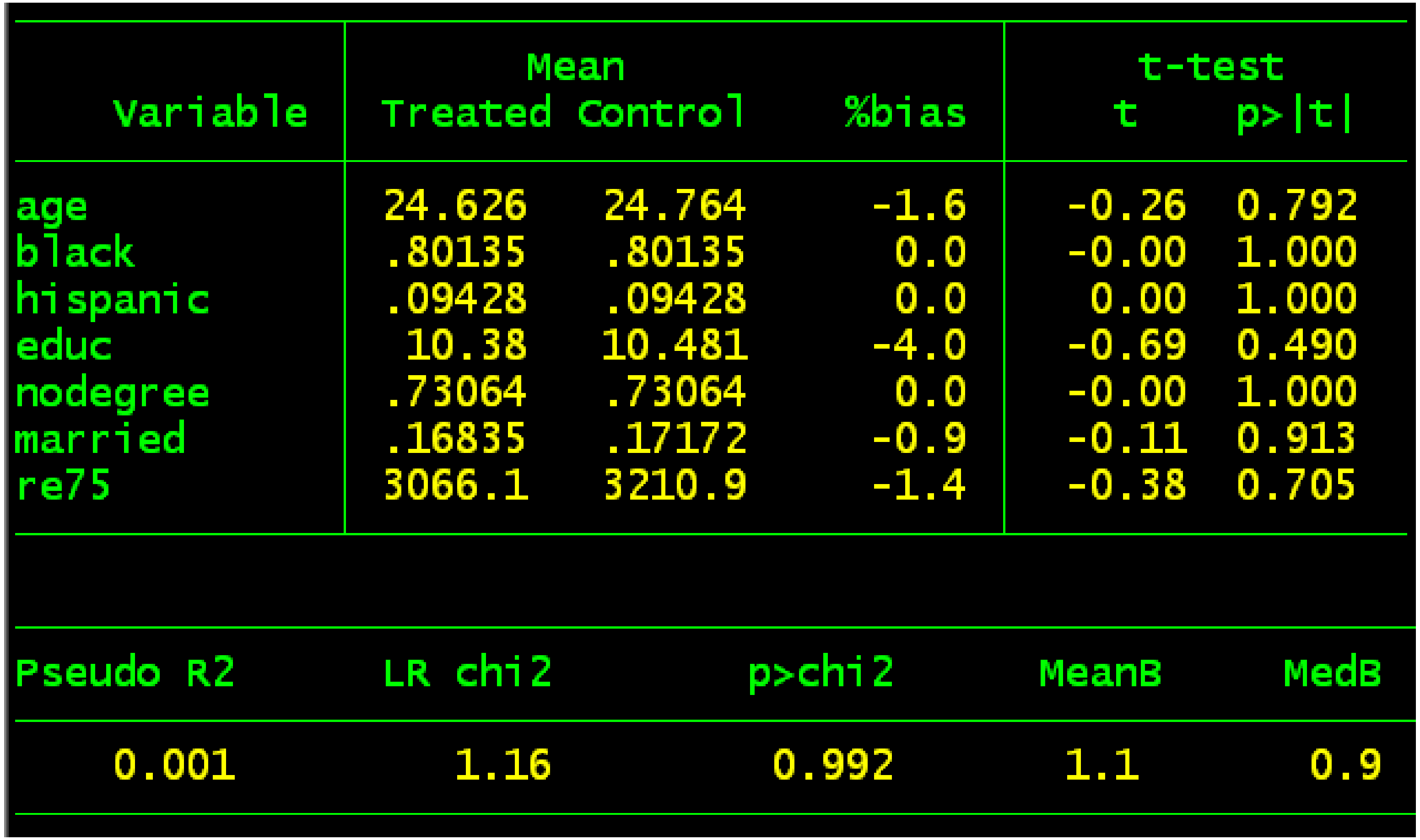
```
        Percentiles      Smallest
 1%      1.19e-17        3.36e-68
 5%      8.52e-11        1.31e-35
10%      1.29e-08        4.62e-34
25%      5.14e-06        1.00e-29

50%      .0005869
                          Largest
75%      .0184245        .8831188
90%      .1239506        .8924563
95%      .2752407        .9135577
99%       .733402        .9172212
```

*NSW trainees vs **matched** PSID comparison group – nearest neighbour (w/ replac)*



```
            |       Mean               |     t-test
   Variable | Treated Control    %bias |    t    p>|t|
------------+--------------------------+----------------
age         | 24.626   24.939     -3.6 |  -0.52  0.606
black       | .80135   .79798      0.8 |   0.10  0.919
hispanic    | .09428   .09091      1.4 |   0.14  0.888
educ        |  10.38   10.189      7.6 |   1.05  0.294
nodegree    | .73064   .69697      7.4 |   0.91  0.365
married     | .16835   .12795     11.3 |   1.39  0.166
re75        | 3066.1   3147.8     -0.8 |  -0.22  0.823
------------+--------------------------+----------------


------------------------------------------------------------
Pseudo R2     LR chi2          p>chi2        MeanB      MedB
------------------------------------------------------------
    0.010        7.88           0.343          4.7       3.6
------------------------------------------------------------
```

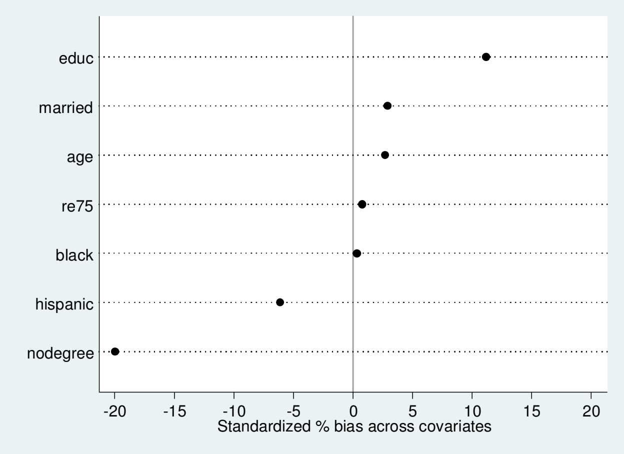
*NSW trainees vs **matched** PSID comparison group – Mahal on X and p(X)*



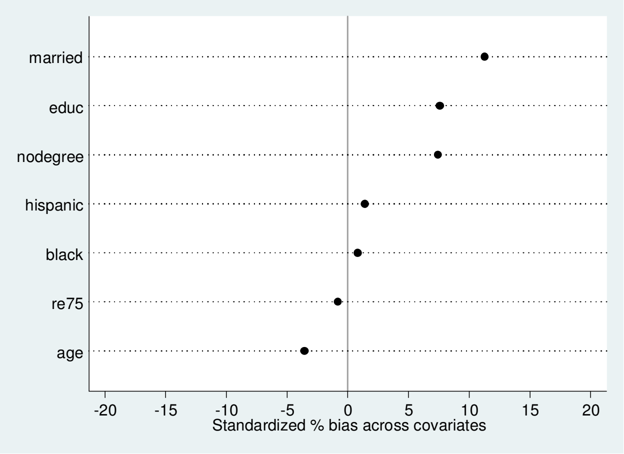
```
               |       Mean               |     t-test
      Variable | Treated Control    %bias |    t    p>|t|
---------------+--------------------------+----------------
age            | 24.626    24.764    -1.6 |  -0.26  0.792
black          | .80135    .80135     0.0 |  -0.00  1.000
hispanic       | .09428    .09428     0.0 |   0.00  1.000
educ           |  10.38    10.481    -4.0 |  -0.69  0.490
nodegree       | .73064    .73064     0.0 |  -0.00  1.000
married        | .16835    .17172    -0.9 |  -0.11  0.913
re75           | 3066.1    3210.9    -1.4 |  -0.38  0.705
--------------------------------------------------------


---------------------------------------------------------------
Pseudo R2       LR chi2        p>chi2      MeanB       MedB
---------------------------------------------------------------
    0.001          1.16         0.992       1.1         0.9
---------------------------------------------------------------
```

Achieved balancing

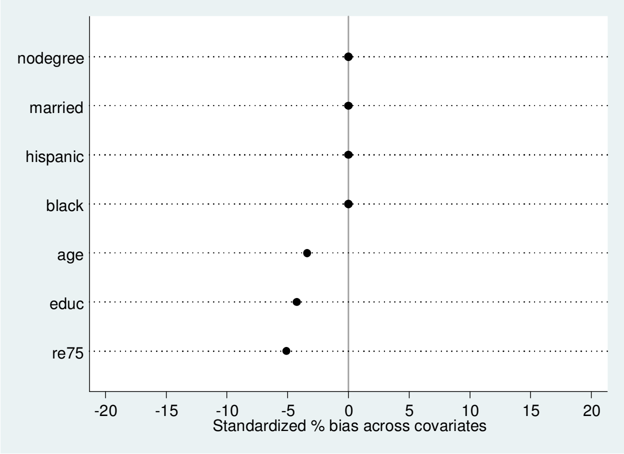
Randomisation



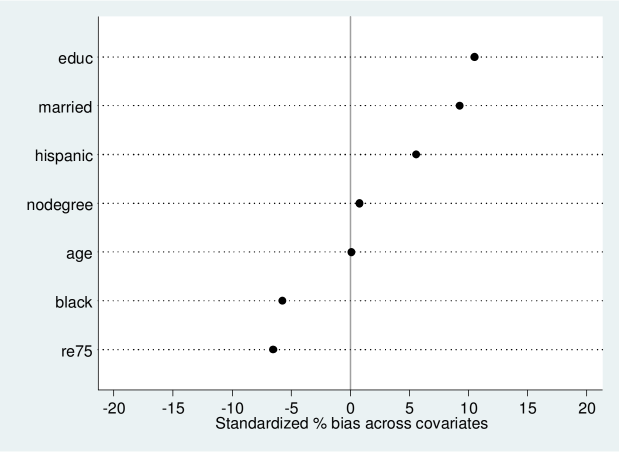
Nearest Neighbour, with replacement



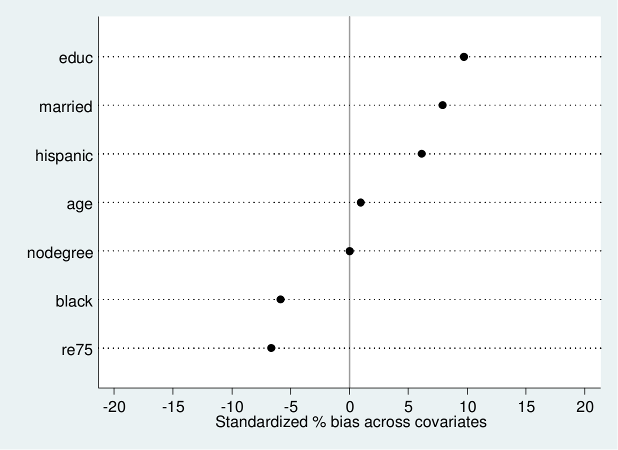
Mahalanobis



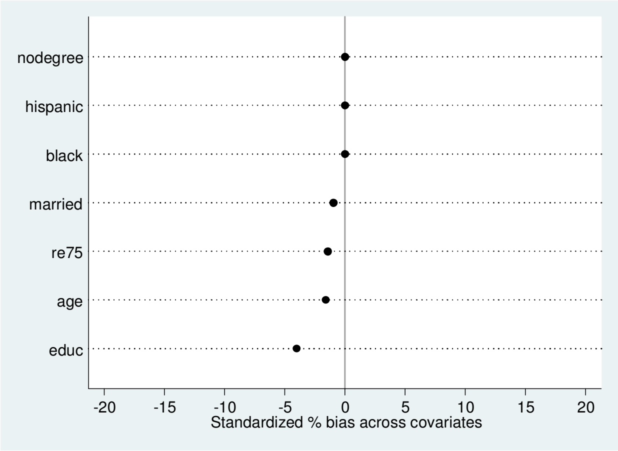
Kernel (epan, *h*=0.01)



Radius (*r*=0.01)

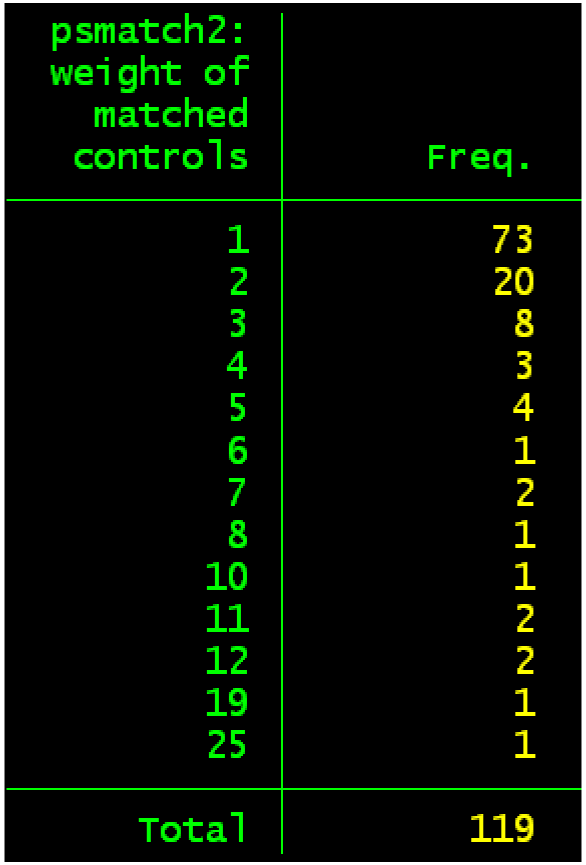


Augmented Mahalanobis



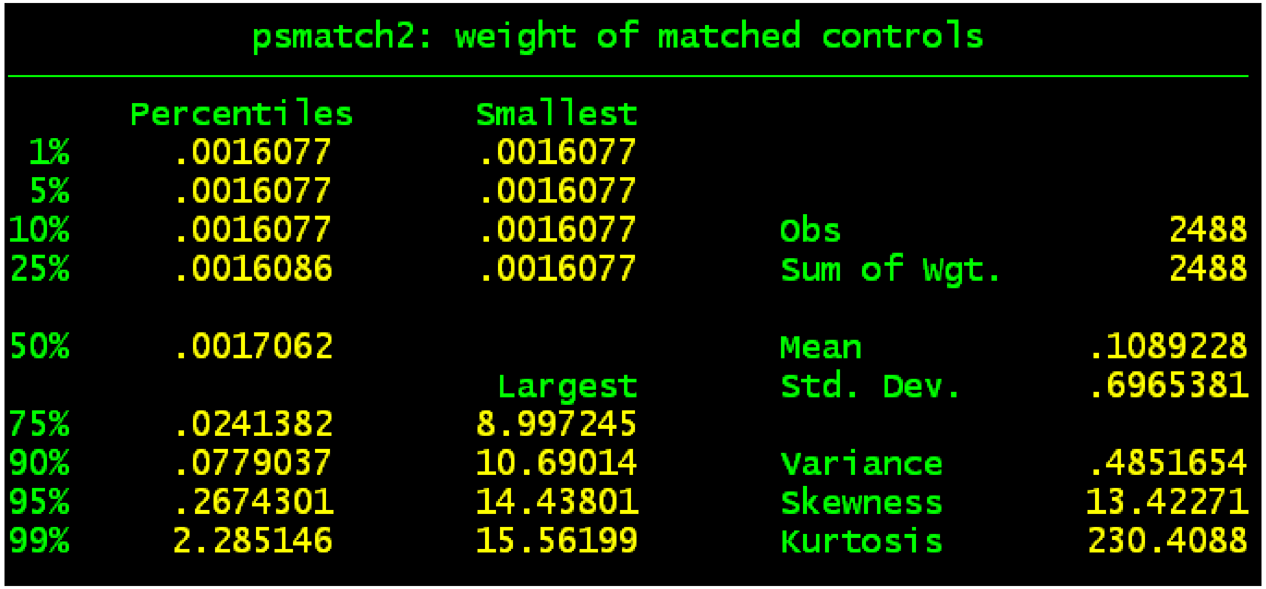
How many PSID members are we *really* using?

Nearest neighbour (w/ replac)



```
psmatch2: |
weight of |
  matched |
 controls |      Freq.
----------+-----------
        1 |         73
        2 |         20
        3 |          8
        4 |          3
        5 |          4
        6 |          1
        7 |          2
        8 |          1
       10 |          1
       11 |          2
       12 |          2
       19 |          1
       25 |          1
----------+-----------
    Total |        119
```

Kernel (Epan, h=0.01)



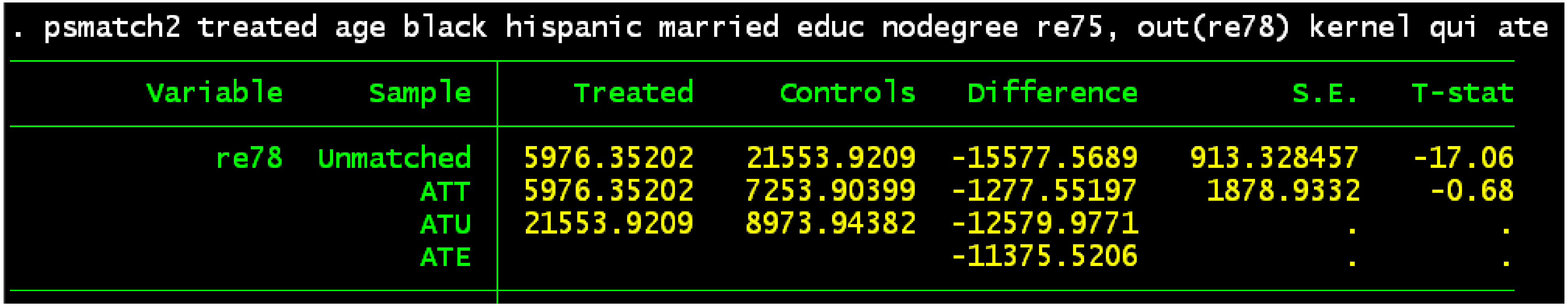
```
            psmatch2: weight of matched controls
-------------------------------------------------------------
      Percentiles      Smallest
 1%      .0016077       .0016077
 5%      .0016077       .0016077
10%      .0016077       .0016077       Obs                2488
25%      .0016086       .0016077       Sum of Wgt.        2488

50%      .0017062                      Mean           .1089228
                        Largest        Std. Dev.      .6965381
75%      .0241382       8.997245
90%      .0779037       10.69014       Variance       .4851654
95%      .2674301       14.43801       Skewness       13.42271
99%      2.285146       15.56199       Kurtosis       230.4088
```

Impact estimates

True ATT (experimental estimator)	886*
Naïve estimator	-15,578***
OLS	-1,458*
FILM	-1,361*
Nearest neighbour (w/ replacement)	551
Kernel (Epan, *h*=0.01)	-737
Augmented Mahalanobis	-830

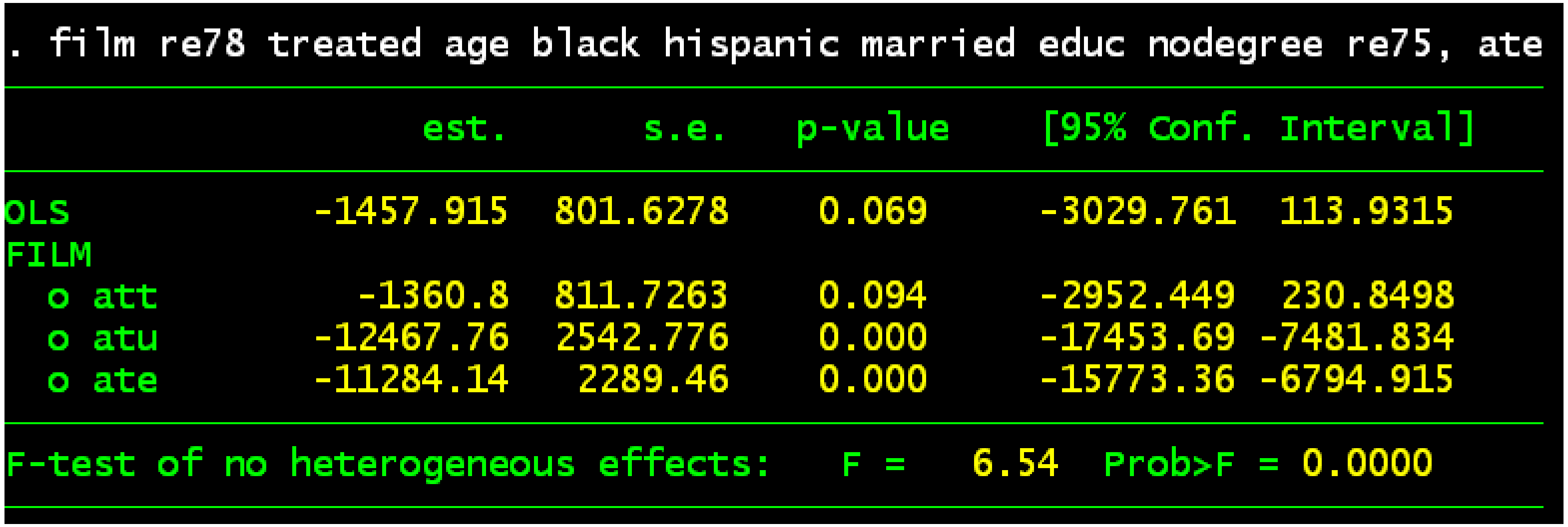
ATNT Average effect of NSW programme had the PSID participated in it

Kernel PS matching (epan, *h*=0.06)



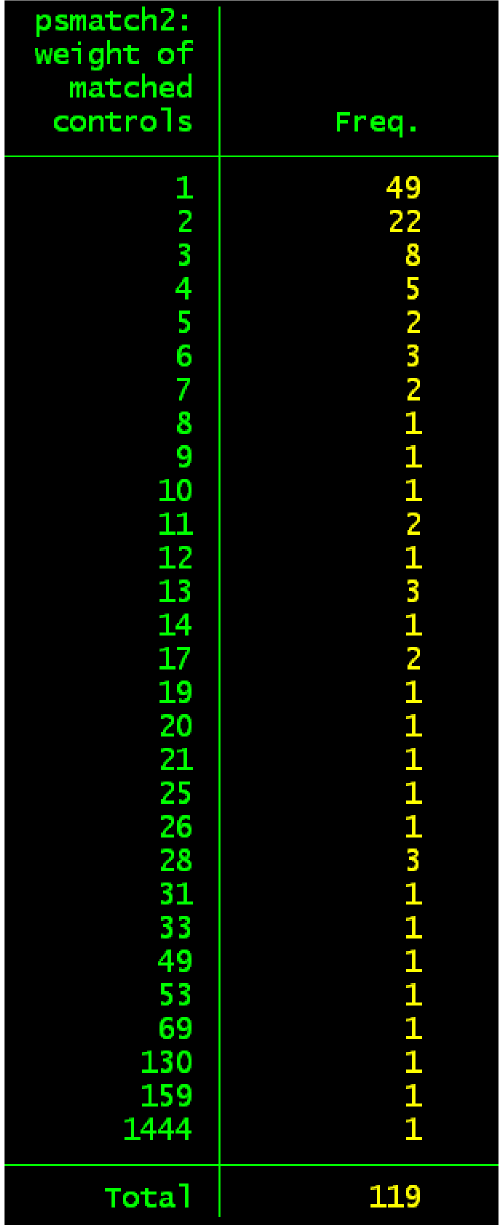
```
. psmatch2 treated age black hispanic married educ nodegree re75, out(re78) kernel qui ate
----------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
----------------------------+-----------------------------------------------------------
            re78  Unmatched | 5976.35202   21553.9209  -15577.5689   913.328457   -17.06
                        ATT | 5976.35202   7253.90399  -1277.55197    1878.9332    -0.68
                        ATU | 21553.9209   8973.94382  -12579.9771            .        .
                        ATE |                          -11375.5206            .        .
----------------------------+-----------------------------------------------------------
```

Fully interacted regression model



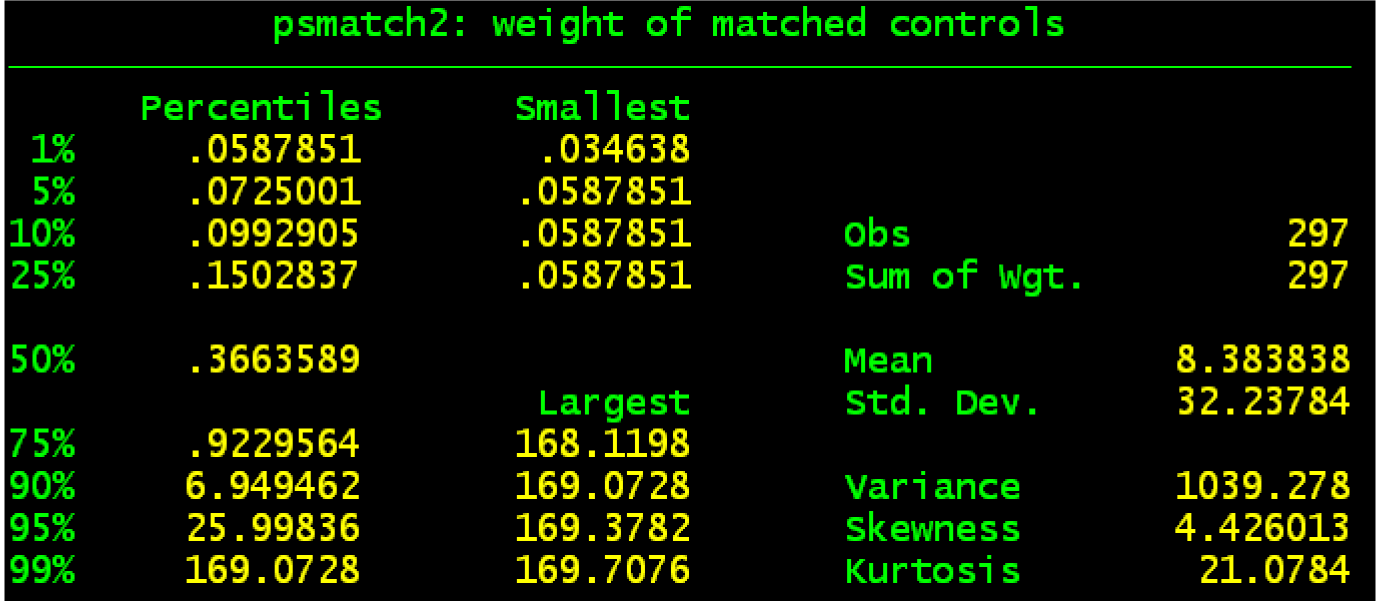
```
. film re78 treated age black hispanic married educ nodegree re75, ate
-----------------------------------------------------------------------
                  est.        s.e.    p-value      [95% Conf. Interval]
-----------------------------------------------------------------------
OLS          -1457.915    801.6278      0.069      -3029.761   113.9315
FILM
  o att        -1360.8    811.7263      0.094      -2952.449   230.8498
  o atu      -12467.76    2542.776      0.000      -17453.69  -7481.834
  o ate      -11284.14     2289.46      0.000      -15773.36  -6794.915
-----------------------------------------------------------------------
F-test of no heterogeneous effects:    F =    6.54   Prob>F = 0.0000
-----------------------------------------------------------------------
```

Nearest neighbour



```
 psmatch2: │
 weight of │
   matched │
  controls │      Freq.
───────────┼───────────
         1 │         49
         2 │         22
         3 │          8
         4 │          5
         5 │          2
         6 │          3
         7 │          2
         8 │          1
         9 │          1
        10 │          1
        11 │          2
        12 │          1
        13 │          3
        14 │          1
        17 │          2
        19 │          1
        20 │          1
        21 │          1
        25 │          1
        26 │          1
        28 │          3
        31 │          1
        33 │          1
        49 │          1
        53 │          1
        69 │          1
       130 │          1
       159 │          1
      1444 │          1
───────────┼───────────
     Total │        119
```

Kernel (Epan, *h*=0.06)



```
                psmatch2: weight of matched controls
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0587851        .034638
 5%     .0725001       .0587851
10%     .0992905       .0587851       Obs                 297
25%     .1502837       .0587851       Sum of Wgt.         297

50%     .3663589                      Mean           8.383838
                        Largest       Std. Dev.      32.23784
75%     .9229564       168.1198
90%     6.949462       169.0728       Variance       1039.278
95%     25.99836       169.3782       Skewness       4.426013
99%     169.0728       169.7076       Kurtosis       21.0784
```

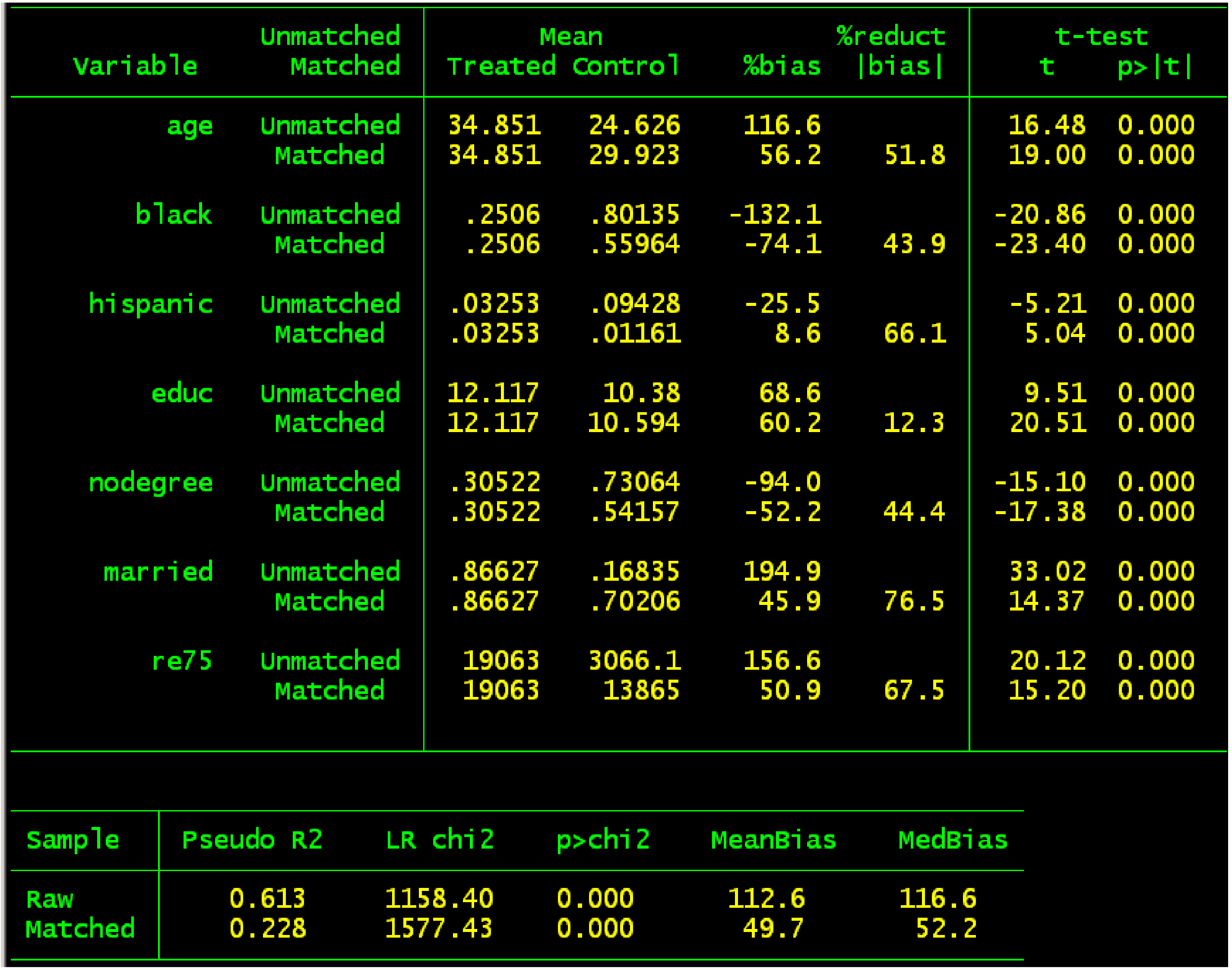
ATNT: –12,580*** (matching) ≈ –12,468*** (film)

Good the PSID did not go into the programme!

Or is it…? ⇨

And now that we are thinking about it…

Do we really want to know the impact the NSW would have had on the full PSID has they participated?!?



```
                Unmatched |       Mean               %reduct |     t-test
    Variable      Matched | Treated Control    %bias  |bias| |    t    p>|t|
--------------------------+----------------------------------+---------------
         age    Unmatched | 34.851   24.626    116.6         |  16.48  0.000
                  Matched | 34.851   29.923     56.2    51.8 |  19.00  0.000
                          |                                  |
       black    Unmatched |  .2506   .80135   -132.1         | -20.86  0.000
                  Matched |  .2506   .55964    -74.1    43.9 | -23.40  0.000
                          |                                  |
    hispanic    Unmatched | .03253   .09428    -25.5         |  -5.21  0.000
                  Matched | .03253   .01161      8.6    66.1 |   5.04  0.000
                          |                                  |
        educ    Unmatched | 12.117    10.38     68.6         |   9.51  0.000
                  Matched | 12.117   10.594     60.2    12.3 |  20.51  0.000
                          |                                  |
    nodegree    Unmatched | .30522   .73064    -94.0         | -15.10  0.000
                  Matched | .30522   .54157    -52.2    44.4 | -17.38  0.000
                          |                                  |
     married    Unmatched | .86627   .16835    194.9         |  33.02  0.000
                  Matched | .86627   .70206     45.9    76.5 |  14.37  0.000
                          |                                  |
        re75    Unmatched |  19063   3066.1    156.6         |  20.12  0.000
                  Matched |  19063    13865     50.9    67.5 |  15.20  0.000
                          |                                  |
--------------------------+----------------------------------+---------------


 ---------------------------------------------------------------------
| Sample  | Pseudo R2   LR chi2   p>chi2   MeanBias   MedBias
|---------+-----------------------------------------------------------
| Raw     |     0.613   1158.40    0.000      112.6     116.6
| Matched |     0.228   1577.43    0.000       49.7      52.2
 ---------------------------------------------------------------------
```

WRAPPING UP…

SELECTION ON UNOBSERVABLES

- Set of conditioning X matters
  $\Rightarrow$ better data help a lot!

SELECTION ON OBSERVABLES

- Avoid use of functional forms in constructing counterfactual
  $\Rightarrow$ (matching $\approx$ fully interacted OLS) $\succ$ simple OLS
  no mis-specification bias

- Compare comparable people
  $\Rightarrow$ matching $\succ$ fully interacted OLS
  highlight – actual comparability of groups,
  – hence reliability (& relevance) of estimates

SELECTED REFERENCES

A comprehensive review

Imbens, G. (2004), ‘Semiparametric estimation of average treatment effects under exogeneity: a review’, *Review of Economics and Statistics*, 86, 4-29.

The propensity score

Rosenbaum, P.R. and Rubin, D.B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika*, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984), “Reducing Bias in Observational Studies Using Sub-Classification on the Propensity Score”, *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P.R. and Rubin, D.B. (1985), “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score”, *The American Statistician*, 39, 1, 33-38.

Dehejia, R.H. and Wahba, S. (1999), “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes”, *Journal of American Statistical Association*, 94, 1053-1062.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997), “Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme”, *Review of Economic Studies*, 64, 605-654.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1998), “Matching as an Econometric Evaluation Estimator”, *Review of Economic Studies*, 65, 261-294.

Mahalanobis-metric matching

Rubin, D.B. (1979), “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies”, *Journal of the American Statistical Association*, 74, 318-328.

Rubin, D.B. (1980), “Bias Reduction Using Mahalanobis-Metric Matching”, *Biometrics*, 36, 293-298.

Multiple treatments

Imbens, G.W. (2000), “The Role of Propensity Score in Estimating Dose-Response Functions”, *Biometrika*, 87, 706-710.

Lechner, M. (2001), Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies*, Heidelberg: Physica/Springer, 43-58.

Inference/Efficiency issues

Abadie, A. and Imbens, G. (2011), “Matching on the Estimated Propensity Score”, mimeo.

Abadie, A. and Imbens, G. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects”, *Econometrica*, 74, 235-267.

Hahn, J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, 66, 315-331.

Hirano, K., G. Imbens, and G. Ridder (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161-1189.