Propensity Score Strategies for Causal Inference

Course Roadmap

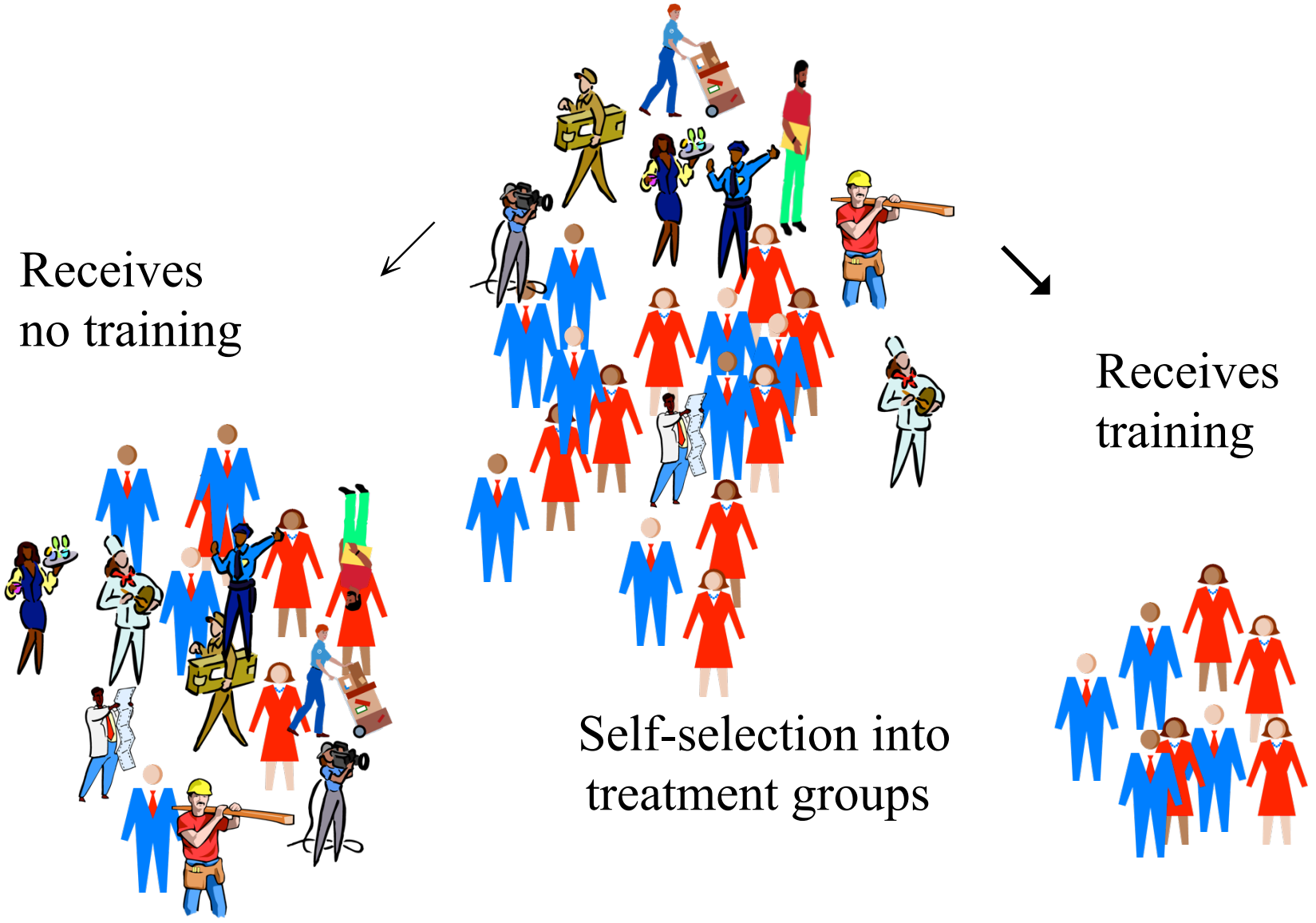
- Defining causal effects
- Estimating causal effects with randomized experiments
- Estimating causal effects with observational studies
  - Stratification/regression
  - Propensity score matching
  - Instrumental variables
  - Difference in differences
  - Fixed effects
  - Regression discontinuity
  - Sensitivity analysis

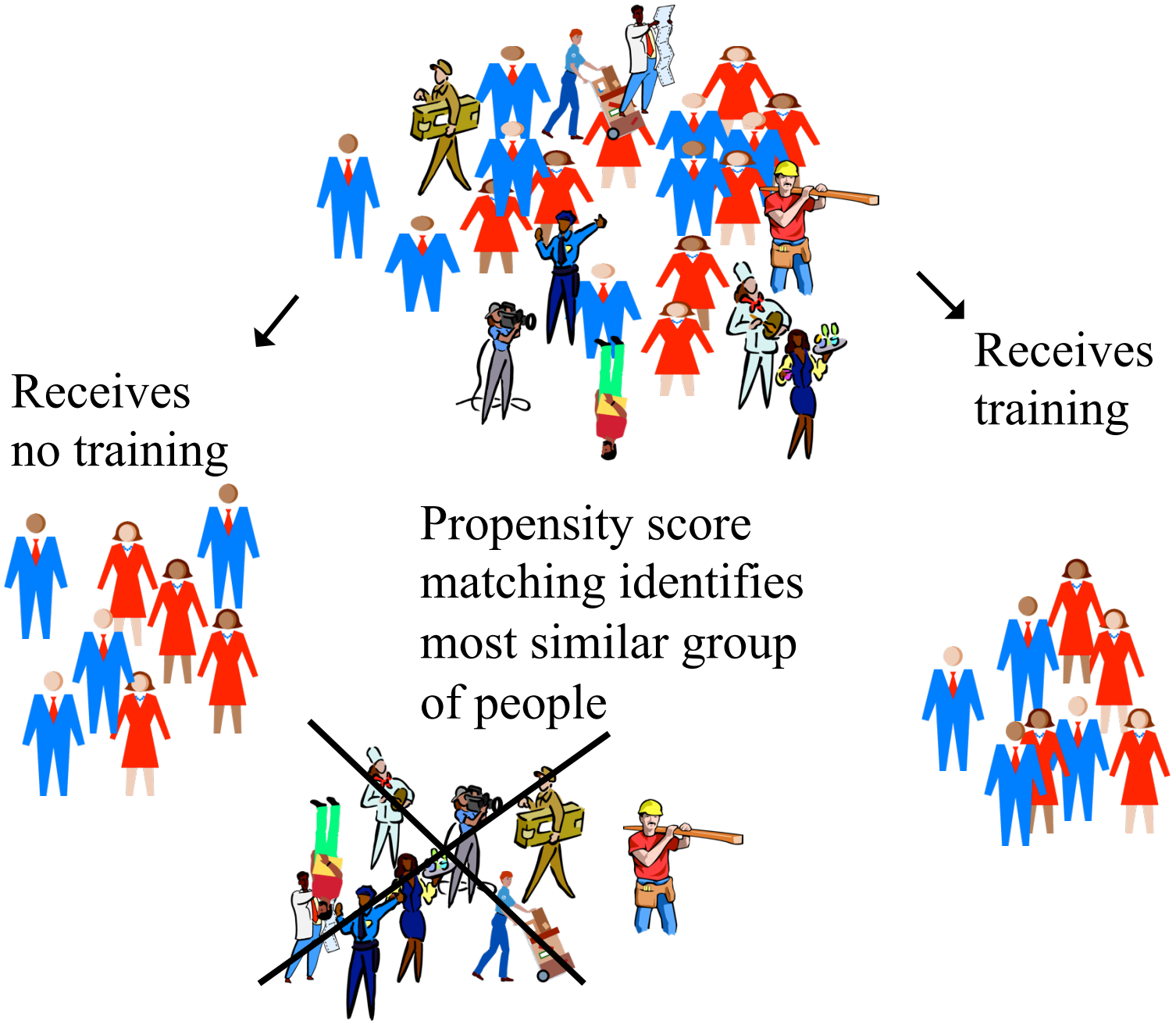
Road Map

- Propensity score matching
  - Intuition
  - Propensity score theory
  - Five principal steps when using propensity score to estimate causal effects
  - Theory
  - Complications and issues to consider
- Other ways to use propensity scores!

  (fancy matching, IPTW, subclassifcation…)
- Related issues

Propensity Score Matching Intuition


Receives
no training
Receives
training
Self-selection into
treatment groups


Receives
no training
Propensity score
matching identifies
most similar group
of people
Receives
training

Hypothetical Example

Person	Treat	Educ.	Age	Y0	Y1	Y	pscore
1	1	1	26	10	14	14	0.68
2	1	1	21	8	12	12	0.65
3	1	1	30	12	16	16	0.71
4	1	1	19	8	12	12	0.63
5	1	0	25	6	10	10	0.33
6	1	0	22	4	8	8	0.24
7	0	0	21	4	8	4	0.22
8	0	0	26	6	10	6	0.36
9	0	0	28	8	12	8	0.43
10	0	0	20	4	8	4	0.19
11	0	1	26	10	14	10	0.68
12	0	1	21	8	12	8	0.65
13	0	0	16	2	6	2	0.12
14	0	0	15	1	5	1	0.10

Matched to	Person	Treat	Educ.	Age	Y0	Y1	Y	pscore
11	1	1	1	26	10	14	14	0.68
12	2	1	1	21	8	12	12	0.65
11	3	1	1	30	12	16	16	0.71
12	4	1	1	19	8	12	12	0.63
8	5	1	0	25	6	10	10	0.33
7	6	1	0	22	4	8	8	0.24
	7	0	0	21	4	8	4	0.22
	8	0	0	26	6	10	6	0.36
	9	0	0	28	8	12	8	0.43
	10	0	0	20	4	8	4	0.19
	11	0	1	26	10	14	10	0.68
	12	0	1	21	8	12	8	0.65
	13	0	0	16	2	6	2	0.12
	14	0	0	15	1	5	1	0.10

!!POP Quiz!!

1. Based on previous hypothetical example, could you create your own matched data and estimate the average treatment effect on the treated (ATT).
2. How do you create your matched data set?
3. How would you decide whether the matching did a good job?
4. Which causal assumption(s) does matching help you to achieve?

Results of matching in our hypothetical example

$$\bar{a}_{T=1} - \bar{a}_{matched} = 23.83 - 23.5 = 0.33$$

$$\bar{e}_{T=1} - \bar{e}_{matched} = 0.67 - 0.67 = 0.0$$

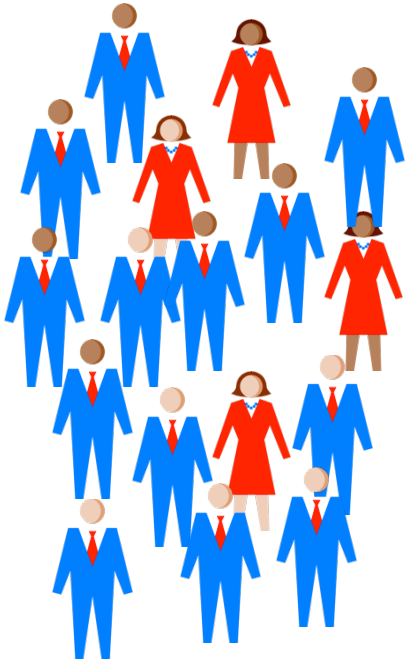
$$\bar{Y}_{T=1} - \bar{Y}_{matched} = 12.0 - 7.67 = 4.33$$

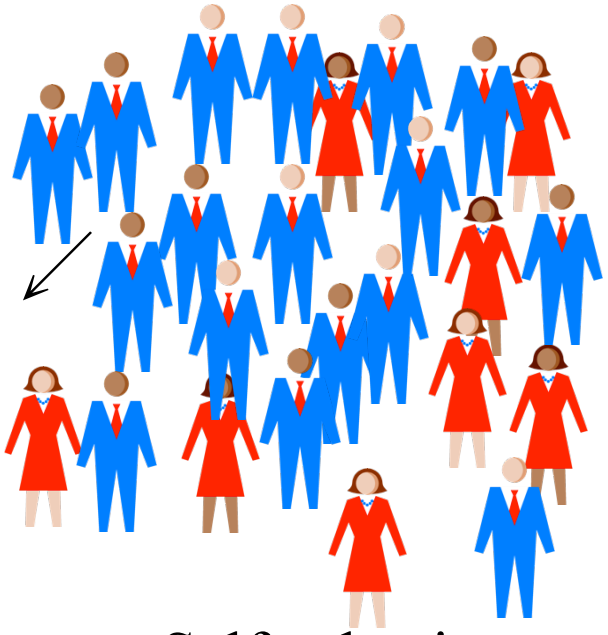
Defining the estimand

Researchers often fail to identify what estimand they are actually trying to estimate (ATE, ATT, ….)

In other words, what group of observations are we trying to make inferences (draw conclusions) about?

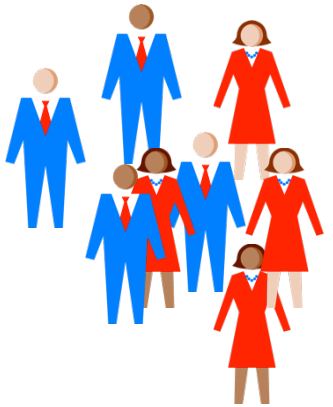
Receives
no training



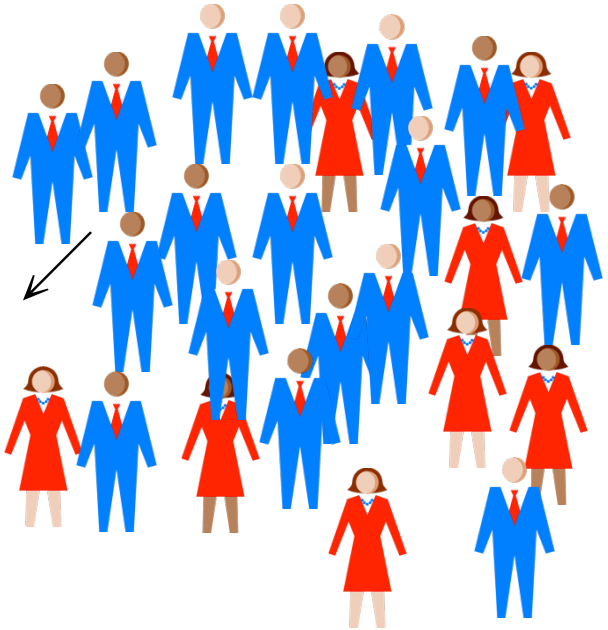


Self-selection
into
treatment groups

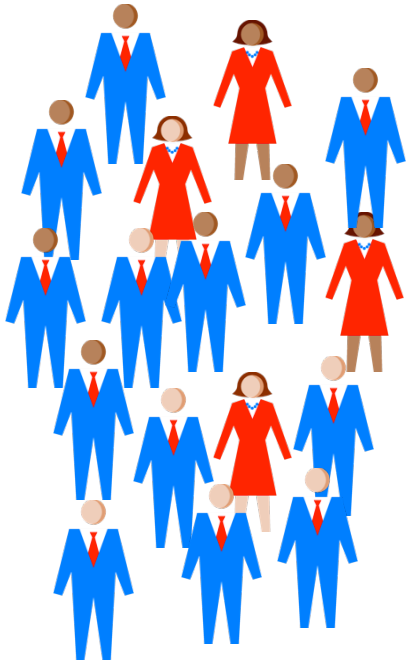
Receives
training



Suppose we observe the following

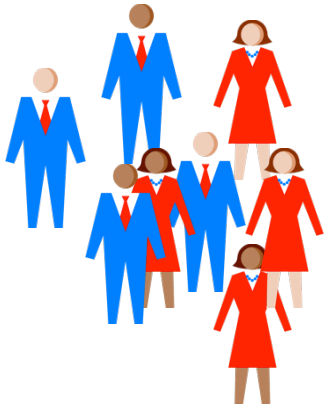


Receives no training

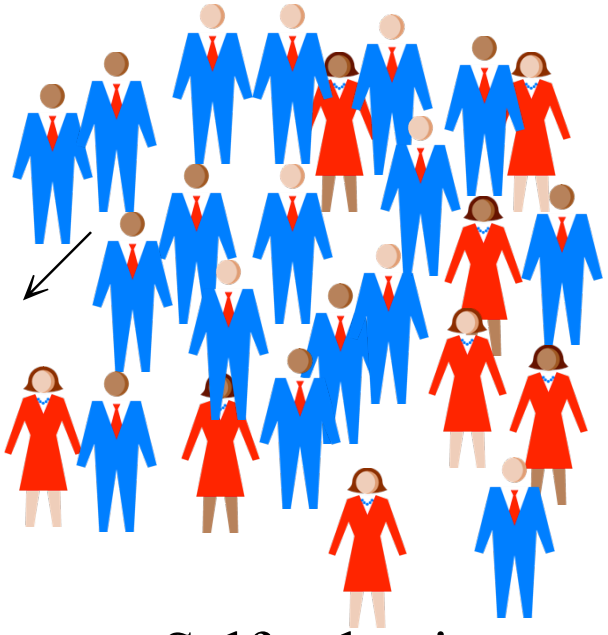


Self-selection into treatment groups

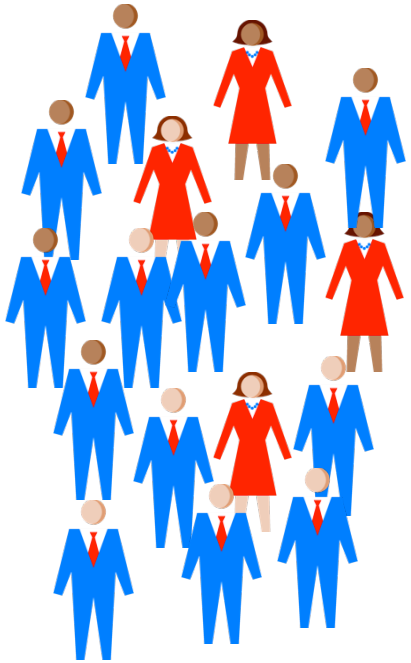
Receives training



Suppose we observe the following

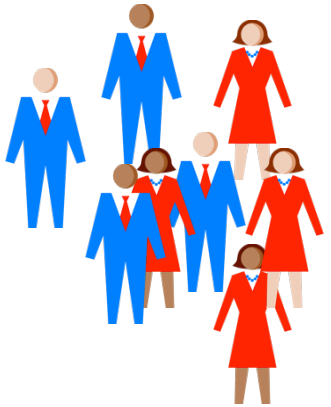


Receives no training



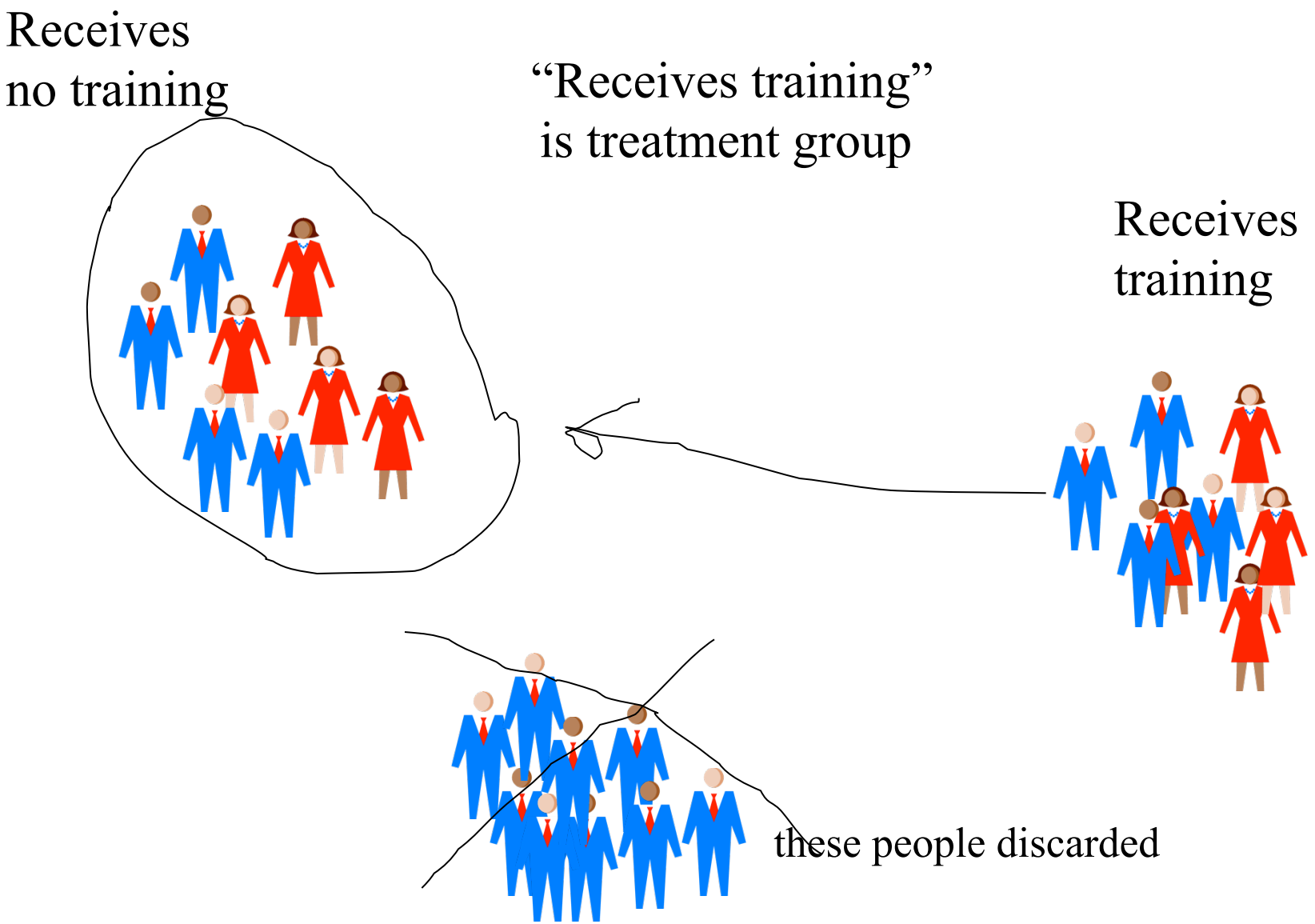
Self-selection into treatment groups

Receives training

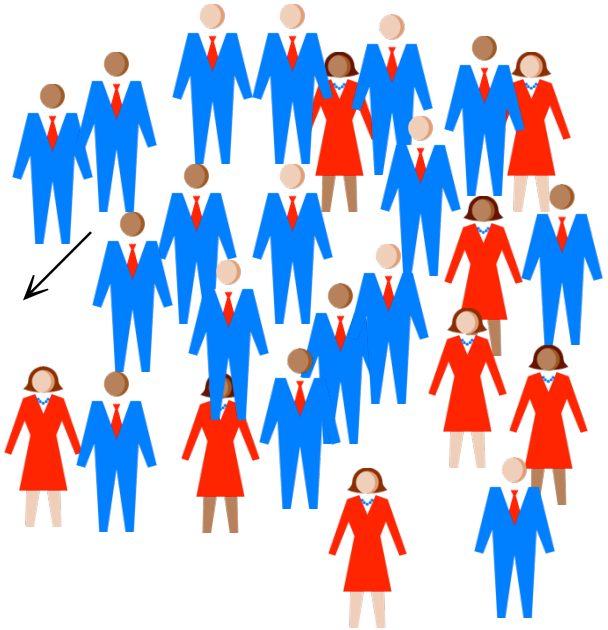


How could we use matching to estimate the effect of the treatment on the treated?

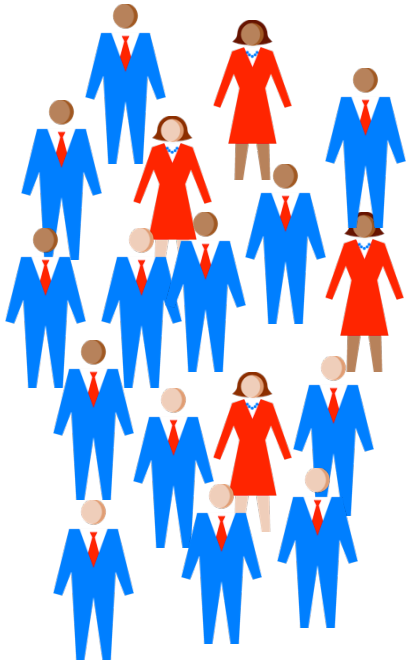
Average Treatment effect on Treated



Suppose we observe the following

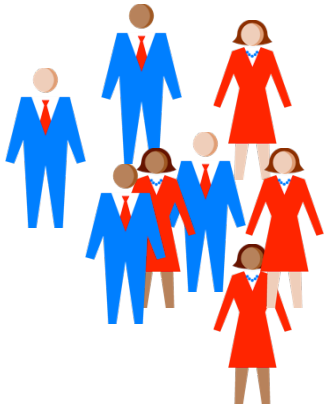


Receives no training



Self-selection into treatment groups

Receives training



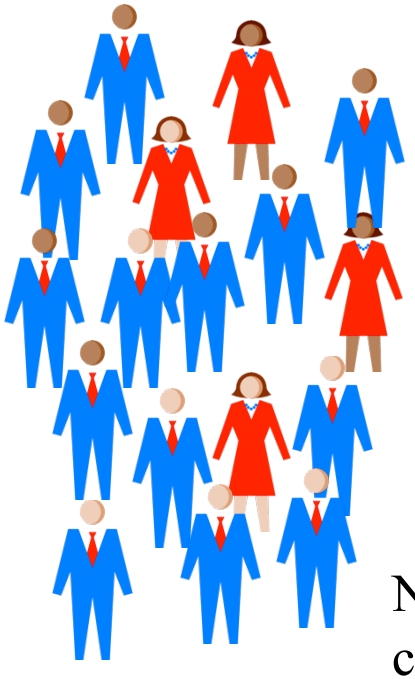
How could we use matching to estimate the effect of the treatment on the controls?

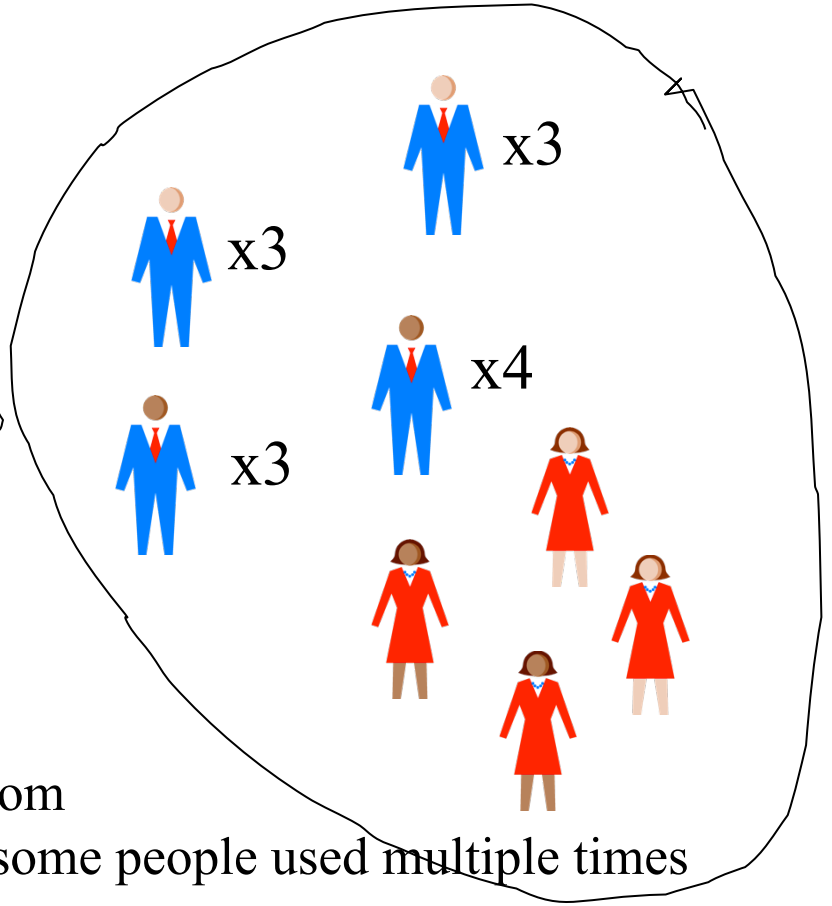
Average Treatment Effect on Control

Does not
get married

"Does not get married"
is inferential group

Gets
married





No one discarded from
comparison group; some people used multiple times

!!Pop Quiz!!

(1)When might we care about ATT more than ATC? Give a real-world example.

(2) When might we care about ATC more than ATT? Give a real-world example.

(3) When might we care about ATE? Give a real-world example.

(4) Under what conditions would ATT = ATC = ATE?

Hypothetical Example

Person	Treat	Educ	Age	Y0	Y1	Y
1	1	0	22	4	10	10
2	1	0	25	6	12	12
3	1	1	21	8	12	12
4	1	1	19	8	12	12
5	1	1	30	12	12	16
6	1	1	26	10	16	14
7	0	0	21	4	14	4
8	0	0	20	4	10	4
9	0	0	19	4	10	4
10	0	0	22	4	10	4
11	0	0	26	6	12	6
12	0	0	20	6	12	6
13	0	1	28	8	12	8
14	0	1	21	8	12	8

What is the effect of the treatment on the treated?
What is the effect of the treatment on the controls?

Propensity Score Theory

What is a propensity score?

It is a one-number summary of the covariates!

What is a propensity score?

- In an ideal world, we would want to control for all confounding covariates, in the vector **X**.
- However, usually these are not all observable.
- Propensity score theory says that rather than controlling for (stratifying on, regressing on, or matching on) all the variables in **X**, it is sufficient to control for just the propensity score, e(**X**).
- Think of e(**X**) as *just a one-number summary* of all the confounding covariates of **X**.

How a propensity score can be useful?

- Definition of a propensity score, e(**X**)
- $e(\mathbf{X})=Pr(Z=1|\ \mathbf{X})=E[Z\ |\ \mathbf{X}]$
- We can estimate e(**X**) with standard software using e.g. logistic or probit regression (there are other possibilities).
- Therefore, **IF IGNORABILITY HOLDS** (there are no other confounding variables) and if we "**properly control**" for our estimate of e(**X**) we should be able to get unbiased treatment effect estimates.

Recall the relationship between ignorability and bias

- Say we want to estimate the effect on the treated,

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

- Even though we have an observational study, if **ignorability holds** (i.e. the covariates in X are the only confounding covariates) and we stratify on all the covariates **X**

$$E[Y(0) \mid Z=1, \mathbf{X}] = E[Y(0) \mid Z=0, \mathbf{X}]$$

- We can unbiasedly estimate

$E[Y(1) \mid Z=1, \mathbf{X}]$ with $\bar{Y}_{Z=1,X}$

and $E[Y(0) \mid Z=1, \mathbf{X}]$ with $\bar{Y}_{Z=0,X}$

and then $E[Y(1)-Y(0) \mid Z=1] = E_x[\, E[\, Y(1)-Y(0) \mid Z=1, \mathbf{X}]]$

- But if X includes many covariates this can be difficult!!

!! Pop Quiz!!

How do we define ignorability?

Implications of propensity score for bias

- We can use PS to match (where matching is one of many strategies to control for confounding variables)
- We want to estimate the effect on the treated,

  $E[Y(1) - Y(0) \mid Z=1] = E[Y(1) \mid Z=1] - E[Y(0) \mid Z=1]$
- If ignorability holds (i.e. the covariates in X are the only confounding covariates) and we match on the propensity score $E[Y(0) \mid Z=1, e(\mathbf{X})] = E[Y(0) \mid Z=0, e(\mathbf{X})]$
- So we can unbiasedly estimate

  $E[Y(1) \mid Z=1, e(\mathbf{X})]$ with $\bar{Y}_{Z=1,e(X)}$ and

  $E[Y(0) \mid Z=1, e(\mathbf{X})]$ with $\bar{Y}_{Z=0,e(X)}$

This just means that we use pscore-adjusted groups to estimate each mean

How might one adjust for the propensity score?

- Matching
- Weighting
- Subclassification
- ~~Regression~~

Benefits/costs of using propensity scores

Benefits

- Adjusting for one variable is (superficially) easier than adjusting for many
- Allows for non-parametric or semi-parametric to causal inference so it's more robust; said another way it allows us to have less strict parametric assumptions
- Diagnostics can allow the researcher to address problems with overlap
- Can be incorporated into a doubly-robust strategy for causal inference (discussed more next week)

Costs

- Requires a decent estimate of the propensity score
- Can be a hassle/time-consuming
- No clear guidelines for success

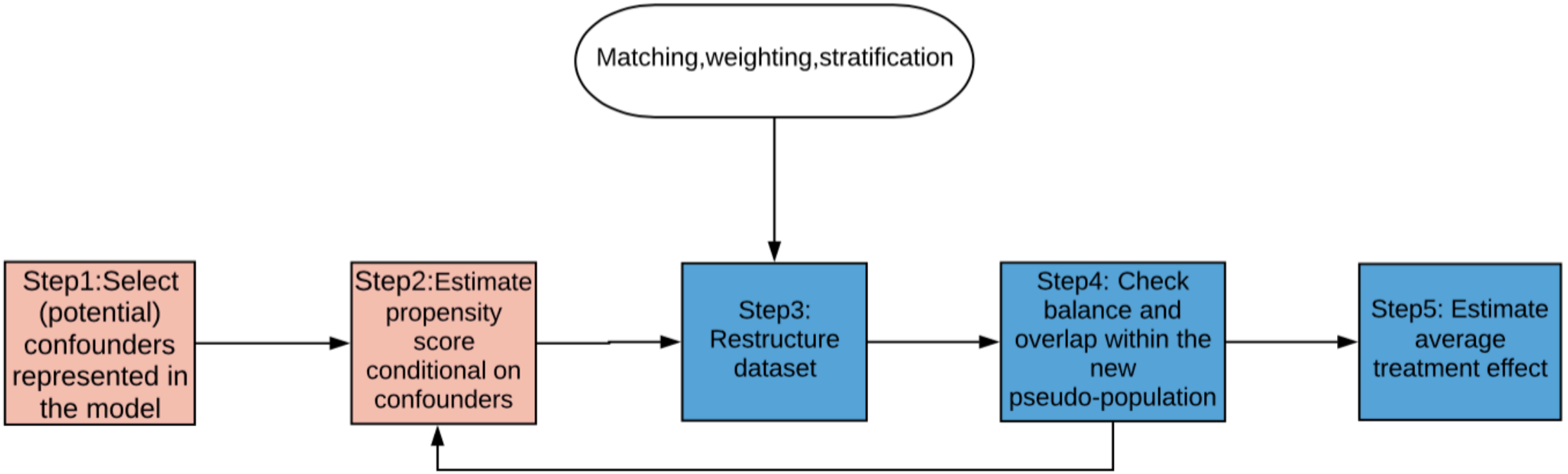
Steps to consider when using propensity score to estimate a causal effect

Five Major Steps

1. Select (potential) confounders represented in the model
2. Estimate propensity score conditional on confounders
3. Restructure dataset based on propensity score. The goal is to create a pseudo population where treated and control groups look very similar
4. Check balance within the new pseudo-population and assess degree of overlap (common support)*
   [Here possibly return to Step 2**]
5. Estimate average treatment effect (ATE, ATT, ATC) on the restructured data

* Some might argue this should happen between 2 and 3

** Then use a different pscore model or restrict sample to those that satisfy overlap (common support)



Red indicates the original data set (unless we restrict to observations with overlap)

Blue indicate the reconstructed data set

Propensity Score Matching
(National Supported Work Example)

Example: Key Features of the Study/Data

As we go through these five principal steps, let's consider an example:

Treatment group

- From National Supported Work (NSW)
- NSW provided job training to disadvantaged men
- Outcome is earnings in 1978 (re78t) in thousands of $

Comparison group

- Two comparisons group pulled from survey data:
  - A **sample** of 2,490 men from the PSID
  - A **sample** of 15,992 from the CPS

!!!Pop Quiz!!!

1. What's the definition of a confounder?

2. What confounders might exist for this study?

3. What causal assumption would excluding confounding variables violate?

Example: Covariates from the Study/Data

Covariates available are:

earnings (in thousands of $) in 1974 and 1975 (re74t, re75t)

age

years of education

marital status

race/ethnicity

Step 1: Determine confounding covariates

- What variables are you worried about that may be predicting both your treatment and your outcome? (confounder)
- Importantly, (as in any causal analysis) you cannot "control for" any variables that are observed *post-treatment* because technically these are outcomes.
- These confounding covariates are the ones you care about balancing across groups.
- In our example they have been defined as: earnings in 1974 and 1975, age, years of education, marital status, and race/ethnicity.

Determining confounding covariates, more generally

- How do we determine confounding covariates in other contexts
- General rule for small to medium # covariates (not universally agreed upon) is to use all of them if you can. Caveats:
  - This could lead to overfit. Overfit is bad. (why?)
  - There are some variables that could be problematic to condition on because if ignorability is not satisfied they can amplify our bias. These bias amplifiers have the general form that they strongly predict the treatment but not the outcome.

Determining confounding covariates, more generally

- What if you have a very large number of covariates available?
- Possible strategies (not mutually exclusive)
  - Theory
  - Variable selection (stepwise, lasso, horseshoe)
  - Method that uses regularization
- Last two can be problematic if
    - only used for treatment assignment model
    - assume a linear model
    - (also beware regularization-induced confounding)

Stepwise and lasso are not recommended

Step 2: Estimate propensity scores

The propensity score is defined as: Pr[Z=1 | X]

Since Z is binary Pr[Z=1 | X] = E[Z | X]

E[Z | X] is just a **regression with a binary dependent variable.**

!!! Pop Quiz!!!

What are models or fitting algorithms that we can use to estimate E[Z | X]?

What might be the problems when you model the propensity score?

[Side note: might be a interesting topic for final project!!]

Step 2: Estimate propensity scores cont'd

There are many methods to estimate propensity scores

- Most common traditionally:
  - logistic regression
  - probit regression

Increasing in popularity are more flexible models

- Generalized additive models(GAM), SVM (support vector machines), random forest, BART, neural net, etc.
- Crucially important not to overfit though.

(What happens if we overfit the pscore model....?)

Stata commands

```
. psmatch2 treat age educ black hisp married re74t re75t
```

- Calculates propensity scores using the first variable as treatment variable and rest as confounding covariates (default is probit regression)
- Uses propensity scores to match “control” units (those for whom treat==0) to “treatment” units (those for whom treat==1) (default is matching with replacement)

Stata

```
. psmatch2 treat age educ black hisp married re74t re75t

Probit regression                                 Number of obs   =      16177
                                                  LR chi2(7)      =     993.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -514.17625                       Pseudo R2       =     0.4915

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.007333   .0048824    -1.50   0.133    -.0169023    .0022363
        educ |  -.0470966   .0157949    -2.98   0.003    -.0780541   -.0161392
       black |   1.888368   .1015384    18.60   0.000     1.689357     2.08738
        hisp |   .7103781   .1573136     4.52   0.000     .4020491    1.018707
     married |  -.4613337   .1095799    -4.21   0.000    -.6761064    -.246561
       re74t |  -.0177822   .0119227    -1.49   0.136    -.0411503    .0055858
       re75t |  -.0897277   .0149838    -5.99   0.000    -.1190954     -.06036
       _cons |  -1.591653   .2343013    -6.79   0.000    -2.050875   -1.132431
------------------------------------------------------------------------------
```

possible warning messages

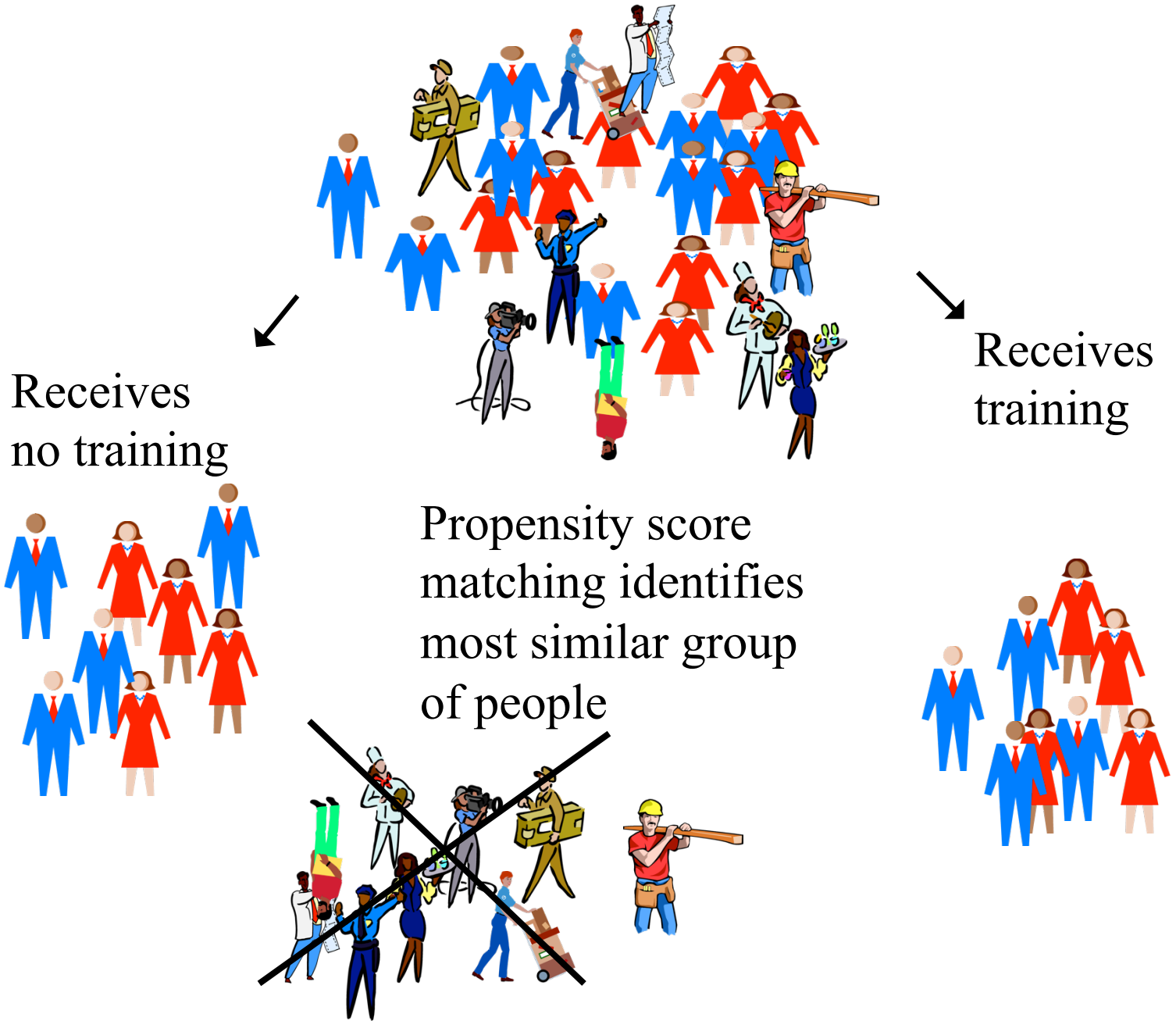
```
Note: 15 failures and 0 successes completely determined.
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
```

What does this mean??

Step 3: Restructure data based on pscore

- Once you have obtained predicted probabilities for each observation, use this information to "restructure" your data to create balanced treatment and control groups
- There are several methods to do so, including
    - **matching** (there are several methods)
    - subclassification/stratification
    - inverse probability of treatment weights

Examples of restructuring through matching


Receives
no training
Propensity score
matching identifies
most similar group
of people
Receives
training

Restructuring in practice. Key idea: Matching is a crude form of weighting

In practice, propensity score software packages perform diagnostics and analyses described in the next steps through weighting rather than creating a new restructured dataset.

Why does this make sense??

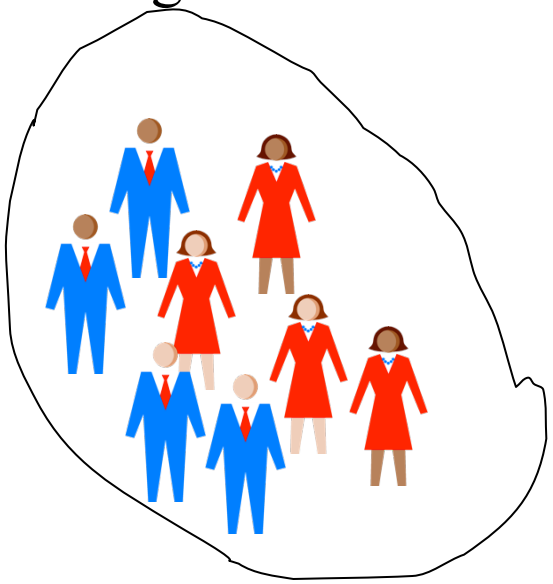
Average Treatment effect on Treated

Receives
no training

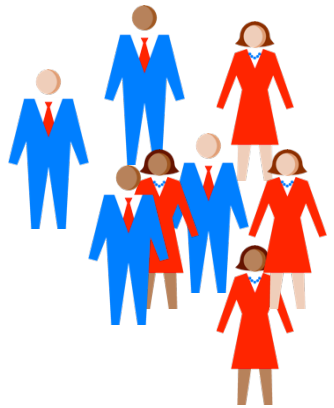
"Receives training"
is treatment group

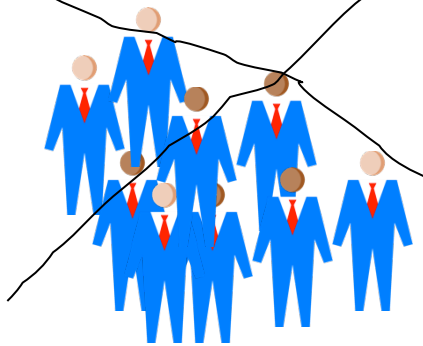
What weights would be applied here?

Receives
training



weight =1





these people discarded

weight = 0

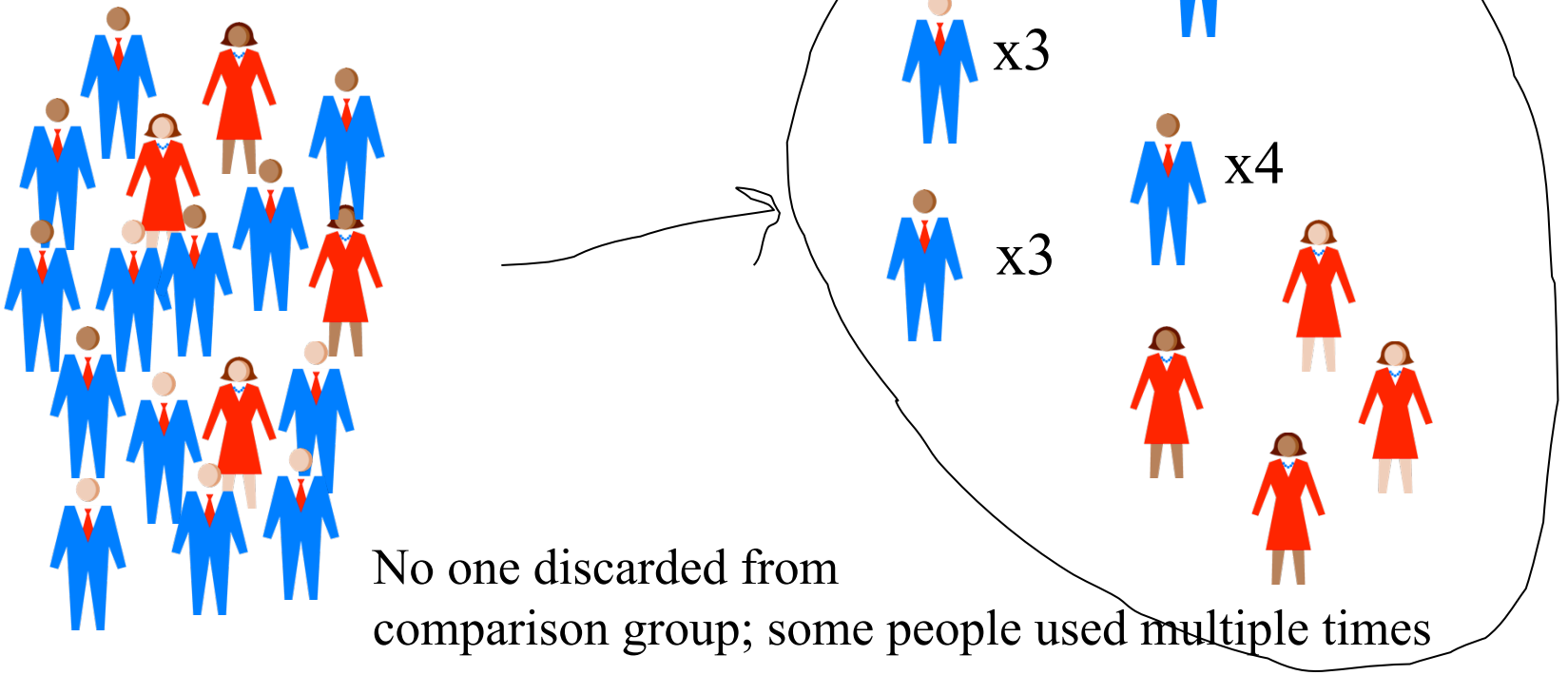
Average Treatment Effect on Control

Does not get married

"Does not get married" is inferential group

Gets married

What weights would be applied here?



No one discarded from comparison group; some people used multiple times

Step 3: Using the propensity score as a Distance Measure

The overall goal when matching is to find people who are similar to each other. How do we define similar?

Using the propensity score, the distance between unit i and unit j can be expressed as:

1. Propensity score : $D_{ij} = |e_i - e_j|$
2. Linear propensity score : $D_{ij} = |\text{logit}(e_i) - \text{logit}(e_j)|$

Examples of other distance measures:

1. Exact measure
2. Mahalanobis (multivariate distance measure)

Step 3: Matching algorithm using propensity scores

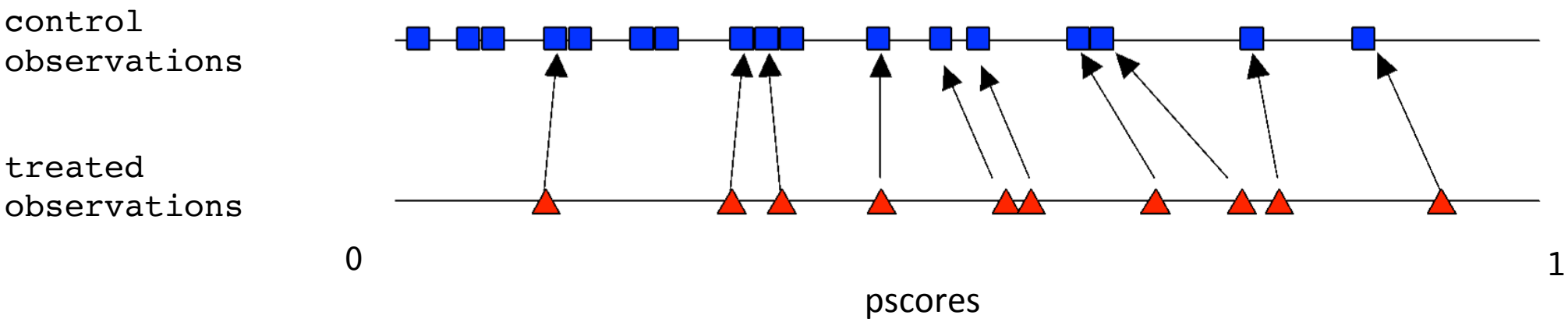
- Nearest Neighbour Matching: For each person in the treatment group, find the person in the control group with the closest propensity score -- this is the treatment group members' match. This can be done with or without replacement.
- K-1 Matching
- Caliper Matching
- Optimal Matching

(we will discuss these more at a later time)

Step 3: With or without replacement?

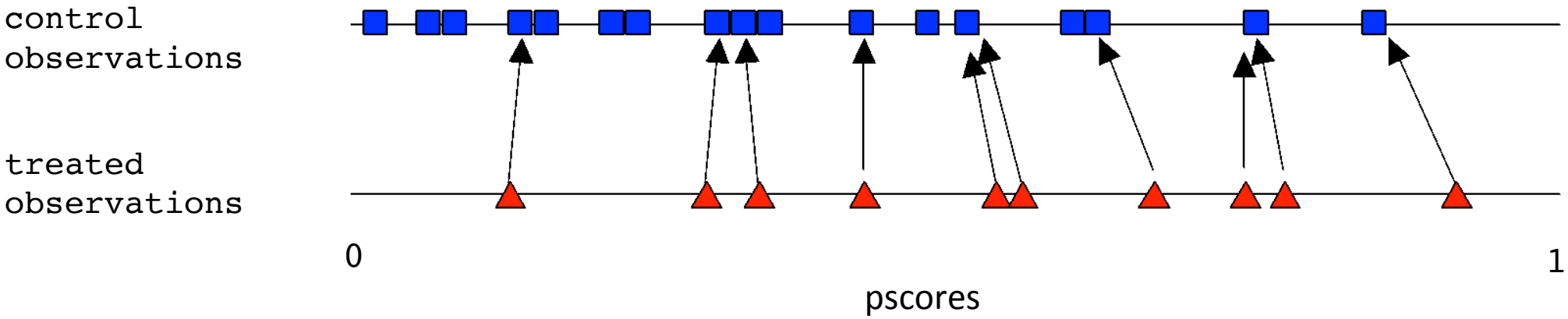
- In standard matching without replacement, this control group member is then removed from the control reservoir and cannot be chosen again.
- In matching with replacement, this control group member is used as many times as long as the it's the best match.
- Matching with replacement tends to reduce bias (though may increase s.e.'s) relative to matching without replacement.
- Also matching without replacement cannot be performed when the control group is smaller than the treatment group.

What type of matching is this?



without replacement

What type of matching is this?



with replacement

Recall how we use these data to estimate the effect of the treatment on the treated. What type of matching did we use?

Matched to	Person	Treat	Educ.	Age	Y0	Y1	Y	pscore
	1	1	1	26	10	14	14	0.68
	2	1	1	21	8	12	12	0.65
	3	1	1	30	12	16	16	0.71
	4	1	1	19	8	12	12	0.63
	5	1	0	25	6	10	10	0.33
	6	1	0	22	4	8	8	0.24
	7	0	0	21	4	8	4	0.22
	8	0	0	26	6	10	6	0.36
	9	0	0	28	8	12	8	0.43
	10	0	0	20	4	8	4	0.19
	11	0	1	26	10	14	10	0.68
	12	0	1	21	8	12	8	0.65
	13	0	0	16	2	6	2	0.12
	14	0	0	15	1	5	1	0.10

Step 4: Check overlap and balance

Step 4: Examine the "overlap" (common support?*)

- Conceptually, we want to make sure that for each treatment group member there is a control group member that is sufficiently similar that we believe they can act as an **empirical counterfactual**.
- To investigate, plot histograms for the propensity scores in each group separately (or on top of each other using different colors) and compare. This should include **all units** in the analysis sample, not just the matched sample.
- Can also create overlaid histograms (or other plots) of the confounders directly)
- Is there enough "overlap" to justify subsequent analyses? e.g. if they don't overlap at all we may surmise that the people in the two groups are just too different to be compared (apples and oranges)

* Some people call this the positivity assumption

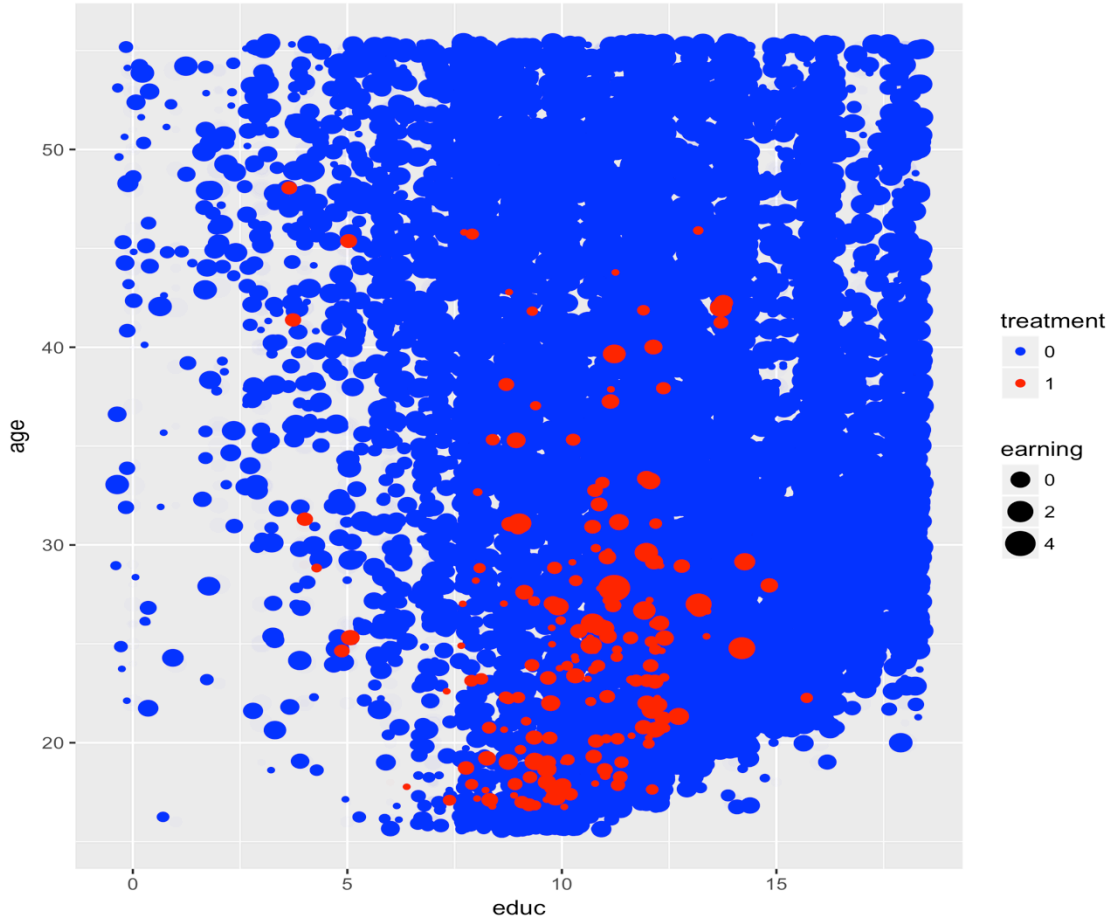
Overlap in age and education: NSW example

PSID Treated and controls **before** matching

The survey controls
look very different from the treated men.

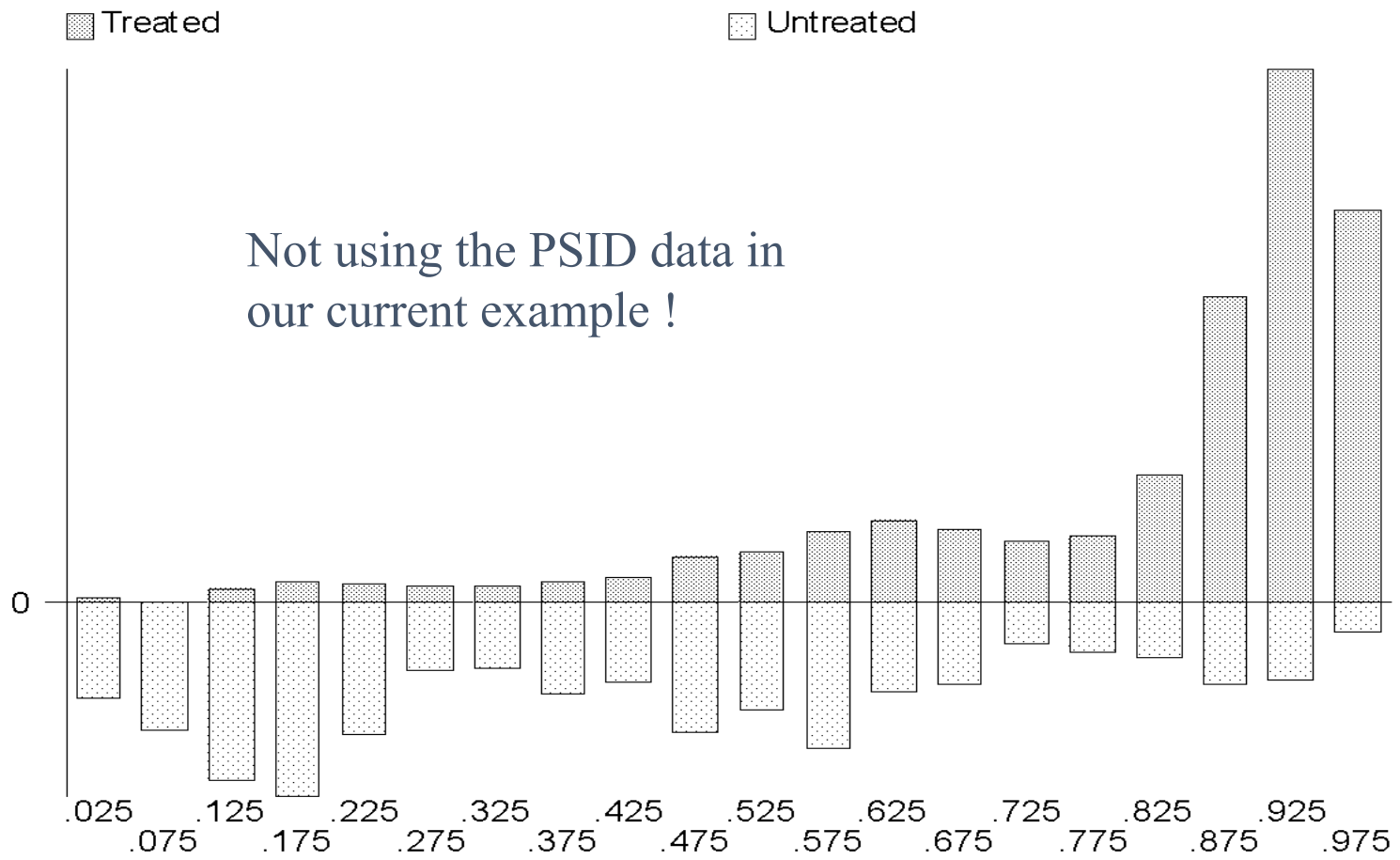
Therefore:
(1) We do not want to use the difference in sample means(-15000) as our treatment effect estimate.

(2) A standard model fit to these data will rely on a good deal of model Extrapolation.



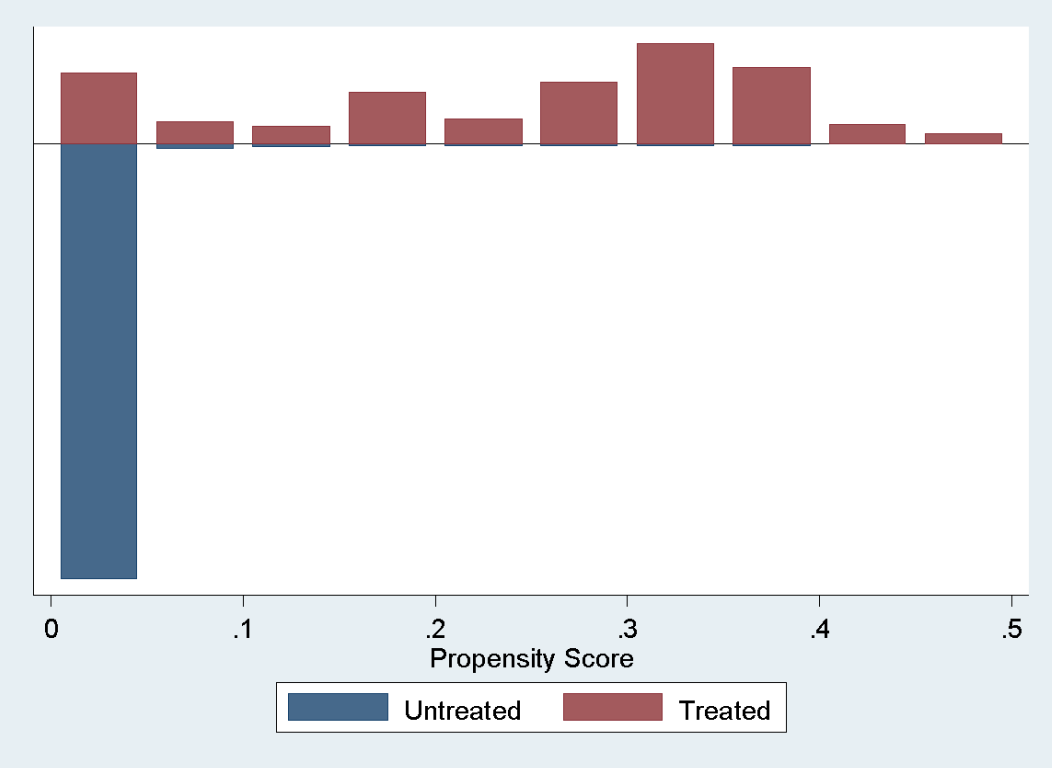
Education vs age (jittered).
Blue: T=0 (controls)
Red: T=1 (treated)
size of dot reflects magnitude of earnings

Now in Stata: Example of a pscore overlap picture in Stata produced by "psgraph"



This is just an example, not our data...

Now our data…….

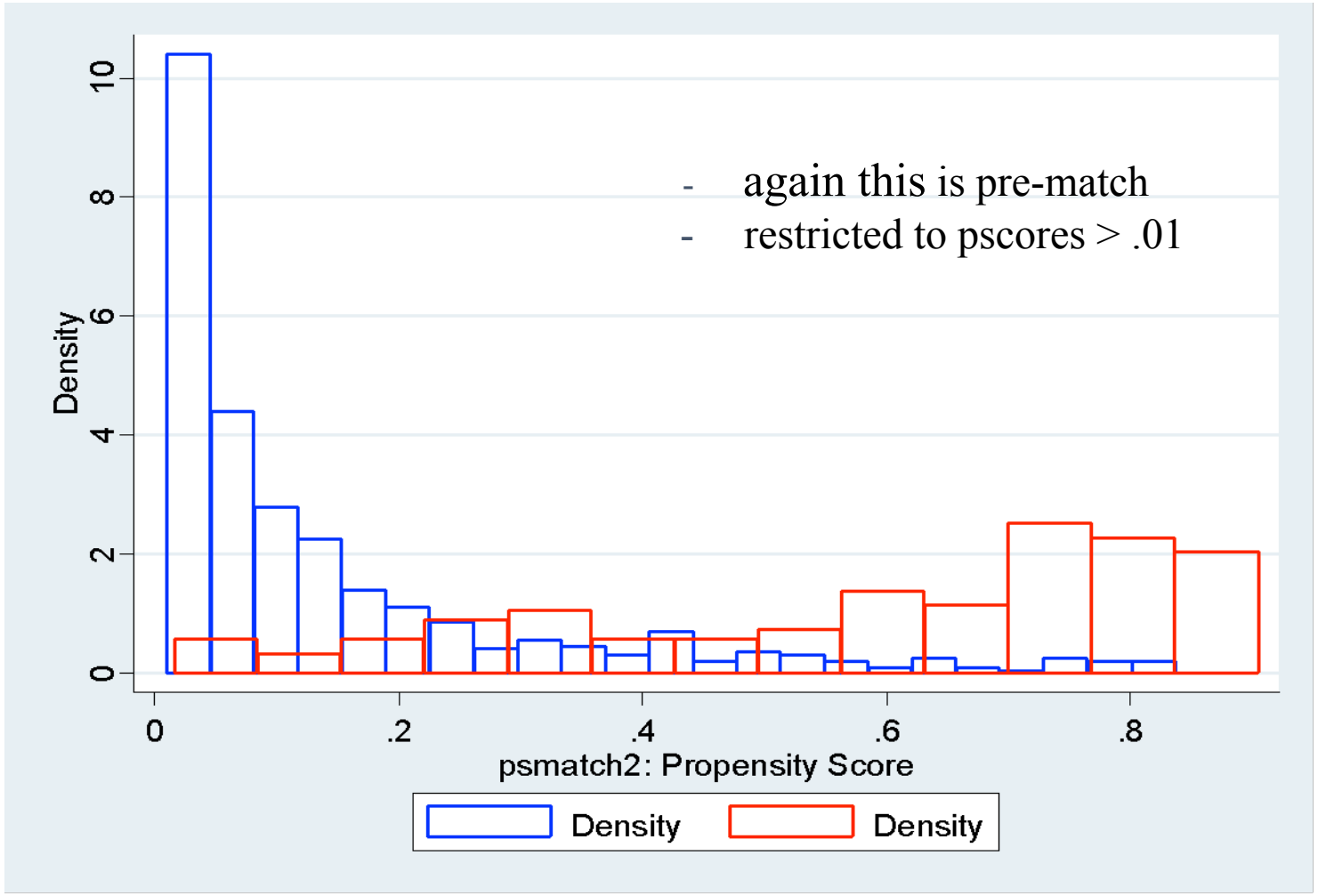


What's going wrong there?

Is there sufficient overlap??

These are the distributions of propensity scores for each group BEFORE MATCHING

My solution:
Check overlap by region to minimize distortions



Overlap

- How do we determine if there is sufficient overlap (common support)?
  - A standard rule is to see if there are units in the inferential group with propensity scores outside the bounds of the comparison group.
  - For instance if the goal is to estimate the ATT (SATT or PATT), then we might exclude from the analysis treated observations with propensity scores higher than the highest control group propensity score.
- What do we do if there isn't sufficient overlap?
  - If you exclude observations from the inferential group then you should
    - "profile" them (e.g. create a table/figure of descriptive statistics); who are these units that lack empirical counterfactuals?
    - describe the new analytic sample (who are we now making inferences about?)

!!! Pop Quiz !!!

- Can we using matching to create overlap?

No, because if the overlapped subjects do not exist, new subjected cannot be magically created.

Inference can only be make within the overlap range of P-scores, assuming the pscores are good estimates.

Balance : what does it mean?

- We want to create a restructured dataset that looks like the output of a randomized experiment.
- What we really want is for the multivariate distribution of the covariates to be same across groups.
- We can only check this for the observed covariates.
- (Recall conceptually that the *real* goal is to make the distributions of Y(0) and Y(1) to be equivalent across groups. Is balance sufficient for this???)

!!POP Quiz!!

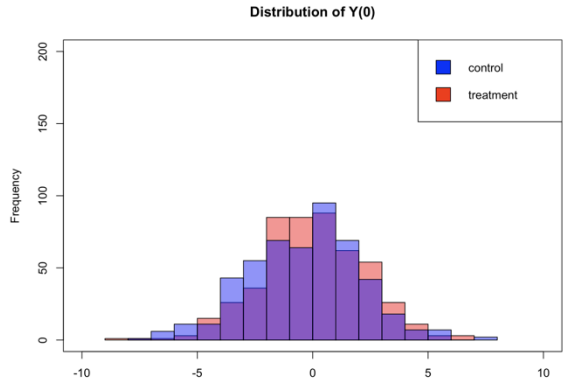
1. What does balance mean?

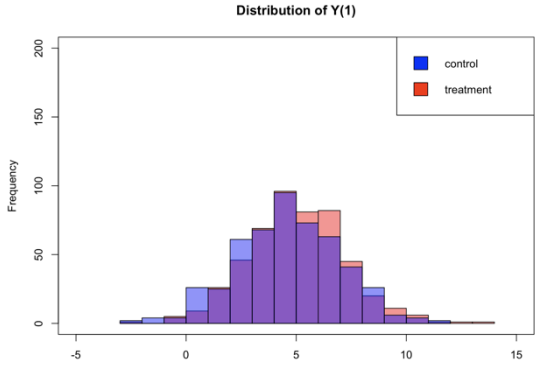
2. How is balance different from overlap?

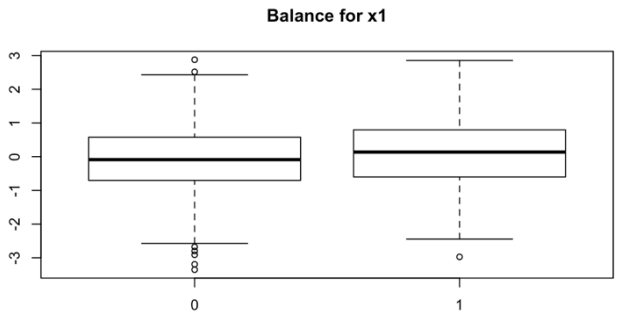
3. How is balance related to ignorability?

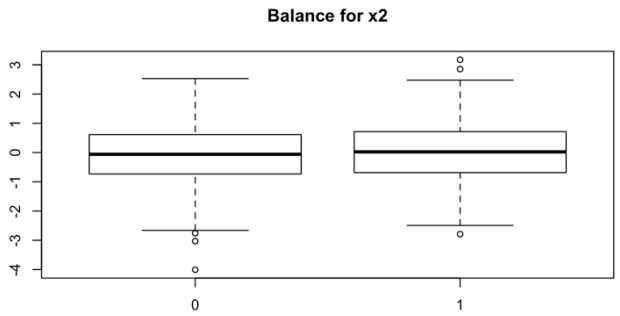
4. Is checking balance for observed covariates sufficient for satisfying ignorability?

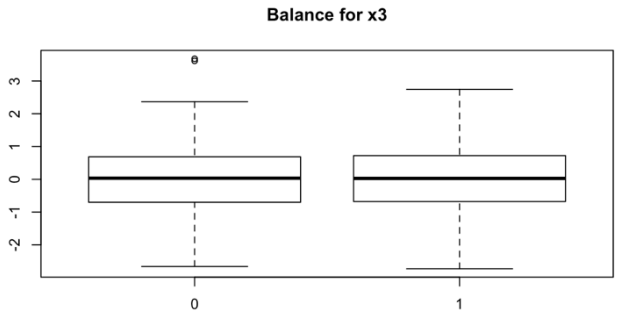
Balance: randomized experiment

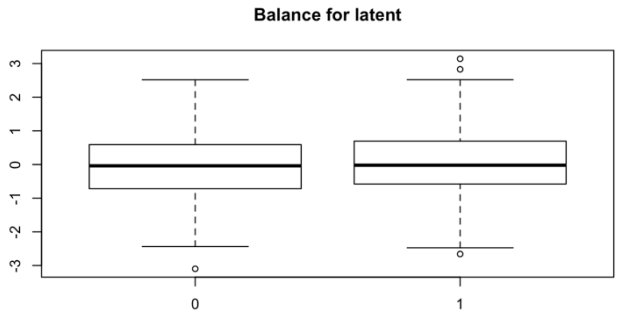












Note that in this simple simulation the randomization achieves balance in all the covariates, not just the observed covariates X1-X3

DGP:

```
x1<-rnorm(1000,0,1)
x2<-rnorm(1000,0,1)
x3<-rnorm(1000,0,1)
latent<-rnorm(1000,0,1)
Y1<-5+x1+x2+x3+latent+rnorm(100
0)
Y0<-x1+x2+x3+latent+rnorm(1000)
Z<-sample(rep(c(0,1),each=500),
1000,replace=FALSE)
Y<-Y1*Z+(1-Z)*Y0
```

Examine balance

- Compare pre-matching balance to post-matching balance
- What does balance mean?
  - It means how close are the confounding covariates' **distributions** between the treatment and control
  - We typically approximate this by looking at moments of those distributions (see below)
- How do we examine whether the sample is balanced?
  - Usually the standardized mean difference (SMD) and variance ratio (VR) are used for continuous variables
  - Difference in percentages make more sense for binary variables (or each level of a categorical variable)
  - In addition: difference in correlations, difference in squared terms or interactions, qqplot for empirical distributions variance ratio

Do not use the psmatch2 default method for examining balance!

```
pstest age educ black hisp married re74t re75t, both
-------------------------------------------------------------------------------
                 Unmatched |       Mean                      %reduct |     t-test
    Variable       Matched | Treated Control    %bias   |bias| |    t    p>|t|
-------------------------+----------------------------------+----------------
           age   Unmatched | 25.816   33.225    -79.6           |  -9.10  0.000
                  Matched  | 25.816   26.065     -2.7    96.6   |  -0.26  0.795
                           |                                    |
          educ   Unmatched | 10.346   12.028    -67.9           |  -7.94  0.000
                  Matched  | 10.346   10.232      4.6    93.2   |   0.43  0.664
                           |                                    |
         black   Unmatched | .84324   .07354    242.8           |  39.66  0.000
                  Matched  | .84324   .84865     -1.7    99.3   |  -0.14  0.886
                           |                                    |
          hisp   Unmatched | .05946   .07204     -5.1           |  -0.66  0.510
                  Matched  | .05946   .02703     13.1  -157.9   |   1.53  0.126
                           |                                    |
       married   Unmatched | .18919   .71173   -123.3           | -15.62  0.000
                  Matched  | .18919   .18919      0.0   100.0   |  -0.00  1.000
                           |                                    |
         re74t   Unmatched | 2.0956   14.017   -156.9           | -16.92  0.000
                  Matched  | 2.0956   2.1406     -0.6    99.6   |  -0.10  0.918
                           |                                    |
         re75t   Unmatched | 1.5321   13.651   -174.6           | -17.77  0.000
                  Matched  | 1.5321   1.4183      1.6    99.1   |   0.38  0.701
                           |                                    |
-------------------------------------------------------------------------------
```

Outstanding issues

- Should you use t-statistics (or other significance tests) to determine whether imbalance is problematic?
- How determine how to make trade-offs particularly when there are many covariates?
- Is it sufficient just to look at difference in means?
- How close is close enough?

Balance Diagnostics – more specific advice

- Create a list of which variables are most important to balance based on how predictive each one is of the outcome (in theory)
- Examine the balance in confounding covariates between treatment groups and compare this to the balance that existed before matching. It's hard to know what is "close enough" so the best advise is to get as close as you possibly can (some newer matching algorithms are able to minimize distance once you provide the metric for measuring imbalance)

How to gauge balance?

- For each continuous variable
  - look at standardized differences in means (difference in means divided by the standard deviation in the inferential group)
  - can also compare variances (or difference in squared terms)
  - and correlations (or differences in interaction terms)
- For each binary variable
  - look at the difference in means (percentage)

Examining balance: Stata (option 2)

USE THIS COMMAND for Stata homework (I provide the ado file)

```
. psbal2 age educ black hisp married re74t re75t
```

Variable	Sample	Mean Treated	Mean Control	SD Treated	SD Control	STD Diff	Ratio of SDs
age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	26.065	7.16	10.86	-0.035	0.66
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.232	2.01	2.93	0.056	0.69
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.849	0.36	0.36	-0.015	1.01
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.027	0.24	0.16	0.137	1.46
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.189	0.39	0.39	0.000	1.00
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.141	4.89	3.37	-0.009	1.45
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.418	3.22	2.41	0.035	1.34

Why isn’t balance in means sufficient?

For different scales, mean difference might not be a good measure; standard difference is a better measure.

Why might balance in means not be sufficient?
Why do we check other "moments" of the distribution?

-Suppose the “true” model looks like

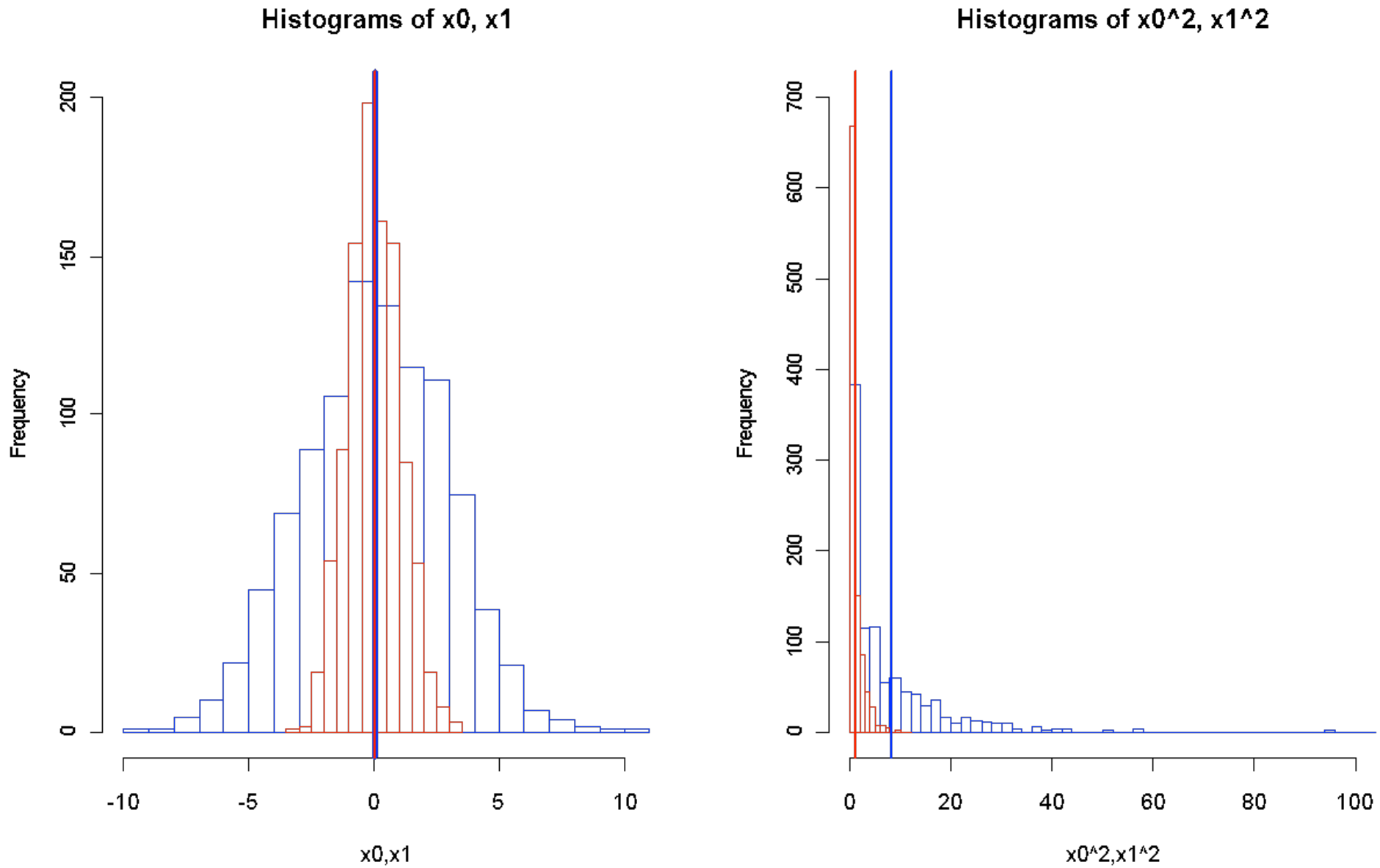
$y \mid z{=}0 = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$

$y \mid z{=}1 = \alpha + \beta_1 x + \beta_2 x^2 + \tau + \varepsilon$

- averaging over each treatment group separately and examining the difference in these means (which would be our estimator for τ if we weren’t fitting a model), we see that this equals

$$\hat{\tau} = \bar{y}_1 - \bar{y}_0 = \beta_1(\bar{x}_1 - \bar{x}_0) + \beta_2(\overline{x_1^2} - \overline{x_0^2}) + \tau$$

- balance in means for x will get rid of part of the bias, but if the variance isn’t equal across groups bias will remain from the last term


Histograms of x0, x1
Frequency
0
50
100
150
200
-10
-5
0
5
10
x0,x1
Histograms of x0^2, x1^2
Frequency
0
100
200
300
400
500
600
700
0
20
40
60
80
100
x0^2,x1^2

Examine balance on **all** confounders chosen in Step 1

Important: We always test balance on all covariates initially designated as confounders even if they aren't in the final estimation model !!!!!

!! POP Quiz!!

- Suppose SMD and Variance Ratio both indicate matched sample are balanced. Could we confirmed our matching is successful? Why or why not?

- Should we check confounders even if they aren’t in the final estimation model? Why?

What if we aren’t happy with our balance?

Return to Step 2!

Back to step 2: Re-estimate the propensity score to try for better balance!

- It's always a good idea to try several different model specifications (for estimating propensity scores) at this stage and to compare the balance achieved under each
- Some suggestions for alternate specifications
  - Add in interactions or squared terms (e.g. if including one variable seems to worsen another's balance try including their interaction)
  - Transform variables (e.g. log of earnings)
  - Remove variables that you think are less important with regard to the outcome variable

Trying for better balance…

- Notice the imbalance in the Hispanic indicator variable. We might try interacting it with other important covariates…

Example of fitting new model to achieve better balance

- noticing the imbalance in the sd's of the pre-treatment earnings variables; it's likely there aren't modeled correctly; perhaps we should log or take the square root of these variables; (can't log 0 values so will have to add a smidge to make work)
- first try with both variables

```
. gen re74tL = log(re74t+.001)
. gen re75tL = log(re75t+.001)

. psmatch2 treat age educ black hisp re74t re75t re74tL re75tL
```

After adding both terms

```
psbal2 age educ black hisp married re74t re75t
```

Variable	Sample	Mean Treated	Mean Control	SD Treated	SD Control	STD Diff	Ratio of SDs
age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	25.768	7.16	11.43	0.007	0.63
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.059	2.01	3.21	0.142	0.63
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.859	0.36	0.35	-0.044	1.04
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.049	0.24	0.22	0.046	1.10
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.146	0.39	0.35	0.110	1.11
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.013	4.89	4.83	0.017	1.01
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	2.030	3.22	4.26	-0.155	0.76

Try removing the log term for re75…

```
. psmatch2 treat age educ black hisp re74t re75t re74tL
. psbal2 age educ black hisp re74t re75t
```

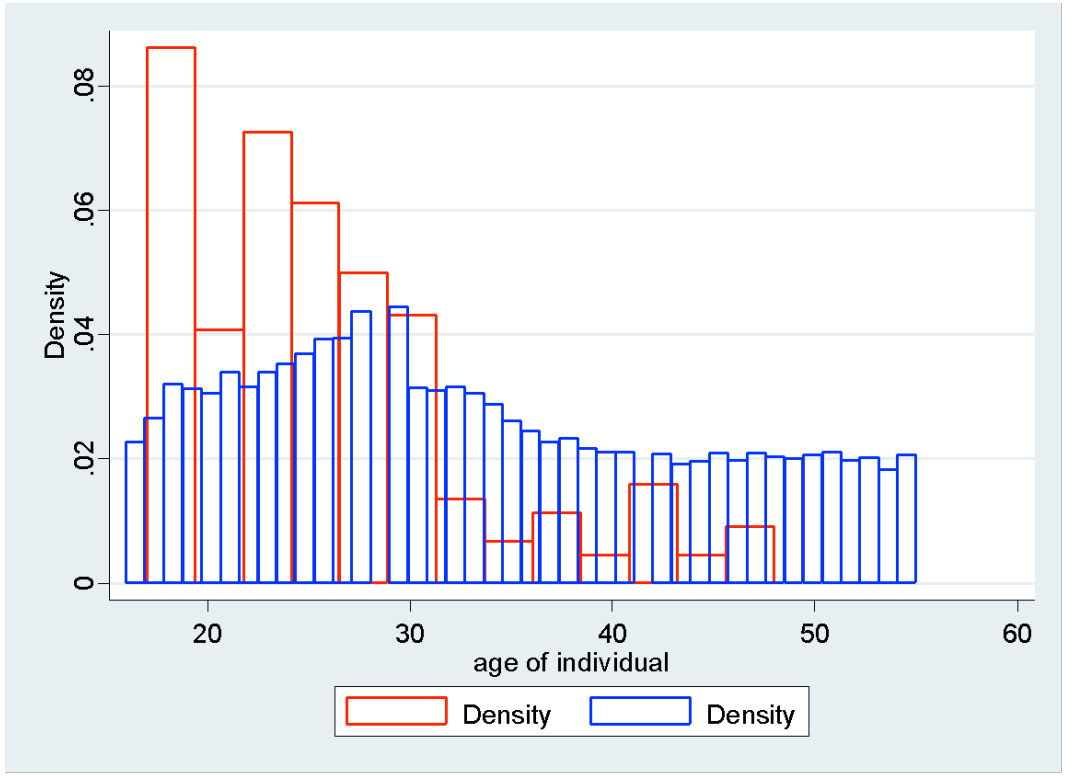
Variable	Sample	Mean Treated	Mean Control	SD Treated	SD Control	STD Diff	Ratio of SDs
age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	24.092	7.16	7.58	0.241	0.94
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.800	2.01	2.30	-0.226	0.87
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.859	0.36	0.35	-0.044	1.04
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.038	0.24	0.19	0.091	1.24
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.330	0.39	0.47	-0.358	0.83
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.503	4.89	5.46	-0.083	0.90
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.674	3.22	3.78	-0.044	0.85

```
xi: psmatch2 treat age educ black hisp re74t re75t re74tL i.hisp*age
i.married*re75t
```

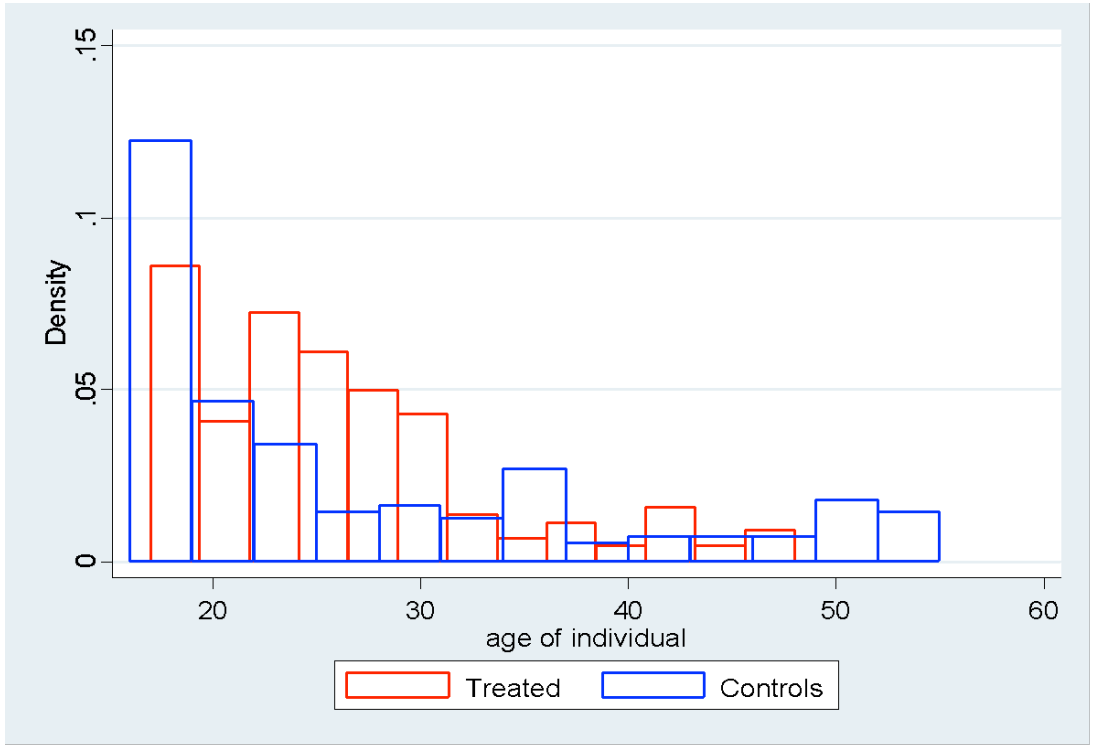
Variable	Sample	Mean Treated	Mean Control	SD Treated	SD Control	STD Diff	Ratio of SDs
age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	26.514	7.16	11.62	-0.097	0.62
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.259	2.01	2.96	0.043	0.68
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.854	0.36	0.35	-0.030	1.03
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.038	0.24	0.19	0.091	1.24
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.184	0.39	0.39	0.014	1.01
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.021	4.89	5.02	0.015	0.97
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.474	3.22	3.48	0.018	0.92

Balance Age

pre-match

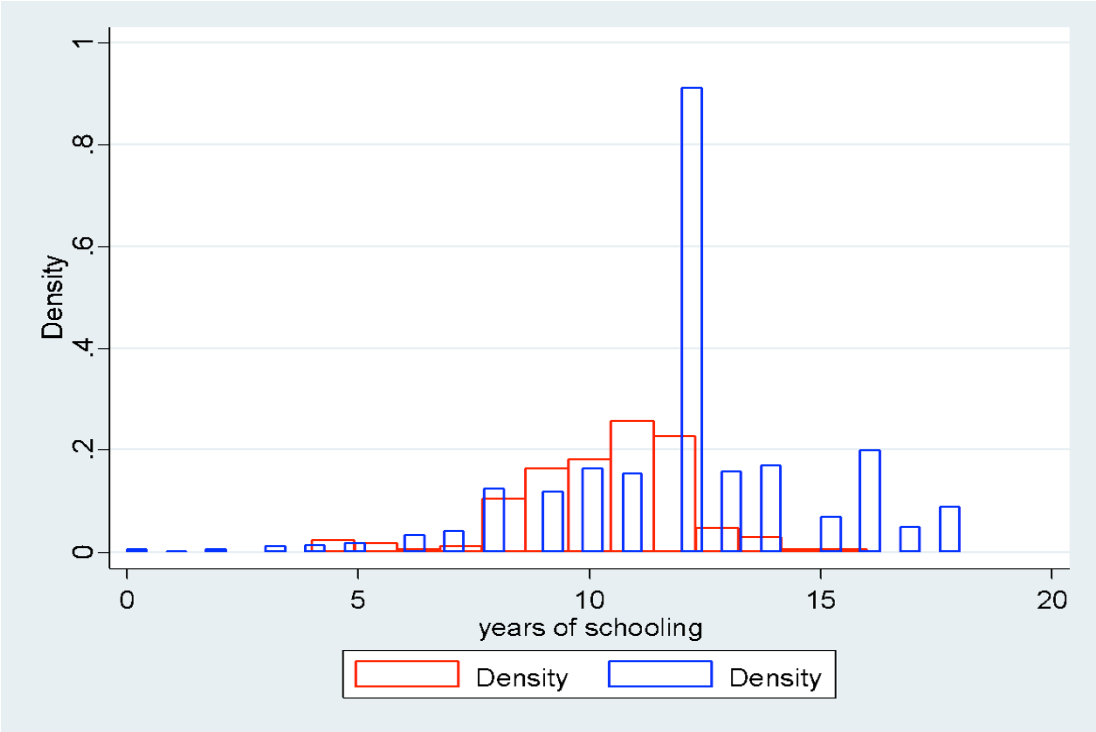


post-match

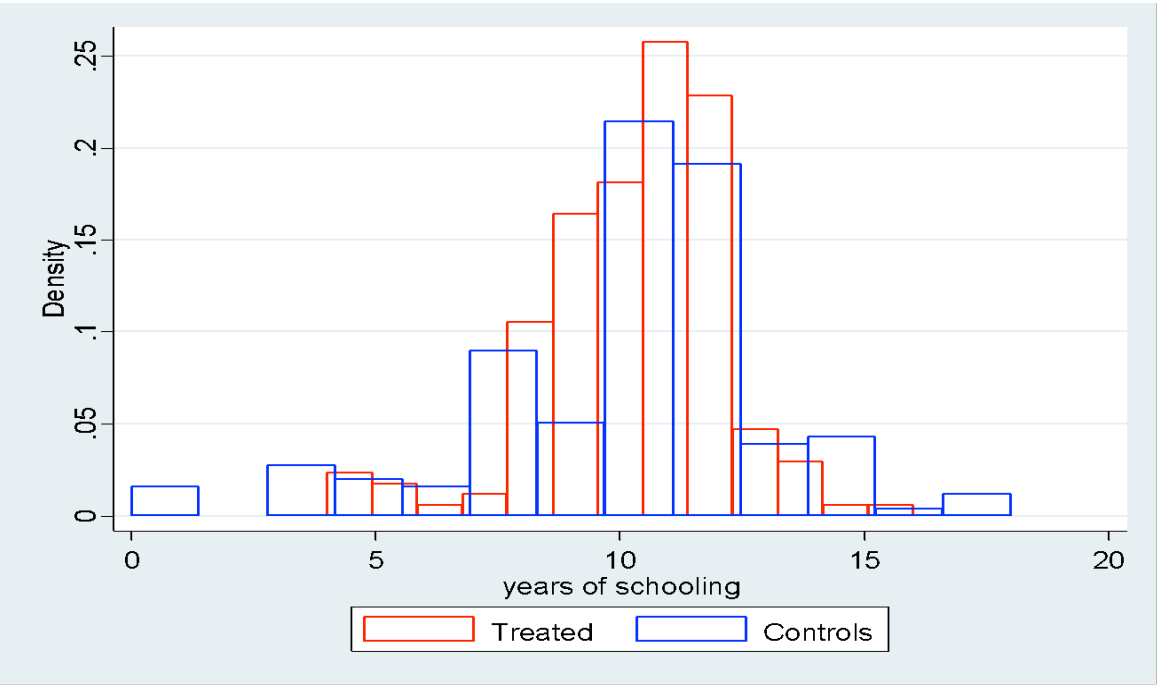


Balance Education

pre-match

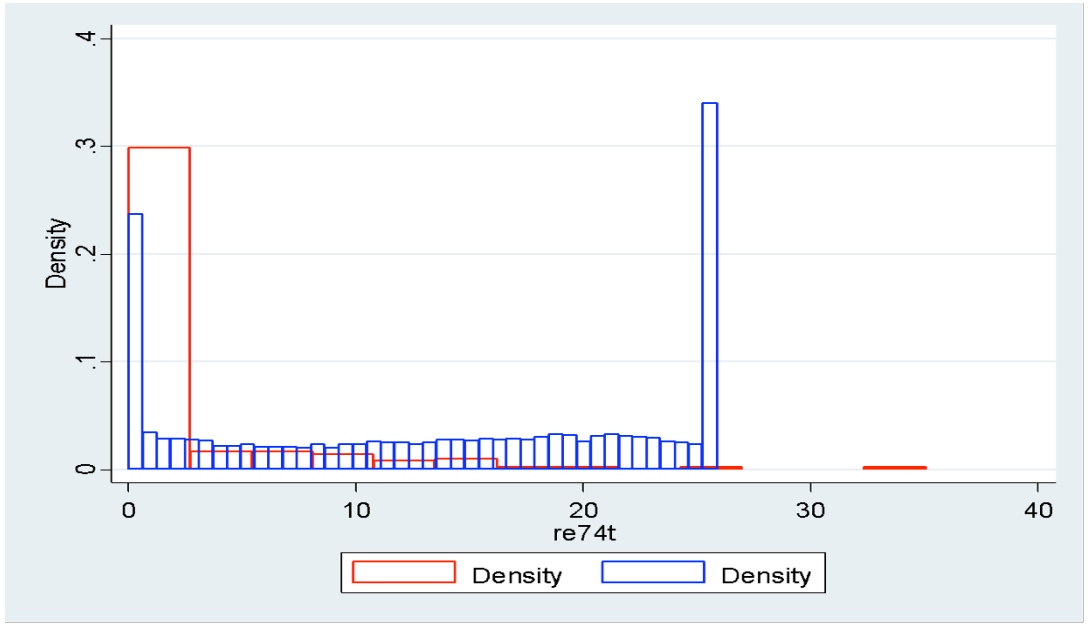


post-match

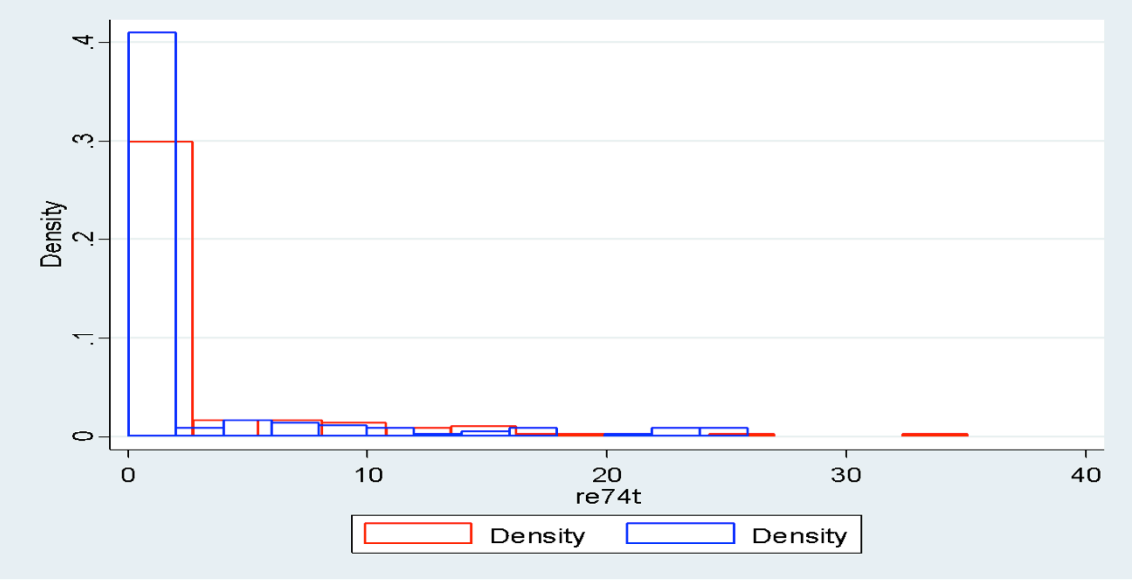


So kill these indicators and go back to earlier fit: look at some more plots

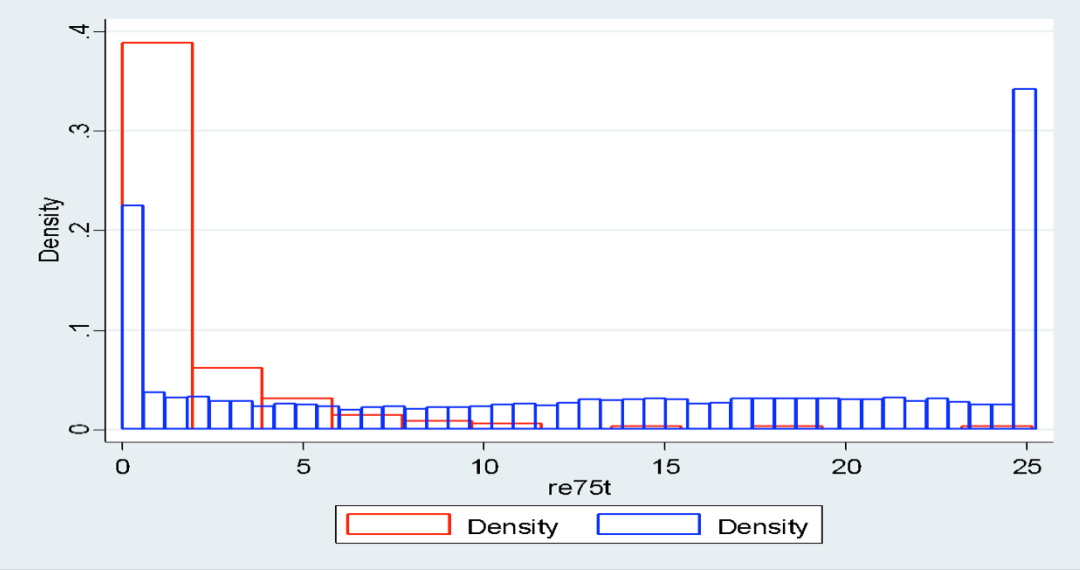
pre-match



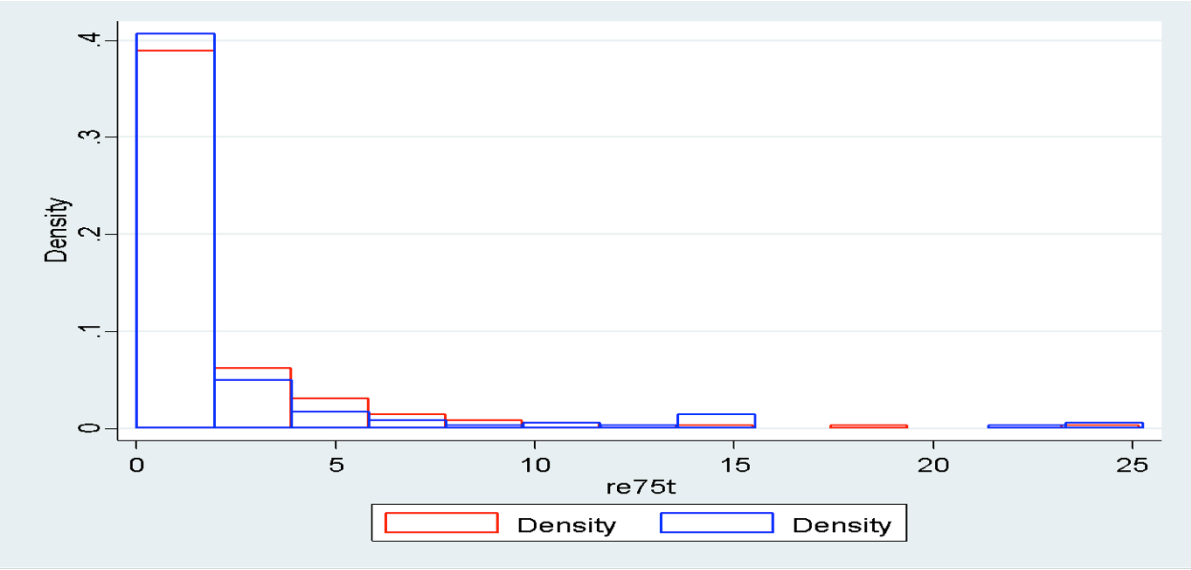
post-match



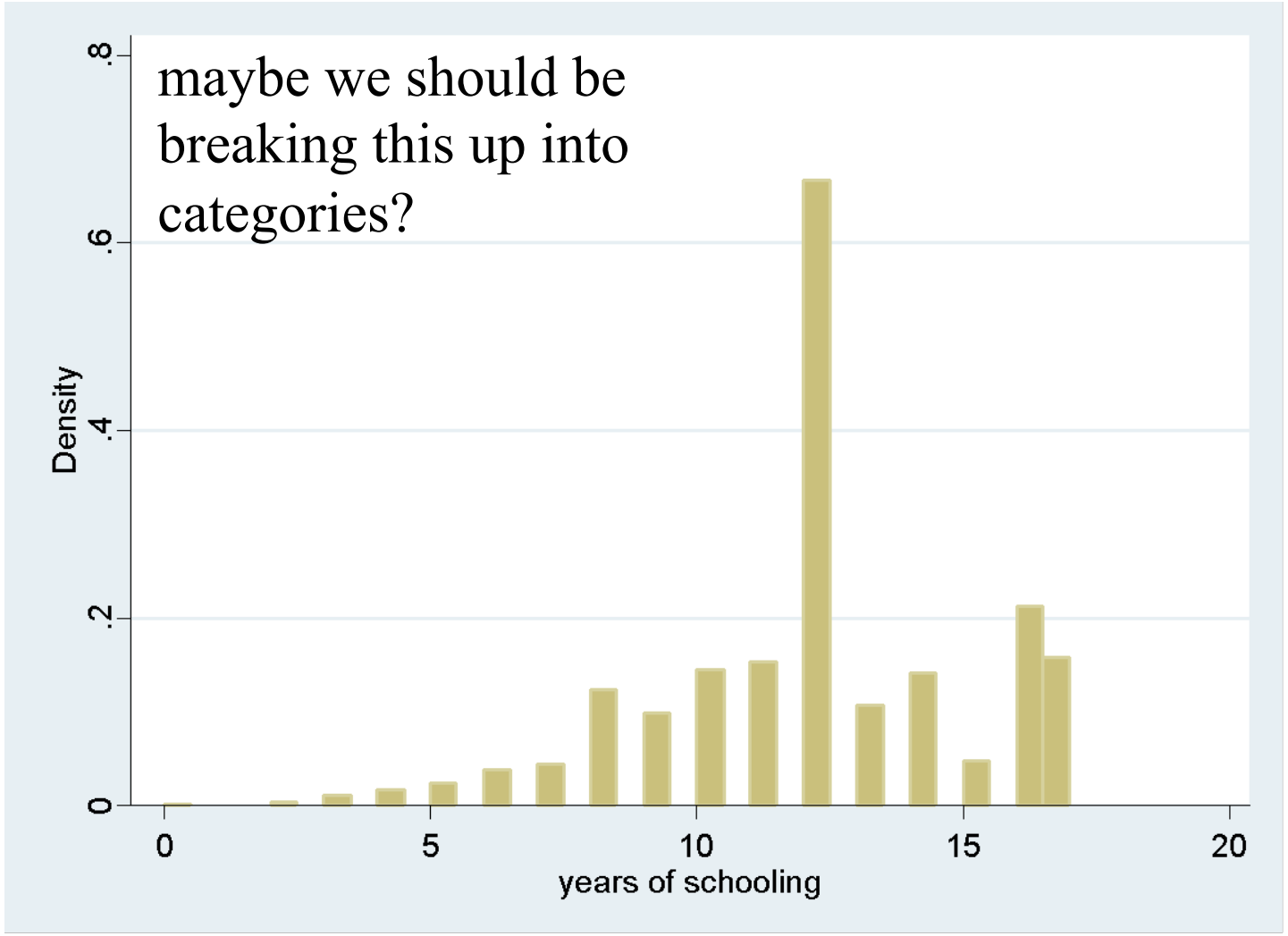
pre-match



post-match



Education Variable



new education categories

```
. hist educ
(bin=42, start=0, width=.42857143)

. gen educ_cat = 1

. replace educ_cat = 2 if educ==12
(6291 real changes made)

. replace educ_cat = 3 if educ>12
(5024 real changes made)
```

Turns out adding this hurt the balance….. oh well!

Stata code for pictures on previous slide
(can also use pull-down menus as shown in lab)

```
twoway (histogram re75t if treat==1, fcolor(none) lcolor(red))
(histogram re75t if treat==0, fcolor(none) lcolor(blue))

twoway (histogram re75t if treat==1, fcolor(none) lcolor(red))
(histogram re75t if treat==0 [fweight = _weight], fcolor(none)
lcolor(blue))
```

matched regression-adjusted treatment effect estimate

```
. regress re78 treat age educ black hisp married re74t re75t [pw=_weight]
(sum of wgt is   3.7000e+02)

Linear regression                               Number of obs =     313
                                                F(  8,   304) =    4.34
                                                Prob > F      =  0.0001
                                                R-squared     =  0.1386
                                                Root MSE      =  6673.9

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1920.304   765.6579     2.51   0.013      413.644    3426.964
         age |  -12.25284   47.06953    -0.26   0.795    -104.8762    80.37049
        educ |   339.0843   176.9679     1.92   0.056    -9.152767    687.3213
       black |  -509.4873   990.8662    -0.51   0.607    -2459.312    1440.337
        hisp |   1629.462   1774.517     0.92   0.359    -1862.429    5121.353
     married |  -853.3803   1114.502    -0.77   0.444    -3046.495    1339.735
       re74t |   167.9214   154.7012     1.09   0.279    -136.4994    472.3422
       re75t |   468.2597    194.053     2.41   0.016     86.40256    850.1168
       _cons |   661.9076   2584.407     0.26   0.798    -4423.683    5747.498
------------------------------------------------------------------------------
```

Return to Step 2 based on poor overlap

If we discover poor overlap what might we do when we return to Step 2??

If you return to Step 2

- Make sure you redo Steps 3 and 4 as well!

Step 5: Treatment Effect Estimation

- Once you are satisfied with the balance achieved by your model/ matching you can estimate treatment effects in several ways:
  - *Difference in means*: Comparing the mean outcomes across matched groups

  - *Regression-adjusted matched estimate*: Running a regression of outcome on treatment indicator and confounding covariates using weights to force sample to represent matched groups (1 if in treatment group, 0 if not matched, and # times matched for matched controls)

Step 5: Difference in means (Typically not a good choice)

- Now run psmatch2 again with out() option specified
  psmatch2 treat age educ black hisp married re74 re75, out(re78)

- Now output includes the following

```
------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------
           re78  Unmatched |  6349.1435   14846.6596  -8497.51612   712.020719   -11.93
                       ATT |  6349.1435   4481.65713   1867.48637   853.180748     2.19
-----------------------------+------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.
```

- Note, you could also use this to calculate the difference in means for covariates

Step 5: Regression-adjusted matched estimates

- There may be some remaining imbalance even after matching. Regression allows you to further adjust for this imbalance which we hope will reduce bias.
- It also helps to reduce standard errors.

Step 5: How do we implement regression adjusted treatment effect estimation?

- We can either created a matched sample (only works for matching without replacement)
- Another (more flexible) alternative is regression with weights. Observations will be weighted based on how many times they are used in the analysis
- So when estimating the effect of the treatment on the treated,
  - the treatment group stays intact, so each treated unit gets a weight of 1.
  - each control gets a weight equal to the number of times it was used as a match.
- Don't confuse this approach with Inverse Probability of Treatment Weighting -- another propensity score approach

- Then use the weights as probability weights in a regression (notice gives same point estimate!)
- This just allows us to reflect our restructured data!

```
regress re78 treat [pw=_weight]
(sum of wgt is 3.7000e+02)

Linear regression                               Number of obs =     319
                                                F( 1,   317) =    5.56
                                                Prob > F      =  0.0190
                                                R-squared     =  0.0176
                                                Root MSE      =  6991.1

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1867.486   791.9337     2.36   0.019     309.3761    3425.597
       _cons |   4481.657   540.6431     8.29   0.000     3417.955    5545.359
------------------------------------------------------------------------------

.
```

- Can also perform additional covariates adjustment

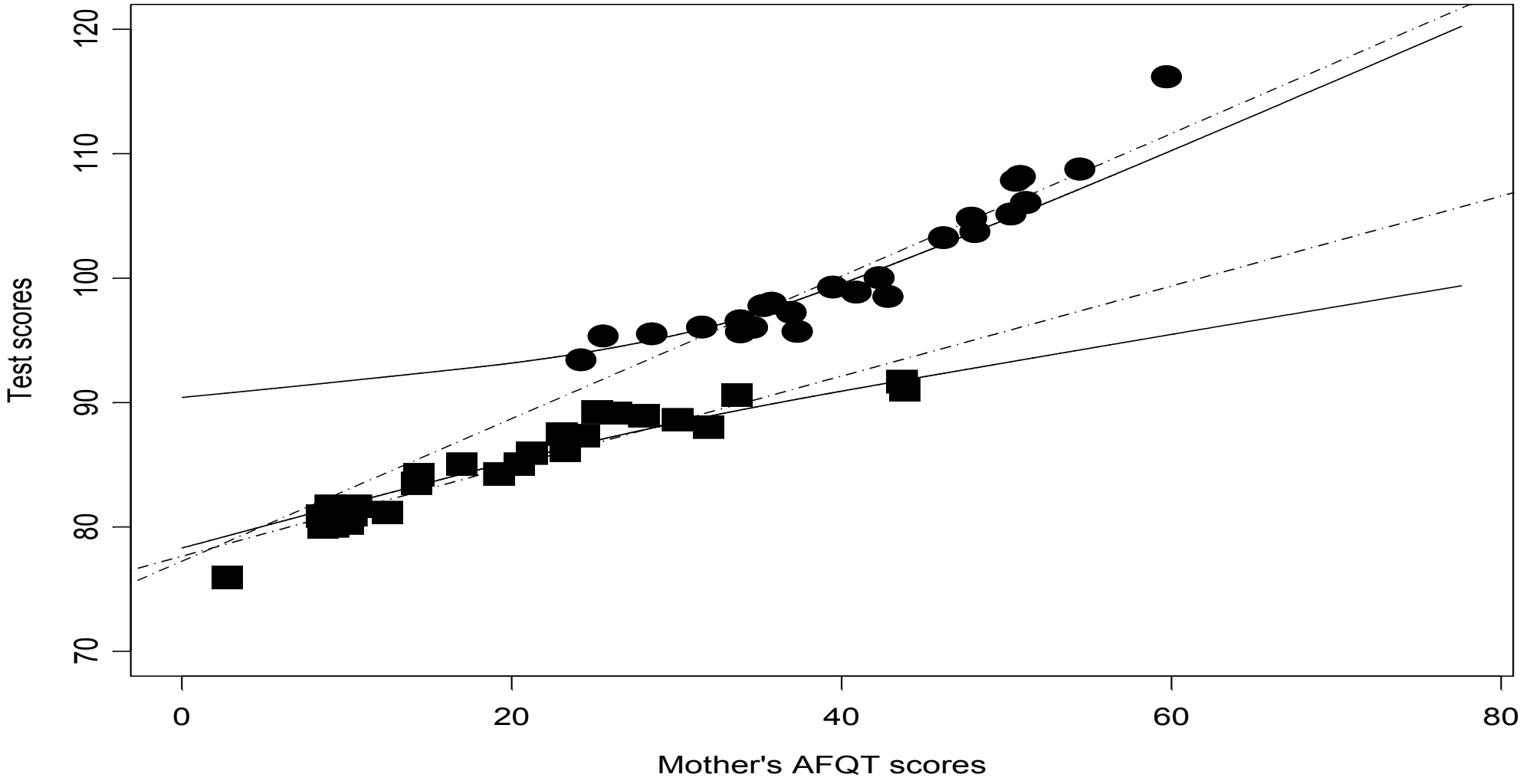
```
. regress re78 treat age educ black hisp married re74t re75t [pw=_weight]
(sum of wgt is   3.7000e+02)
Linear regression                                       Number of obs =     319
                                                        F(  8,   310) =    2.47
                                                        Prob > F      =  0.0132
                                                        R-squared     =  0.0740
                                                        Root MSE      =  6863.9
------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1851.513   764.7679     2.42   0.016     346.7205    3356.306
         age |  -22.72532   39.02057    -0.58   0.561    -99.50398    54.05333
        educ |   271.1451   157.9516     1.72   0.087    -39.64781     581.938
       black |  -2053.403   1170.323    -1.75   0.080    -4356.184    249.3782
        hisp |  -2440.454    2022.89    -1.21   0.229    -6420.784    1539.877
     married |   80.27376   1022.811     0.08   0.937    -1932.257    2092.804
       re74t |   55.08028    180.252     0.31   0.760    -299.5919    409.7524
       re75t |    440.362   228.0364     1.93   0.054    -8.332856    889.0568
       _cons |   3350.415   2580.098     1.30   0.195    -1726.305    8427.136
------------------------------------------------------------------------------
```

Wait a second….!

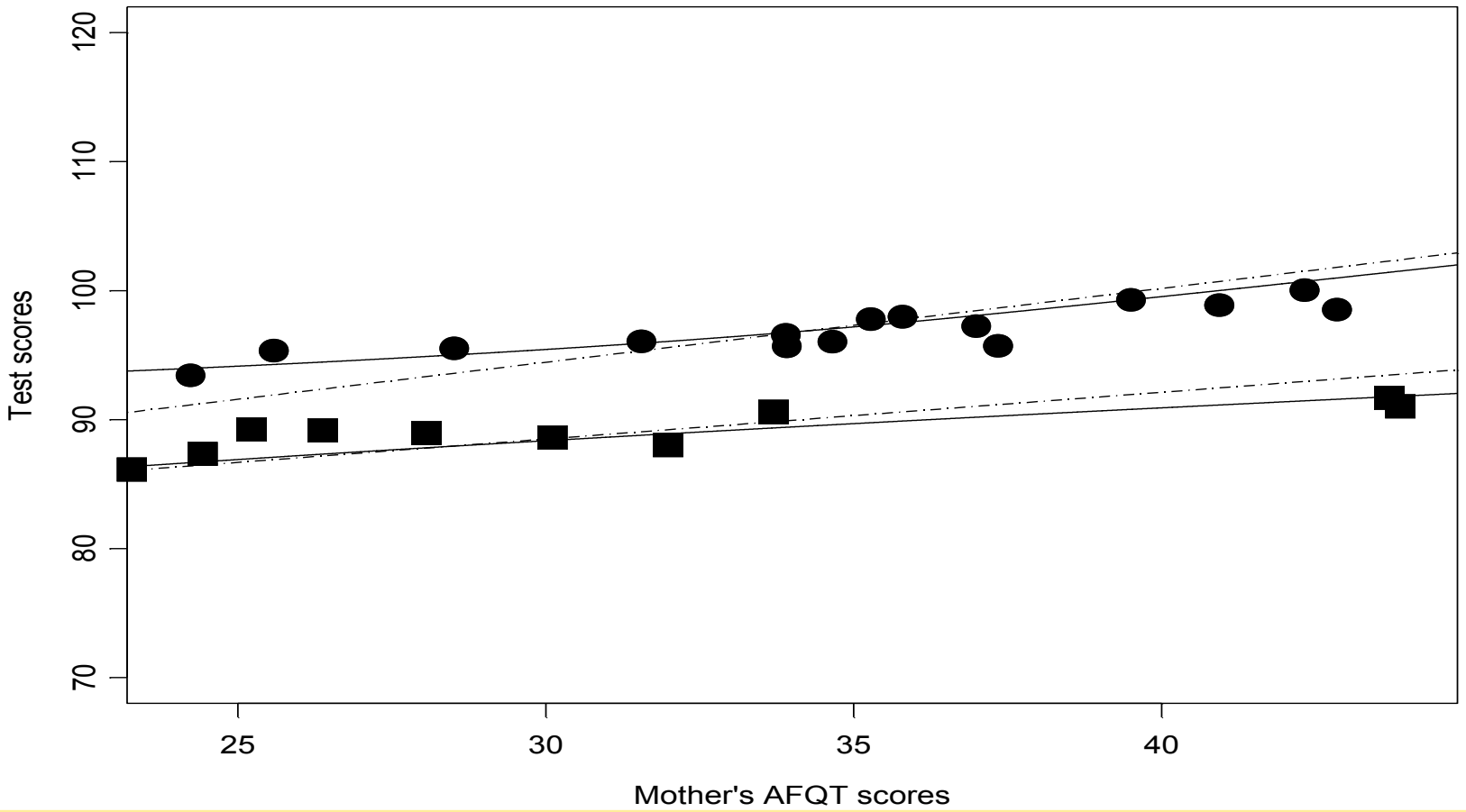
- I thought the whole reason we were using a propensity score strategy was to avoid the restrictive parametric assumptions of methods like linear regression?
- Regression is justifiable here because overlap means we shouldn't be extrapolating over parts of the covariate space with no data

The lack of overlap and imbalance are the main issues causing trouble for regression.

Recall how regression causes trouble…



Justification for regression after matching



Simply limiting the sample (these are the same data) to the subsample where there's overlap makes a huge difference in what we can reliably estimate

- Compare the matched results (t.e. =) to the same regression without weights (so a standard analysis of unmatched data)

```
.    regress re78 treat age educ black hisp married re74t re75t
```

- Where the results are:

```
. regress re78 treat age educ black hisp married re74t re75t

      Source |       SS       df       MS              Number of obs =   16177
-------------+------------------------------           F(  8, 16168) = 1833.55
       Model |  7.1967e+11     8  8.9959e+10           Prob > F      =  0.0000
    Residual |  7.9324e+11 16168    49062556           R-squared     =  0.4757
-------------+------------------------------           Adj R-squared =  0.4754
       Total |  1.5129e+12 16176  93528158.4           Root MSE      =  7004.5

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   754.2433   547.0619     1.38   0.168    -318.0586    1826.545
         age |  -101.6858   5.880837   -17.29   0.000    -113.2129   -90.15875
```

Constructed observational study: redux

Success stories…? Turns out……

- We just used a classic dataset for evaluating the performance of observational methods was first used by LaLonde (1983) to test the efficacy of a variety of econometric techniques
- He started with data from a randomized field experiment (National Supported Work) and "constructed" an observational study by using the treatment group from this experiment and adding comparison groups created from contemporaneous survey data (PSID, CPS)
- Thus the observational estimates can be evaluated using the experimental estimate as a reliable benchmark of the "truth". Based on the randomized data, the estimate of the average treatment effect is about 1800 dollars/year.
- This study was repeated (Dehejia & Wahba, 1999) using propensity score approaches which faired well but sparked a big debate

Constructed observational studies

- This concept of the "constructed observational study" has since been used more broadly
- Randomized experiments (where the true treatment effect is "known") are generally used as the starting point, but then the control group is replaced by a control group constructed from another survey, another location or cohort in the same experiment, etc. Then observational methods are used to see how close they can get to the experimental benchmarks.
- In two such cases (Dehejia & Wahba, 1999; Hill, Reiter & Zanutto, 2004 -- both presented today!) propensity scores were found to replicate experimental estimates well in settings where regression failed
- In a few others the results have been more mixed. In some case these studies were not well-constructed or p-scores strategies not carried out as they should have been. In others it may reflect situations where pscores were indeed not appropriate (e.g. if ignorability is not satisfied)

!!!Pop Quiz!!!

Why would we use propensity scores instead of linear regression?

_Less limitation in assumptions, can be considered as a more robust linear regression;
_Linear regression estimates the average effect of a combination of covariates, it is not aiming to estimate the causal effects.

Overview of most important assumptions for propensity score matching

The most important assumptions required for propensity score matching to yield valid causal inferences are

- Ignorability
- Sufficient overlap
- Appropriate specification of the propensity score model/ balance achieved
- SUTVA

Interpreting an estimate causally

For those who received job training among the analysis sample, their average income were about $1800 higher than had they not received the job training.