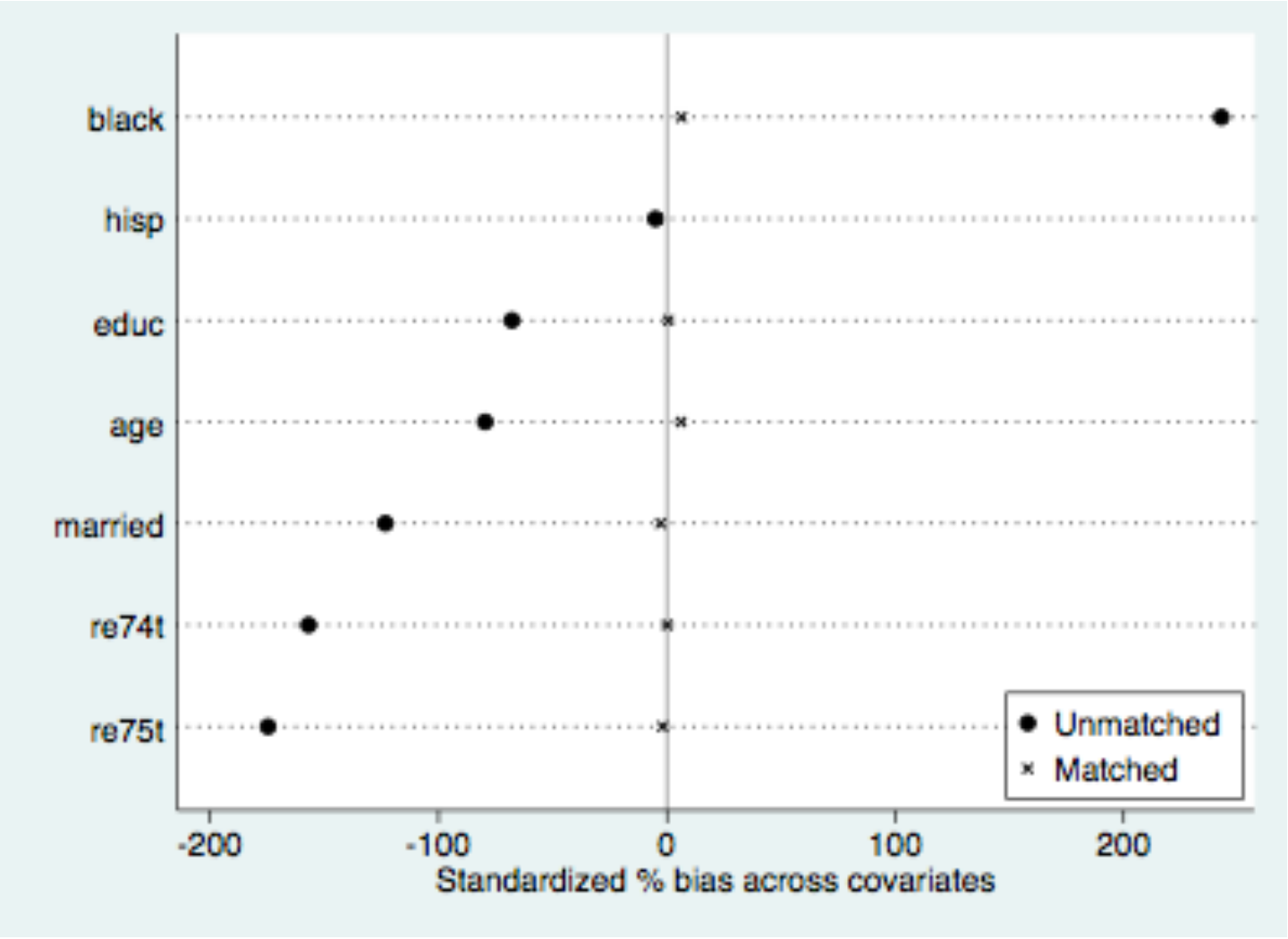
Propensity Score Strategies for Causal Inference

(Further uses and further issues)

additional balance check display available

pstest age educ black hisp married re74t re75t, both graph



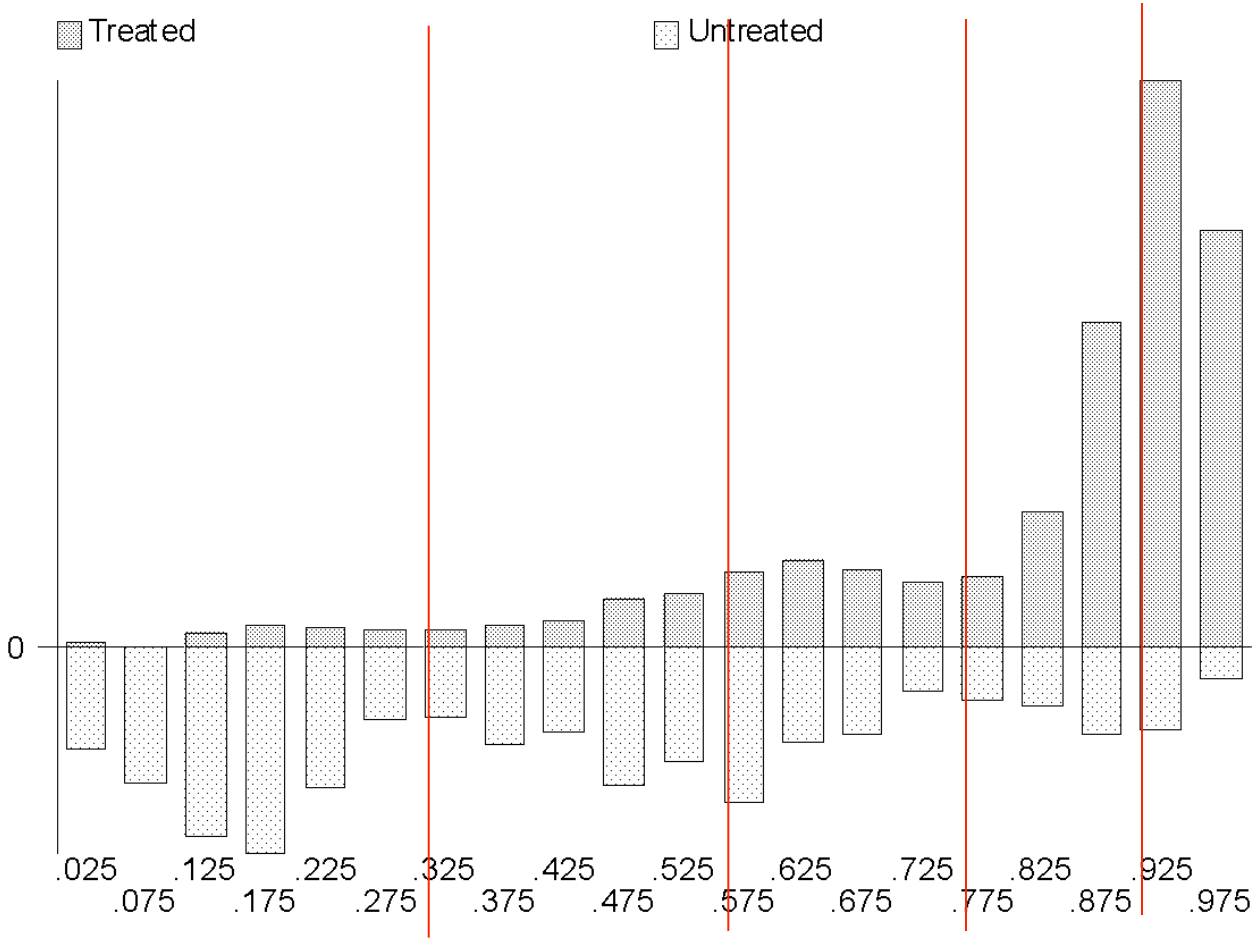
Propensity-score subclassification

Subclassification

- Another way of using propensity scores to make causal inferences
- Instead of using the propensity scores to match participants, we use them to divide our sample into strata, or subclasses
- Within each subclass we hope that our covariates are balanced across treatment groups (that is $X \perp Z$)
- Advantages: Can usually retain a much bigger sample (good for efficiency)
- Disadvantages: May incur greater bias

Subclassification

estimand: ATE



Five Steps to implement

- Calculate propensity scores as usual
- Identify the quartiles (quintiles, deciles, …) of the pscore distribution
- Use these as cut-points that determine subclasses
- Calculate treatment effect estimates, τ_j (differences in means, reg.estimates) within each subclass, i
- If desired, these can be weighted to form average treatment effects:
  - Overall average: $\Sigma\, \tau_i n_i / \Sigma n_i$
  - Treatment on treated: $\Sigma\, \tau_i n_i^{T} / \Sigma n_i^{T}$

Diagnostics

- Balance (e.g. difference in means across treatment groups for each variable) should be compared within each subclass

- Or, to look at balance overall you can use a two-way ANOVA with treatments and subclasses as factors

!!! Pop Quiz!!!

Compare subclassification with a classic randomized block design. What are the differences and similarities?

Inverse Probability of Treatment Weighting (IPTW)

Inverse probability of treatment weighting (IPTW)

- The goal of IPTW is intuitively the same as matching
  - We want to re-weight the control group to look like the treatment group (ATT)
  - We want to re-weight the treatment group to look like the control group (ATC)
- Or we want to re-weight both treatment and control groups to look like the full sample (ATE)

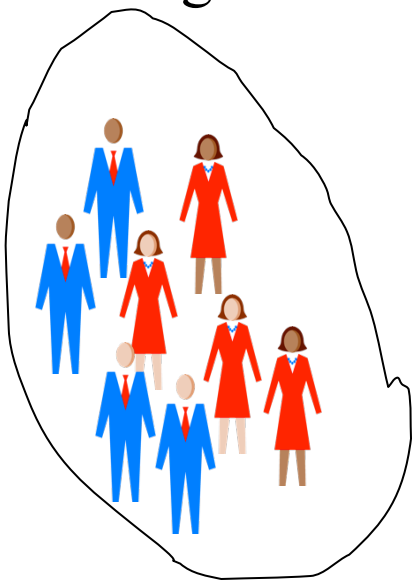
Average Treatment effect on Treated

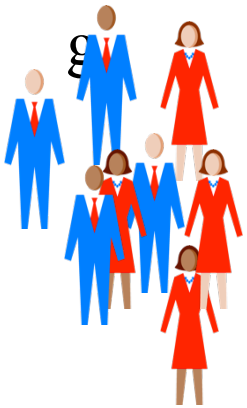
Receives no training

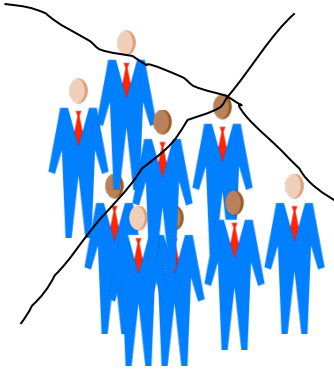
"Receives training" is treatment group

Recei ves trainin g

What weights would be applied here?







these people discarded

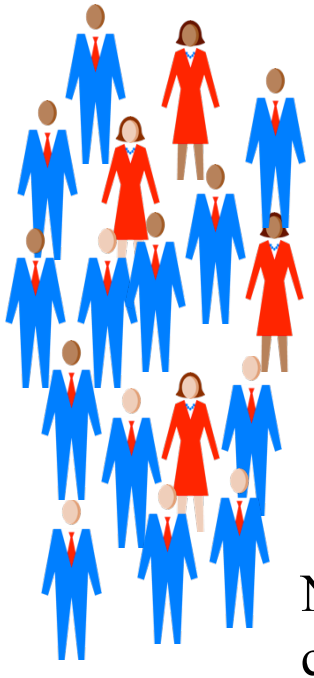
Average Treatment Effect on Control

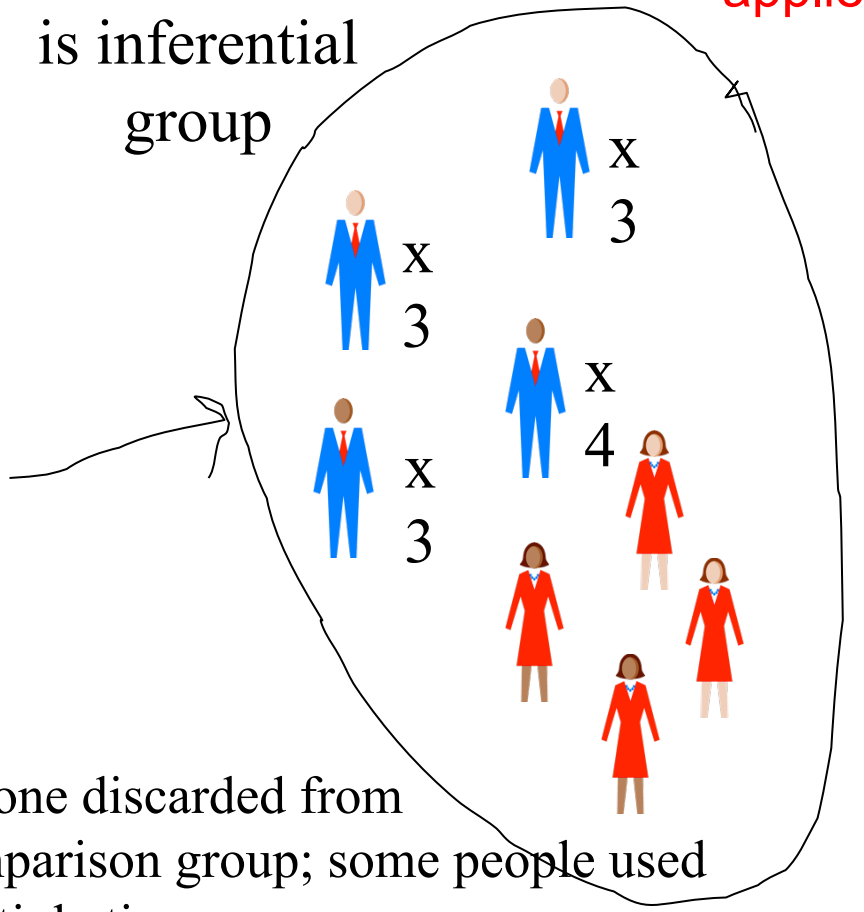
Does not get married

"Does not get married" is inferential group

Gets married

What weights wou applied here?





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

Weights for ATE: informal justification

- We want to re-weight each group so that they represent the full sample
- So, similar to typical survey weighting applications, we weight each group by the inverse of the (estimated) probability that they ended up in that group, i.e.
  - $1/\hat{e}(x)$ for the treatment group
  - $1/(1-\hat{e}(x))$ for the control group

(where $\hat{e}(x)$ is the estimate of $Pr(Z=1 \mid X)$)

!! Pop Quiz!!

If we can treat matching as weighting process, how do we assign weight in matching?

How does the matching weight different from inverse propensity weight here?

Weights for ATE: formal justification

It can be shown that, under ignorability

$$\mathrm{E}\left[\frac{YZ}{e(X)}\right] = \mathrm{E}\left[\frac{Y(1)Z}{e(X)}\right] = \mathrm{E}\left[\mathrm{E}\left[\frac{Y(1)Z}{e(X)} \mid X\right]\right] = \mathrm{E}\left[\frac{\mathrm{E}[Y(1) \mid X]e(X)}{e(X)}\right] = \mathrm{E}[Y(1)]$$

$$\mathrm{E}\left[\frac{Y(1-Z)}{1-e(X)}\right] = \mathrm{E}\left[\frac{Y(0)(1-Z)}{1-e(X)}\right] = \mathrm{E}\left[\mathrm{E}\left[\frac{Y(0)(1-Z)}{1-e(X)} \mid X\right]\right] = \mathrm{E}\left[\frac{\mathrm{E}[Y(1) \mid X](1-e(X))}{1-e(X)}\right] = \mathrm{E}[Y(0)]$$

(where e(X) is the propensity score: Pr(Z=1|X))

This can be used to justify weighted estimates of the following form:

Treated observations receive weights of 1/ê(x)

Control observations receive weights of 1/(1-ê(x))

[where ê(x) is the *estimated* propensity score]

Weighting for other estimands

- Average effect of the treatment on the treated

  E[Y(1)-Y(0) | Z=1] = E[Y(1) | Z=1] - E[Y(0) | Z=1]

  Using similar derivations as before

$$\mathrm{E}\left[Y(0) \mid Z=1\right] = \mathrm{E}\left[\frac{Y(1-Z)e(X)}{1-e(X)}\right]$$

  Therefore

  Treated get weights of 1 (don't need to re-weight)

  Controls get weights equal to ê(x)/(1-ê(x))

Weighting for other estimands

- Average effect of the treatment on the controls

E[Y(1)-Y(0) | Z=0] = E[Y(1) | Z=0] - E[Y(0) | Z=0]

$$\mathrm{E}\left[Y(1) \mid Z=0\right] = \mathrm{E}\left[\frac{YZ(1-e(X))}{e(X)}\right]$$

Therefore
Controls get weights of 1 (don't need to re-weight)
Treated get weights equal to (1-ê(x))/ê(x)

Weighting -- how it works

Example: ***effect of treatment on the treated***

Male	Z	Y(0)	Y(1)	Y	Wt
1	0	10	12	10	2
1	0	10	12	10	2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
1	1	10	12	12	1
1	1	10	12	12	1
1	1	10	12	12	1
1	1	10	12	12	1
0	1	4	10	10	1
0	1	4	10	10	1

$$\Pr(Z=1|\text{Male}=0)=1/3$$
$$\Pr(Z=1|\text{Male}=1)=2/3$$

So if Male=0, the (unnormalized) weight equals (1/3)/(2/3) = 1/2

If Male=1, the (unnormed) weight equals (2/3)/(1/3) = 2

Weights total to 6

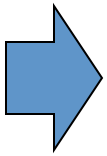
E[Y(0)|Z=1] = (2*10 + 2*10 + .5*4 + .5*4 + .5*4 + .5*4)/6

E[Y(1)|Z=1] = (12 +12+12+12+10+10)/6

Weighting -- how it works

Example: ***effect of treatment on the treated***

Male	Z	Y(0)	Y(1)	Y	Wt
1	0	10	12	10	2
1	0	10	12	10	2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
0	0	4	10	4	1/2
1	1	10	12	12	1
1	1	10	12	12	1
1	1	10	12	12	1
1	1	10	12	12	1
0	1	4	10	10	1
0	1	4	10	10	1



Male	Z	Y(0)	Y(1)	Y
1	0	10	12	10
1	0	10	12	10
1	0	10	12	10
1	0	10	12	10
0	0	4	10	4
0	0	4	10	4

In effect, the weights create a "pseudo-population" of controls that look like the treatment group

Weighing tradeoffs

- If used without balance checks this could lead to less robust estimates due to heavier reliance on the propensity score model
- This should be alleviated by
  - checking balance on re-weighted samples (just as you would with matching)
  - more robust estimation of the propensity score
- Weights can be unstable (there are ways to address through trimming or stabilized weights) [note as well that the "weights" produced by matching with replacement can also be unreasonable in some settings]
- Uses more of the data which can lead to greater precision

Trim or stabilize extreme weights?

- Weights can get "unstable" (large) in situations when the probability in the denominator is quite small (e.g. when someone unlikely to be in the treatment group ends up there anyway)
- These can be stabilized by e.g. changing the numerator (e.g. by Pr(Z=1))
- Others scholars advise "trimming" weights in this case. (Cole and Hernan, 2008) This is also called winsorizing and has a long tradition in the survey world. However there is (e.g. Lee et al.) conflicting evidence regarding the efficacy of this approach
- General advice is to focus on appropriate estimation of the propensity score model

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight Trimming and Propensity Score Weighting. *PLoS ONE*, *6*(3), e18174. http://doi.org/10.1371/journal.pone.0018174

Example in Stata: estimating the effect of the treatment on the treated

```
xi: psmatch2 treat age educ black hisp married re74t re75t re74tL
gen wts=1
replace wts=_pscore/(1-_pscore) if treat==0

* now normalize so the the weights in the control group
* sum to the number of control observations
egen sum_wts=sum(wts) if treat==0
egen nc=sum(1-treat) if treat==0
replace wts=wts*(nc/sum_wts) if treat==0

*** you might notice here that some weights too large!
*** not a problem for this example, but if it were
*** you could do something like:
replace wts=50 if wts>50
*** this next line crucial to trick psmatch into
*** calculating balance for us
replace _weight=wts
psbal2 age educ black hisp married re74t re75t
```

Balance using pscores from simpler model to create weights

```
. psbal2 age educ black hisp married re74t re75t
-------------------------+-----------------------------------+------------------
         age  Unmatched | 25.816   33.225     7.16    11.05  | -1.035       0.65
                Matched | 25.816   25.243     7.16    10.73  |  0.080       0.67
                        |                                    |
        educ  Unmatched | 10.346   12.028     2.01     2.87  | -0.836       0.70
                Matched | 10.346   10.331     2.01     2.81  |  0.008       0.72
                        |                                    |
       black  Unmatched |  0.843    0.074     0.36     0.26  |  2.111       1.40
                Matched |  0.843    0.823     0.36     0.38  |  0.056       0.96
                        |                                    |
        hisp  Unmatched |  0.059    0.072     0.24     0.26  | -0.053       0.92
                Matched |  0.059    0.070     0.24     0.25  | -0.043       0.93
                        |                                    |
     married  Unmatched |  0.189    0.712     0.39     0.45  | -1.331       0.87
                Matched |  0.189    0.201     0.39     0.40  | -0.030       0.98
                        |                                    |
       re74t  Unmatched |  2.096   14.017     4.89     9.57  | -2.440       0.51
                Matched |  2.096    2.079     4.89     4.76  |  0.003       1.03
                        |                                    |
       re75t  Unmatched |  1.532   13.651     3.22     9.27  | -3.764       0.35
                Matched |  1.532    1.662     3.22     3.46  | -0.040       0.93
                        |                                    |
-------------------------------------------------------------------------------
```

Not getting great balance in sd's for age

(maybe try more complicated model)

- `gen age2=age^2`
- `xi: psmatch2 treat age age2 educ black hisp married re74t re75t re74tL i.educ_cat`

(generate weights as before)

```
. regress re78 treat age educ black hisp
married re74 re75 [pw=_weight]
```

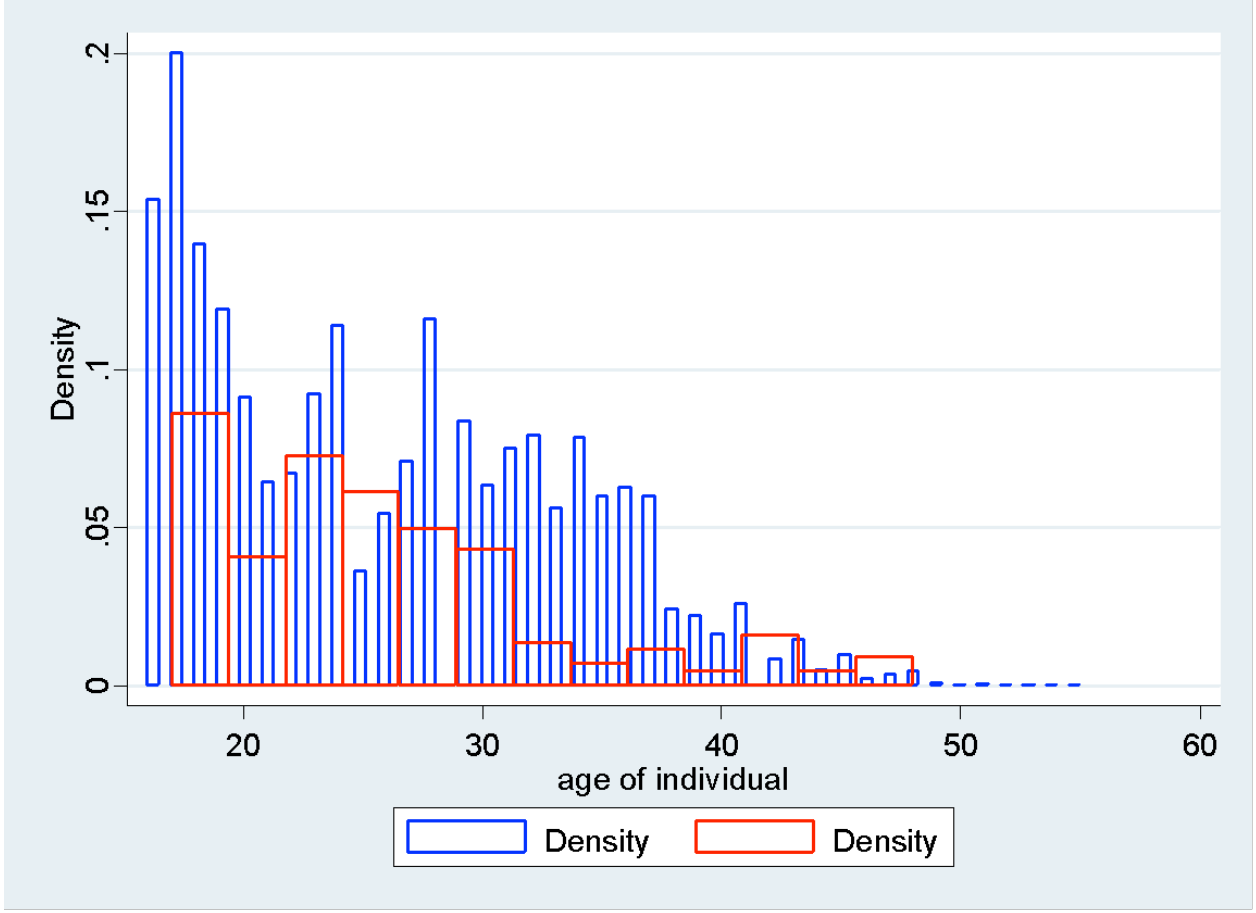
Balance using pscores from more complicated model to create weights

```
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	25.830	7.16	7.61	-0.002	0.94
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.496	2.01	2.03	-0.075	0.99
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.818	0.36	0.39	0.068	0.95
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.068	0.24	0.25	-0.037	0.94
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.237	0.39	0.43	-0.121	0.92
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.642	4.89	5.06	-0.112	0.97
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.929	3.22	3.63	-0.123	0.89

better

Post-weighting Balance in Age



```
. regress re78 treat age educ black hisp married re74 re75 [pw=_weight]
(sum of wgt is   1.4256e+04)

Linear regression                                      Number of obs =   16177
                                                       F(  8, 16168) =   46.32
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2566
                                                       Root MSE      =  5518.2

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1575.722   667.5531     2.36   0.018     267.2443      2884.2
         age |   -100.773   35.76951    -2.82   0.005    -170.8852    -30.6608
        educ |   309.7614   116.2508     2.66   0.008       81.897    537.6259
       black |  -842.9317   395.5205    -2.13   0.033    -1618.196   -67.66779
        hisp |   1035.691   481.6796     2.15   0.032     91.54537    1979.836
     married |  -717.0243    624.765    -1.15   0.251    -1941.633    507.5844
        re74 |    .209974    .079527     2.64   0.008     .0540924    .3658557
        re75 |   .6759959    .090615     7.46   0.000     .4983804    .8536113
       _cons |   3479.413   1499.571     2.32   0.020     540.0884    6418.737
------------------------------------------------------------------------------
```

!!! Pop Quiz !!!

- True or False?

- If I use a weight option in the model I use to estimate my treatment effect then I am using IPTW.

Standard errors

Standard Errors

- Correctly calculating standard errors is a problem for several reasons:
  - after matching the observations are no longer independent of each other (even with regular matching)
  - matching is based on *estimated* propensity scores
- Matching with replacement creates the additional complication of including some units multiple times but this is easily remedied by standard software (e.g. by using the pweights option in Stata)
- Abadie & Imbens probably have best solution out there but it is labor intensive (involves further matching)
- Quick question: Is the standard error smaller or larger in matched samples? why?

Correlation between matched pairs

- Units in our original sample presumably are *independent* of each other (this is one of the implications of performing a random sample)
- Units in our matched sample, however, were chosen specifically because they are *similar* to one another.
-
- Consider the variance calculations for the differences in means,

$$\mathrm{Var}(\bar{Y}^T - \bar{Y}^{MC}) = \mathrm{Var}(\bar{Y}^T) + \mathrm{Var}(\bar{Y}^{MC}) - 2\mathrm{Cov}(\bar{Y}^T, \bar{Y}^{MC})$$

- If the means are independent, the last term is equal to 0
- If they are positively correlated (as we would expected them to be) then the last term>0
- Therefore, all else equal, standard calculations should overestimate the true variance.

Not a good solution: Matched pairs

- Some researchers advocate using matched pairs analyses to account for the *within-pair correlation*
- However pscore matching isn't really optimized for finding close pair matches (better at finding balance overall)
- Therefore this can be an inefficient choice

Better alternative: Fitting a model

- If we fit a model for E[Y | Z, X] (and it's the right model) then we may be able to account for these correlations if they are solely driven by similarity to to X
- After all, we just need our errors (what's left over after adjusting for X) to be independent

Back to Step 3: Other ways to restructure using propensity scores

- Subclassification
- Inverse probability of treatment weighting

Other issues

Non-binary treatment variables

Multiple treatment groups

Here is one potential solution (Imbens, 2001)

- First predict the probability of belonging to each of the K treatment (dosage) groups using polytomous regression or multinomial logit (or similar)
- Each person will have K probabilities attached to them (one for each of the treatment groups)
- Expected values of the outcome for the k^{th} group can be calculated by weighting the outcomes of people in that group using the inverse of each person's probability of being in that group

- (Rosenbaum et al. 2001 present a similar, though more complicated, alternative)

$$E\left[\frac{Y \cdot I(Z=k)}{\Pr(Z=k \mid X)}\right] = E\left[Y(Z=k)\right]$$

Continuous treatment variables

Here is a possible solution (Imai & van Dyk, 2004)

- Calculate propensity scores by running a standard linear regression of the treatment variable, Z, on confounding covariates, X (can actually use any of a variety of models that might model this relationship more accurately) and getting predictions for Z
- Subclassify the data based on these propensity scores
- Calculate regression (e.g.) estimates of the treatment effect within each subclass
- Combine appropriately for average effects

Binary outcomes and collapsibility

Binary (and other problematic) outcomes

- Consider what happens when we have a binary outcome
- A common model for treatment effect estimation (assuming we want to perform additional covariance adjustment) would be a logistic regression
- The treatment effect estimated with the standard logistic regression coefficients (exponentiated) would be an odds ratio
- *The odds ratio is not collapsible*

Collapsibility

- Consider the type of estimand we tend to focus on
  $E[Y(1) - Y(0)]$
- A nice property of this is that
  $E[Y(1) - Y(0)] = E[E[Y(1) - Y(0) \mid X]]$
- Conceptually, this means that the average treatment effect that you get by getting average treatment effects at each level of X and averaging over the distribution of X is the same as the average treatment effect marginally
- Which is why we can use additional covariance adjustment post-matching without changing what we are trying to estimate
- This property does not hold for odds ratios etc, thus the marginal and conditional odds ratios actually *mean* something different

Collapsibility: linear example

X=x	E[Y(0)|X=x]	E[Y(1)|X=x]	E[Y(1)-Y(0) | X=x]	Pr(X=x)
x=1	8	12	4	.2
x=2	4	12	8	.7
x=3	10	12	2	.1
	E[Y(0)] = 5.4	E[Y(1)] = 12	E[Y(1)] – E[Y(0)] = 6.6	

$$E[Y(1)-Y(0)] = E_x[E[Y(1)-Y(0) \mid X=x]]$$
$$= \Sigma_x E[Y(1)-Y(0) \mid X=x]Pr(X=x)$$
$$= 6.6$$

Collapsibility: Odds ratio example

$$OR = [Pr(Y=1 \mid Z=1)/(1-Pr(Y=1 \mid Z=1)]/$$
$$[Pr(Y=1 \mid Z=0)/)1-Pr(Y=1 \mid Z=0)]]$$

X=x	p_0=E[Y(0)|X=x]	p_1=E[Y(1)|X=x]	E[OR | X=x]	Pr(X=x)
x=1	.4	.8	6	.2
x=2	.2	.8	16	.7
x=3	.6	.8	2.67	.1
	p_0= .28	p_1= .8	10.28≠12.6668	

$$E[OR] \neq E_x[E[OR \mid X=x]]$$
$$E[OR] \neq \Sigma_x E[OR \mid X=x]Pr(X=x)$$
$$10.28 \neq 12.6667$$

Implications when you have a binary outcome

- The marginal odds ratio and conditional odds ratios actually *mean* something different.
- So even after perfect matching or in a randomized experiment the coefficient on the treatment variable in a logistic regression that doesn't adjust for anything else does not represent that same population quantity as the coefficient on the treatment variable in a logistic regression that includes the treatment variable as well as other covariates
- Same problem holds for probit regression.
- "Solutions"?
  - Fit a linear regression (estimand goes back to a difference in "means" (probabilities)
  - Get predicted probabilities for each group and then take the difference (will have to bootstrap the standard error)

Propensity scores in the context of data with survey weights

Incorporating survey weights

- Suppose your sample data are from a survey which has calculated weights, that allow the sample to be representative of a larger population of interest
- Let I be an indicator for whether an observation was included in the sample. These weights, $w(x) = 1/Pr(I=1 \mid X)$
- Let $e(x) = Pr(Z=1 \mid X)$
- Suppose we are using weights to calculate the ATE for the population. These weights should be equal to
- $1/Pr(I=1 \text{ and } Z=z)$
- $Pr(I=1 \text{ and } Z=z) = Pr(I=1)Pr(Z=z \mid I=1)$

  $= (1/w(x))*e(x)$ (when Z=1)

  $= (1/w(x))*(1-e(x))$ (when Z=0)

Incorporating survey weights (continued)

Therefore weights for ATE that incorporate survey weights would look like

- $1/ [(1/w(x))*e(x)]$ for the treated (Z=1)
- $1/ [(1/w(x))*(1-e(x))]$ for the controls (Z=0)

- Note also that the propensity score, e(x), should be estimated using a method that reflects the survey weights as well

Propensity scores and multilevel data

Propensity scores and multilevel data

What is the best approach for calculating and using propensity scores with grouped/multilevel data?

Recent research suggests doing one or more of the following:

1) Fit the propensity score model using a fixed effects or multilevel model
2) Match within group if possible
3) Fit the analysis model to estimate treatment effects (on matched or weighted data) using a fixed or random effects model

Caveat to (3): Random effects probably not the best choice for (3) if there may be unobserved group structure that is correlated with the outcomes

Also...

- Why would we use propensity score approaches rather than a linear regression?
- What is it buying us?

Overall Conclusions

- Propensity score matching DOES NOT….

  solve the "omitted variable bias" or "selection bias" problem.

  – Recall that propensity score matching still relies on a very **strong assumption**. We need to assume that we have controlled for all the potential confounders.

- Propensity score matching DOES…

  – rely on weaker (i.e. more plausible) assumptions about the way that Y and X are related to each other (for instance as compared to linear regression)

- While implementation is easy to do in theory, in practice there are lots of choices to make and not a lot of guidance in the literature about "best practice"

Reproducibility

- Look at all the decisions that are being made here!
- Would it be easy for someone to reproduce what you've done afterwards?
- If you are using this method (particularly for a research paper) I would urge you to (at the very least) document not only what commands you are using but why!
- You might even want to come up with a pre-analysis plan documenting in advance how you will make decisions.
- OR you could use a more automated method that is geared towards finding good balance like twang or genmatch.

(if time allows)

More matching methods

Other Matching methods

Beyond simple one-to-one nearest neighbor type matching algorithms there are host of others (these methods could be investigated further as part of a final project)

- Matching with multiple controls
  - K-nearest (Stata)
  - *Within calipers (radius matching) (Stata)
  - Kernel matching (Stata)
- **Full matching/Optimal matching
- *Mahalanobis matching (within propensity score intervals)
- **Genetic matching (genmatch in R)
- (Coarsened matching) (wouldn't recommend)

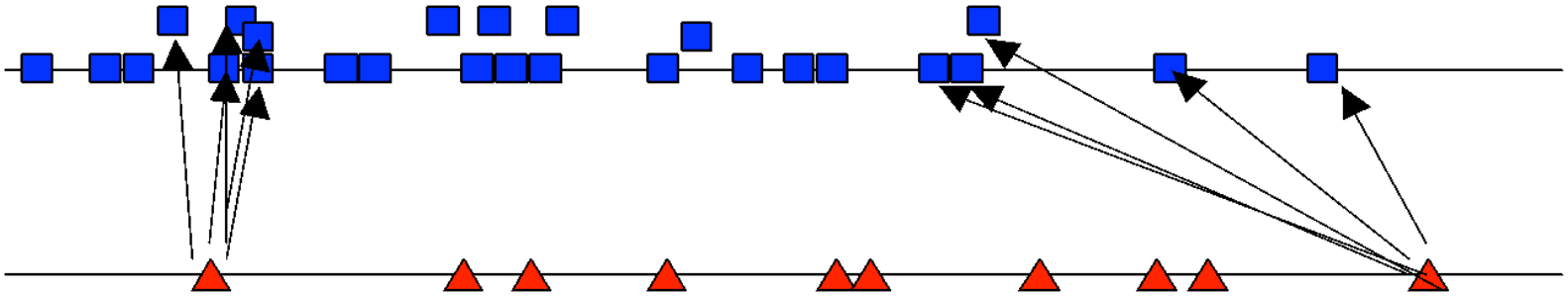
K-closest matching

- Choose the *k* controls units with the closest propensity scores to be matches
- Control units can be used more than once
- Match-specific treatment effects can be calculated as

$$Y_i^T - (\Sigma\, Y_{ij}^C)/k$$

  where *i* indexes treatment unit "stratum", *j* indexes control match within the i^{th} stratum
- These individual treatment effects can be averaged to get the effect of the treatment on the treated
- When k=1 this is matching with replacement

Matching with multiple controls: k-closest



K-closest (n=2)

estimate is $1510

```
. psmatch2 treat age educ black hisp married re74t re75t, n(2)
. psbal2 age educ black hisp married re74t re75t
```

age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	25.841	7.16	10.48	-0.003	0.68
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.324	2.01	3.00	0.011	0.67
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.846	0.36	0.36	-0.007	1.01
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	0.189	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	2.139	4.89	3.53	-0.009	1.38
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.422	3.22	2.44	0.034	1.32

K-closest (n=5)

estimate is $1467

```
. psmatch2 treat age educ black hisp married re74 re75, n(5)
. 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	25.477	7.16	10.78	0.047	0.66
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.394	2.01	2.90	-0.024	0.69
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.848	0.36	0.36	-0.012	1.01
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	0.176	0.39	0.38	0.033	1.03
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.159	4.89	3.66	-0.013	1.34
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.519	3.22	2.67	0.004	1.20

K-closest (n=15) estimate is $1321

```
. psmatch2 treat age educ black hisp re74 re75, n(15)
. psbal2 $covs
```

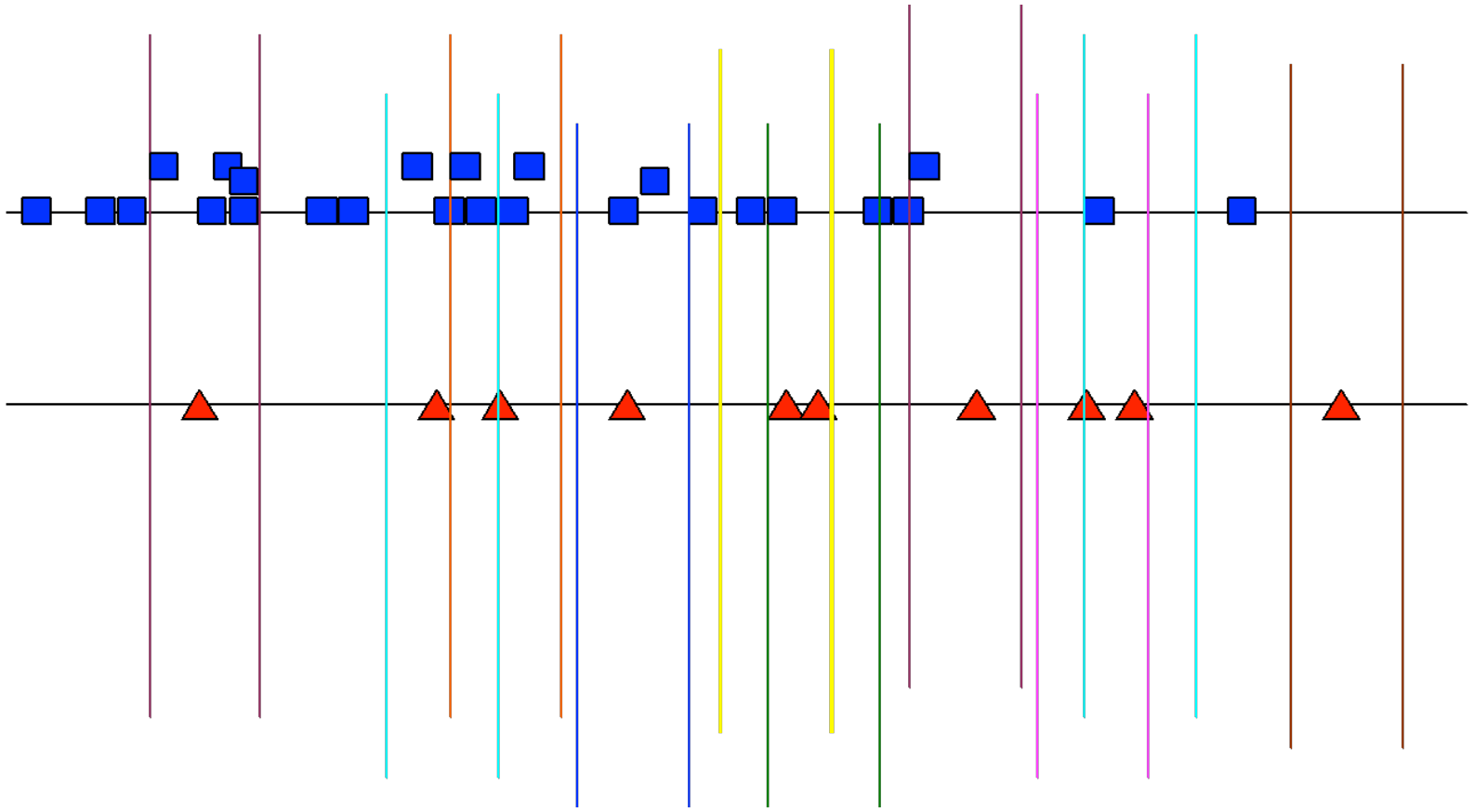
Variable	Sample	Mean Treated	Mean Control	SD Treated	SD Control	STD Diff	Ratio of SDs
age	Unmatched	25.816	33.225	7.16	11.05	-1.035	0.65
	Matched	25.816	24.644	7.16	10.25	0.164	0.70
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.346	10.483	2.01	2.78	-0.068	0.72
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.843	0.837	0.36	0.37	0.018	0.99
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.059	0.044	0.24	0.21	0.065	1.16
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.189	0.155	0.39	0.36	0.086	1.08
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.096	2.109	4.89	3.58	-0.003	1.37
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.532	1.496	3.22	2.63	0.011	1.23

.

Radius matching

- Choose a caliper width, c
- Within a distance of $c/2$ units on either side of the control units propensity score, keep all controls as matches
- Control units can be used more than once (i.e. for more than one treatment unit)

Matching with multiple controls within a certain caliper (radius matching)



Radius matching (standard estimation)

- Match-specific treatment effects can be calculated as

$$Y_i^T - (\Sigma\, Y_{ij}^C)/n_i$$

  where i indexes treatment unit "stratum", j indexes control match within the i^{th} stratum, and n_i denote the number of control units that fall in stratum i
- These individual treatment effects can be averaged to get the effect of the treatment on the treated

Radius Matching (estimate is $1458)

```
. psmatch2 treat age educ black hisp married re74t re75t, radius caliper(.001)
. 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	25.894	25.250	7.22	10.19	0.090	0.71
educ	Unmatched	10.346	12.028	2.01	2.87	-0.836	0.70
	Matched	10.509	10.435	1.84	2.87	0.037	0.64
black	Unmatched	0.843	0.074	0.36	0.26	2.111	1.40
	Matched	0.820	0.821	0.39	0.38	-0.002	1.00
hisp	Unmatched	0.059	0.072	0.24	0.26	-0.053	0.92
	Matched	0.068	0.042	0.25	0.20	0.112	1.26
married	Unmatched	0.189	0.712	0.39	0.45	-1.331	0.87
	Matched	0.205	0.207	0.40	0.40	-0.004	1.00
re74t	Unmatched	2.096	14.017	4.89	9.57	-2.440	0.51
	Matched	2.252	2.493	5.08	4.30	-0.049	1.18
re75t	Unmatched	1.532	13.651	3.22	9.27	-3.764	0.35
	Matched	1.671	1.803	3.40	3.58	-0.041	0.95

Radius matching (kernel estimation)

- Kernel estimation weights responses for the control units differentially depending on how far they are from the treatment unit in terms of their propensity scores
- Match-specific treatment effects can be calculated as

$$Y_i^T - (\Sigma w_i \, Y_{ij}^C)/n_i$$

where w_i is a weight calculated assuming a particular distributional form for the control observations

- These individual treatment effects can be averaged to get the effect of the treatment on the treated

Radius matching -- kernels

```
psmatch2 treat age educ black hisp re74 re75, kernel
caliper(.0001)

      Robust
re78  Coef.      Std. Err.  t    P>t    [95% Conf. Interval]

treat 1159.666   897.728  1.29   0.197  -600.6489  2919.981
age   117.7453   43.71652 2.69   0.007  32.02354   203.4671
```

Mahalanobis matching

- Mahalanobis matching was the most popular form of nearest neighbor matching before pscore matching started being used
- It uses a distance measure in multivariate space that takes into account variances of variables as well as covariances between them
- Treatment effect estimation just as if MWR had been performed
- It finds closer matches in the (scaled) covariate space as compared to propensity score matching
- In most cases it privileges higher order interactions across variables more than it needs to
- It is calculated in this setting as

$$m^2 = (x_T - x_C)' \Sigma_{CR}^{-1} (x_T - x_C)$$

(CR stands for control reservoir)

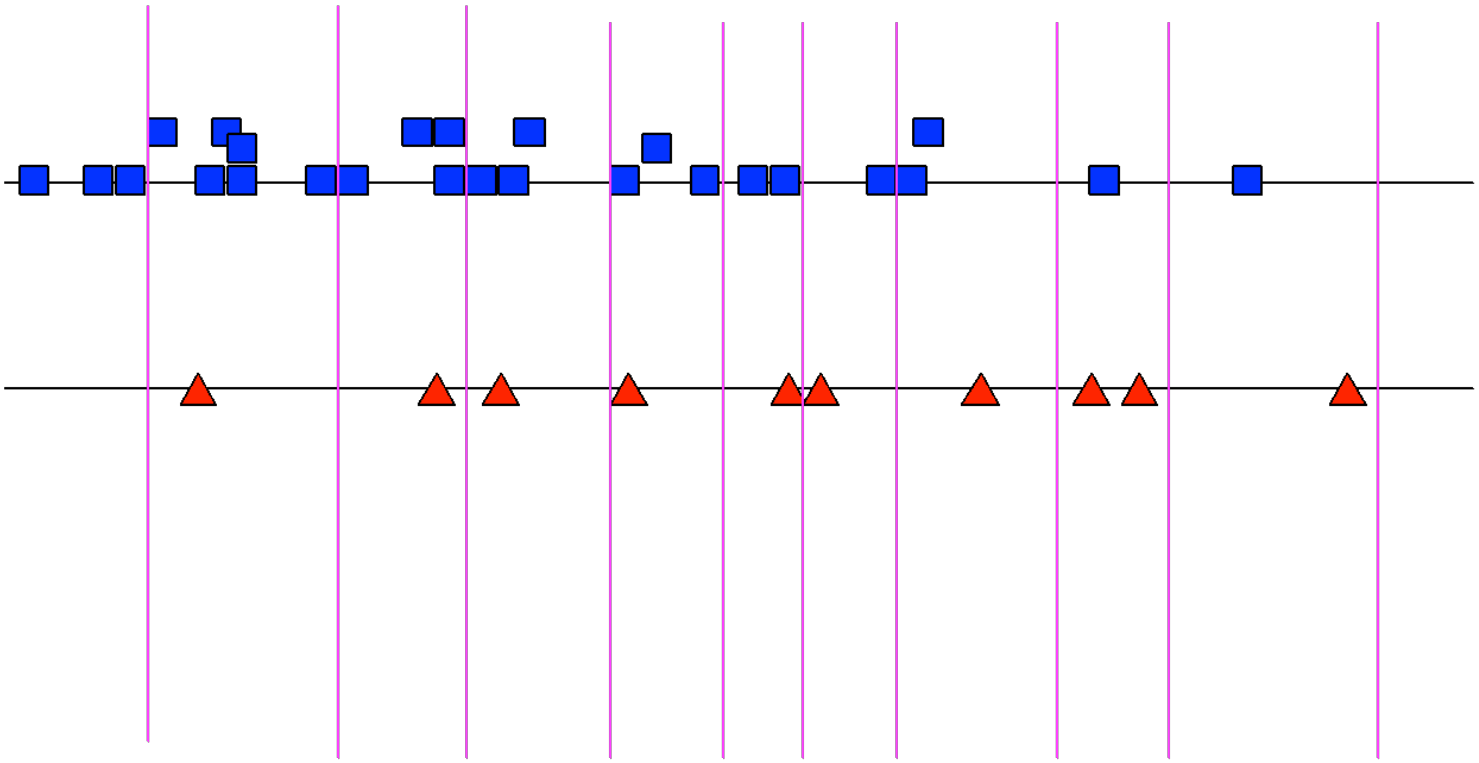
Matching methods not in psmatch2

- Mahalanobis matching within propensity score calipers
- Full matching
- Matching using genetic algorithms

Full matching (Rosenbaum)

- In full matching, the observations are subdivided into strata such that each stratum contains either:
  - one treatment unit and multiple controls, or
  - one control unit and multiple treateds, or
  - one control unit and one treated unit
- Satisfies certain optimality criteria (minimizes the total distance between treats and controls; strata are non-overlapping)
- Requires sophisticated algorithms to choose these strata appropriately

Full matching



Full matching

- Stratum-specific treatment effects can be calculated as

$$\tau_i=(\Sigma\ Y_{ij}^{T})/n_i^{T}-(\Sigma\ Y_{ij}^{C})/n_i^{C}$$

  where n_i^T and n_i^C are the number of treatment and control units, respectively, in the i^{th} stratum
- These individual treatment effects can be averaged to get the effect of the treatment on the treated,

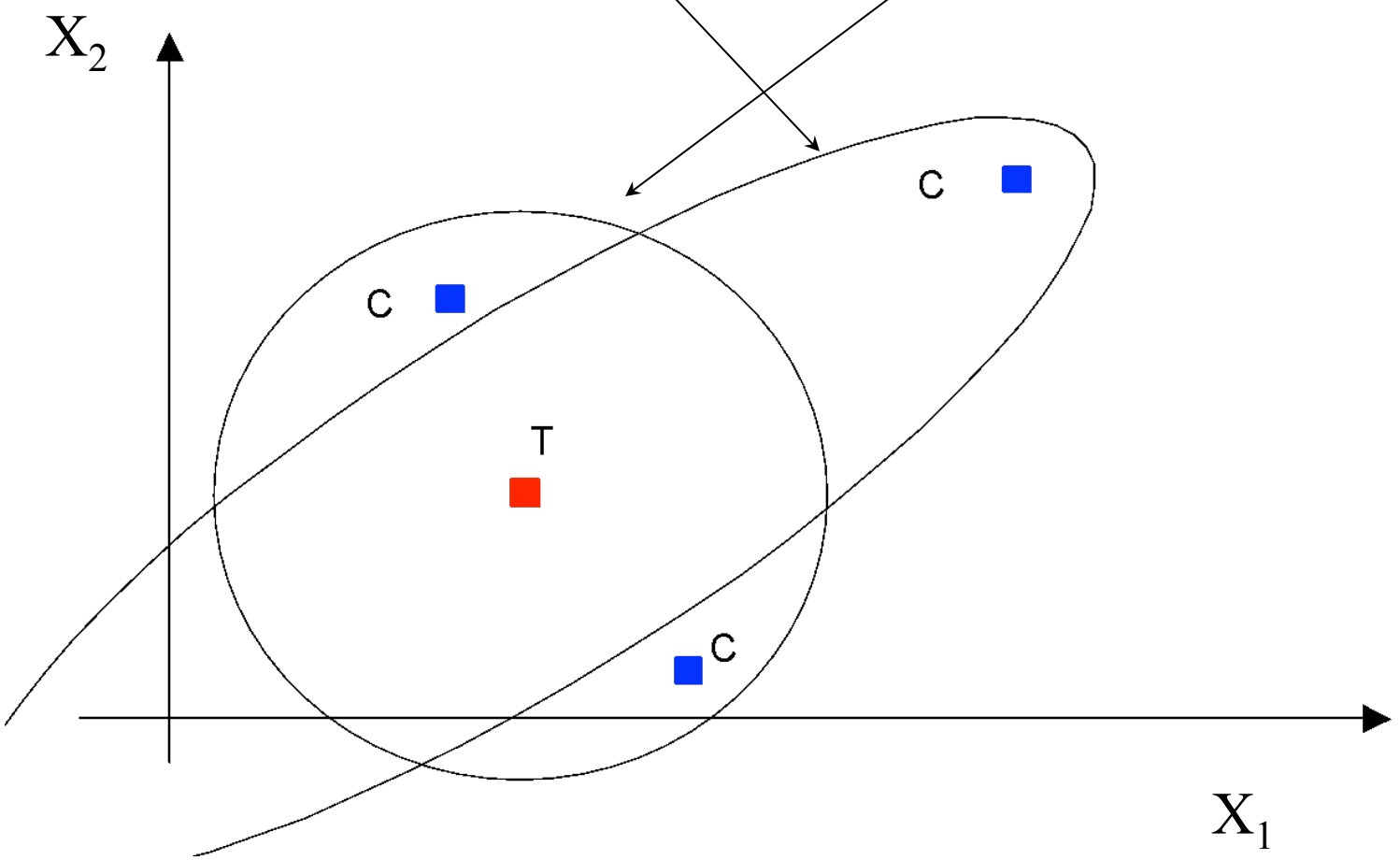
$$\tau = \Sigma\ n_i^T\ \tau_i$$

- Rosenbaum recommends using randomization-based inference (e.g. Hodges-Lehmann aligned rank test) though which will yield both a treatment effect estimate and confidence interval

Mahalanobis distance

Better than Euclidean distance

- scale-free
- takes into consideration associations among variables



In this example X_1 and X_2 are positively associated

Mahalanobis matching within p-score calipers

- First calculate propensity scores
- For each treatment unit find all control units whose p-score are within a certain distance (e.g. .05 s.d.s) of the treated unit's propensity score
  - If only one is within this caliper, that is the match
  - If no units are within, choose the closest control unit (not in caliper) as the match
- Assuming more than one control unit is within the caliper, then choose the unit that is closest in terms of Mahalanobis distance with respect to just a few key variables

Diagnostics

- In each case, perform evaluations of balance in the same way that you estimate treatment effects (just replace each covariate for the outcome variables shown in the formulas)

How to decide among matching methods

- If you are just considering the best treatment effect estimate, you should use the option that creates the biggest sample without incurring greater bias
- This may take a closer examination of the distributions of pscores within your population and perhaps some shearing at the ends of the distributions
- If you want to get the variance right too life gets more complicated as we'll see
- Probably important to look at a range of estimates across different specifications

Conclusions

- Propensity score matching DOES NOT….

  solve the "omitted variable bias" or "selection bias" problem.

  - Recall that propensity score matching still relies on a very **strong assumption**. We need to assume that we have controlled for all the potential confounders.
- Propensity score matching DOES…
  - rely on weaker (i.e. more plausible) assumptions about the way that Y and X are related to each other (for instance as compared to linear regression)
- While implementation is easy to do in theory, in practice there are lots of choices to make (we'll see more soon) and not a lot of guidance in the literature about "best practice"