Week 8: Matching estimators and propensity scores

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- Defining treatment effects (ATE, ATET, ATEC)
- Identifying treatment effects
- Estimating treatment effects
- Diagnosing and dealing with lack of complete overlap
- Using the propensity score to deal with overlap issues (assuming ignorability):
  1. Stratification
  2. Matching
  3. Inverse probability weighting (IPW)
- Stratification, matching, and IPW under strong ignorability: alternatives to estimating treatment effects

Important

- We are assuming ignorability (no unmeasured confounders, etc)
- We will cover propensity scores as a way to 1) define and then 2) diagnose overlap problems
- The we will use propensity score matching (PSM), inverse probability weighting (IPW), and stratification as ways to solve overlap problems by restricting estimation to a region where overlap is better
- But they are also alternative ways of performing regression adjustment when strong ignorability holds (ignorability plus overlap)
- **This has important practical implications**. One of them being that in many cases, when overlap holds, we are going to get very similar results to regression adjustment, although some methods have additional advantages, like IPW having "doubly" robust properties
- But again, in terms of causal inference, NONE of these methods solves ignorability. **It must be assumed**

Defining treatment effects

- We defined causal effects as a **comparison of potential outcomes** for unit i and for a group of N units, which we could measure in terms of expected values, although we saw that this is more general: we could compare other quantities (median, odds ratios, etc):
  1. Average treatment effect (ATE):
     $E[Y_{1i}] - E[Y_{0i}] = \frac{1}{N}\sum_{i=1}^{N} Y_{1i} - \frac{1}{N}\sum_{i=1}^{N} Y_{0i} = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) = E[Y_{1i} - Y_{0i}]$
  2. Average treatment effect on the treated (ATET):
     $E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i Y_{1i} - \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i Y_{0i} = \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i(Y_{1i} - Y_{0i}) = E[Y_{1i} - Y_{0i}|D_i = 1]$
  3. Average treatment effect on the control (ATEC):
     $E[Y_{1i}|D_i = 0] - E[Y_{0i}|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 0]$
  4. $\frac{1}{\sum_{i=1}^{N}(1-D_i)}\sum_{i=1}^{N}(1 - D_i)Y_{1i}$
- The above expressions look esoteric but it's quite simple when you realize that $\sum_{i=1}^{N} D_i$ is the number of treated units (could denote it by N^T instead) and $D_i Y_{1i}$ is Y_{1i} for treated and zero for controls (same for Y_{0i})
- For ATEC, all would be $(1 - D_i)$ since we only want to include controls. So $\sum_{i=1}^{N}(1 - D_i) = N^C$ and $(1 - D_i)Y_i$ is Y_i for controls and zero for treated

Defining treatment effects

- We can also define the average treatment effect as a function of ATET and ATEC:

  $ATE = \frac{N^T}{N} ATET + \frac{N^C}{N} ATEC$
- We also saw that with randomization $ATE = ATET = ATEC$
- We can define treatment effects conditioning for covariates. We saw that this would give us estimates of causal effects in cases like conditional (block) randomization; randomization is based on the value of a covariate (or more than one)
- In that case, equations are similar, but we need to condition for the vector $\mathbf{X}$:

  $ATE = E[Y_{1i} - Y_{0i}|\mathbf{X}_i] = E[Y_{1i}|\mathbf{X}_i] - E[Y_{0i}|\mathbf{X}_i]$

  $ATET = E[Y_{1i} - Y_{0i}|\mathbf{X}_i, D_i = 1] = E[Y_{1i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 1]$

  $ATEC = E[Y_{1i} - Y_{0i}|\mathbf{X}_i, D_i = 0] = E[Y_{1i}|\mathbf{X}_i, D_i = 0] - E[Y_{0i}|\mathbf{X}_i, D_i = 0]$

Estimating treatment effects

- So far all we did is **define** treatment effects in an abstract way
- We can now discuss how to **estimate** treatment effects when we can argue that they are **identified** given data and experiment
- Remember, the fundamental problem is that for each unit i we only observe Y_{1i} or Y_{0i} but not both
- We link observed and potential outcomes with $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$, which we could rewrite in a simpler way: $Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i$
- For controls the **observed** outcome $Y_i = Y_{0i}$ and for treated units the observed outcome $Y_i = Y_{1i}$

Estimating treatment effects

- Under which circumstances a simple comparison of observed outcomes could give us estimates of treatment effects?
- We can decompose the **observed** conditional difference $E[Y_i|\mathbf{X}_i, D_1 = 1] - E[Y_i|\mathbf{X}_i, D_1 = 0]$ into two pieces:
  $E[Y_{1i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 1] + E[Y_{0i}|\mathbf{X}_i, D_i = 1] - E[Y_{0i}|\mathbf{X}_i, D_i = 0]$
- The first difference is the definition of $ATET$, the second one is the part we called the **selection bias**
- If the selection bias is zero, then a comparison of observed expected values is an estimate of treatment effects since $E[Y_i|\mathbf{X}_i, D_i = 0] = E[Y_{0i}|\mathbf{X}_i, D_i = 1]$
- The left-hand side is observed, the right-hand side is a potential outcome: the outcome for the treated group had they not been treated, which we don't observe. But if the observed outcome in the control group is the same as the unobserved outcome for the treated group had they not been treated, then the selection bias is zero
- In other words, the selection bias is zero when the control group provides a good prediction of what would have happened to the treated had they not been treated (and the treated is a counterfactual for the control), conditional on $\mathbf{X}$

Estimating treatment effects

- We called the main assumption relating selection bias being zero as **ignorability of treatment assignment** (or the conditional independence assumption, CIA, selection on observables, no unmeasured confounder):

  $(Y_{0i}, Y_{1i}) \perp D_i | \mathbf{X}_i$

- That is, treatment assignment, conditional on a vector of covariates $\mathbf{X}_i$, is independent of potential outcomes
- Now, this leaves us in a good place. If we can argue that the selection bias is zero, which is equivalent as saying that ignorability holds, all we have to do is find a statistical method to find two conditional expectation functions using **observed** data:

  $E[Y_i | \mathbf{X}_i, D_1 = 1]$ and $E[Y_i | \mathbf{X}_i, D_1 = 0]$

- Whether we need to condition on the vector $\mathbf{X}$ depends on the data generating process. Under simple randomization, D and $\mathbf{X}$ are independent, which makes them **mean independent** as well. So we could find treatment effects without having to condition on $\mathbf{X}$. Under conditional randomization, we do need to condition on $\mathbf{X}$
- And, of course, we need SUTVA

Big picture

- It's helpful to follow Heckman and Vytlacil (2007a) discussion on empirical research as three separate steps (paraphrasing to match our language)
  1. **Define** causal effects using potential outcomes
  2. **Identify** causal effects from a hypothetical population data and situation (experiment or natural experience or quasi-experiment)
  3. **Estimate** models [parametric, nonparametric] from observed samples
- The introduction to today's class follows this framework
- We will focus on step 3 now: estimate models
- So how do we estimate $E[Y_i|\mathbf{X}_i, D_1 = 1]$ and $E[Y_i|\mathbf{X}_i, D_1 = 0]$ assuming that causal effects are identified? (In other words, assuming ignorability holds)
- Those are two conditional expectation functions, but we can combine them: $E[Y_i|\mathbf{X}_i, D_i]$

Regression adjustment, parametric

- This is the old fashioned, vanilla linear/OLS regression model:
  $Y_i = \beta_0 + \beta_1 D_i + \mathbf{X}_i'\beta + \epsilon_i$ or $E[Y_i|\mathbf{X}_i, D_i] = \beta_0 + \beta_1 D_i + \mathbf{X}_i'\beta$
- **Can this model estimate causal effects even if they are identified**? Well, it depends
- First, the model must be correctly specific. That includes the assumption of homogeneous treatment effects. We could add interactions and try other model fits, but we never have certainty that the model is correctly specified (should we add quadratic terms, multiple interactions? Are effects additive and separable?)
- Second, the assumption that $Y_i \sim N(0, \sigma^2)$ could be wrong. So could be other assumptions about the model, like iid errors and homoskedasticity. We saw a bunch of alternatives: logit, probit, and GLMs in general. The regression model needs to consider characteristics of the data generating process, which is very important for inference (standard errors)
- Third, when we use observational data, we need to worry about the assumption we haven't mentioned yet: **overlap**. Implicitly, we **extrapolate information** from controls to treated and vice versa

Regression adjustment, semiparametric

- We also saw that we could estimate **nonparametric or semiparametric** models. This follows straight for the idea that we could estimate models separately for $E[Y_i|\mathbf{X}_i, D_1 = 1]$ and $E[Y_i|\mathbf{X}_i, D_1 = 0]$
- That's what we did with the command -teffects ra-
- The advantage of this method is that it runs stratified models, so it's equivalent to fully interacted models that take into account treatment heterogeneity
- They are also didactically great because it makes it explicit that estimating treatment effects is a **prediction problem**
- We used linear/OLS models, but we could have estimated $E[Y_i|\mathbf{X}_i, D_1 = 1]$ and $E[Y_i|\mathbf{X}_i, D_1 = 0]$ using other models
- Alternatives are logit, probit, GLMs, or even nonparametric or semiparametric models like kernel estimators or series estimators: commands -npregress kernel- and -npregress series-

Observational data and overlap

- Overlap: for all $\boldsymbol{X}_i \in \varphi$, where φ is the support (domain) of the covariates $\mathbf{X}_i$, $0 < P(D_i = 1|\boldsymbol{X}_i) < 1$
- This just means that for each variable in the vector $\mathbf{X}_i$, the probability of treatment should not be 0 (or close to 0) or 1 (or close to 1). Note that $\mathbf{X}_i$ could contain **interactions** between two or more variables
- This would rule out cases in which, say, treated units are old and control units are young. Something like $P(D_i = 1|age_i = 20) \approx 0$
- We don't worry about overlap with randomization because randomization guarantees that the **distribution** of covariates is the same in treated and control units. Note that this is a stronger result than just mean independence. Treated and controls could have the same mean (balance) but they could have bad overlap, affecting variance
- Overlap is tied to the concept of the **propensity score**. The propensity score is defined as $e(\mathbf{X}_i) \equiv P(D_i = 1|\boldsymbol{X}_i)$, the propensity to receive treatment

The propensity score and overlap

- We already saw that we can use the propensity score to diagnose overlap problems since we *define* overlap using the propensity score
- The propensity score is a summary score: if a group of control and a group of treated units have the same propensity score, then they have the same distribution of $\mathbf{X}$, where $\mathbf{X}$ are the variables used to estimate the propensity score (we will see more formally that it's also a balancing score)
- Once we diagnose the problem, we can use the propensity score to find a solution for the overlap problem. All are versions of the same idea: **restrict estimation to the region where there is good overlap**:
  1. Stratification by the propensity score
  2. Matching using the propensity score
  3. Inverse probability weighting (IPW) – with some restrictions
- IPW requires a bit more thought. If overlap doesn't hold, IPW wouldn't be defined for some observations. Recall that $IPW_i = \frac{1}{P(D_i=1|\mathbf{X}_i)}$ if $D_i = 1$ and $IPW_i = \frac{1}{1-P(D_i=1|\mathbf{X}_i)}$ if $D_i = 0$. So $P(D_i = 1|\mathbf{X}_i)$ can be 1 or 0 in the denominator (or very close to 1 or 0)

Data

- We will use a dataset from Gelman, Hill, and Vihtari (2020) (it's a fantastic book)
- Data for children born in the 80s, 290 received special services early in life; 4091 are controls. Children were targeted because they were born prematurely or had low birth weight (≤ 2500) and lived in an intervention city
- Outcome is a cognitive score (ppvtr36)

```
desc ppvtr36 bwg hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                  first bw preterm momage dayskidh

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------------
ppvtr36         float   %9.0g                 ppvtr.36
bw              byte    %8.0g
hispanic        byte    %8.0g
black           byte    %8.0g
bmarr           byte    %8.0g                 b.marr
lths            byte    %8.0g
hs              byte    %8.0g
ltcoll          byte    %8.0g
workdur         byte    %8.0g                 work.dur
prenatal        byte    %8.0g
male            byte    %8.0g
first           byte    %8.0g
bw              float   %9.0g
preterm         float   %9.0g
momage          byte    %8.0g
dayskidh        float   %9.0g
```

Checking balance

- We could compare means and standard deviations or any other metric as a typical Table 1 of any paper
- A convenient summary is to use standardized differences (normalized differences) and variance ratios
- **Standardized difference**: $\Delta_X = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_0^2 + S_1^2}}$
- **Variance ratio**: $\frac{S_1^2}{S_0^2}$
- Rule of thumb is that a standardized difference greater than 0.25 means that a regression model adjusting for covariates would be sensitive to model specification (because of lack of balance, overlap)
- No rule for variance ratios (ideally, close to 1). Differences in variances but good balance is not a major problem
- Note that the standardized difference is similar to the two-sample t-test: $T = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_0^2/N_0 + S_1^2/N_1}}$. Larger sample sizes would decrease T, but larger sample sizes would not make a difference in terms of model specification problems

Checking balance: -teffects-

- In Stata, we can use the -teffects- command to check for balance, but it's unfortunate that the only way to use it is to actually estimate a propensity score type of model or some other tool like matching
- We don't want to see the outcome when we try different models to check for balance
- There are some user-written commands out there but we will stick with -teffects- but will run it quietly
- We will also use a user-written command -coefplot- (type "findit coefplot" to install it) to display standardized differences and variance ratio plots

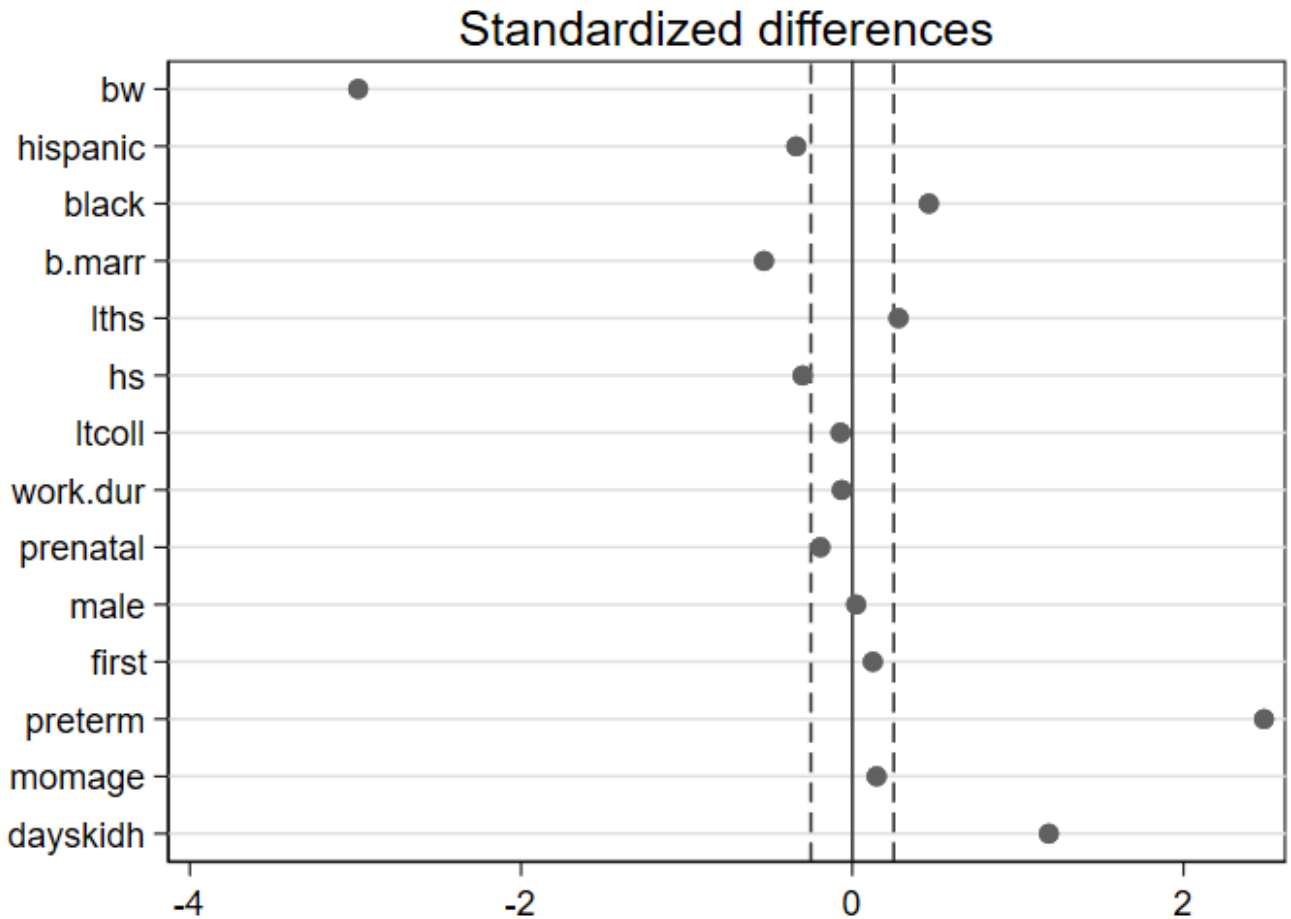
Balance

- Just once so you see what I mean. Note the pstolerance option. We are doing inverse probability weighting, but just because we want to check balance; we would get an error term because some propensity scores are close to zero

```
teffects ipw (ppvtr36) (treat bwg hispanic black bmarr lths hs ltcoll workdur prenatal ///
male first bw preterm momage dayskidh), pstolerance(1e-50)
tebalance summarize
------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw    Weighted           Raw   Weighted
  ----------------+-------------------------------------------------
              bw |  -2.983154   -1.864095      .2545874   .1523872
        hispanic |  -.3384636   -.1374384      .5046336   .7920301
           black |   .4620816    .1761349      1.235165   1.111173
           bmarr |  -.5327536   -.2559087      1.143919   1.113478
            lths |   .2789109    .1534797      1.171242   1.106074
              hs |  -.3002187   -.0130743      .8326499   .9966036
          ltcoll |  -.0712848     -.15487      .8901424   .7567162
         workdur |  -.0638442   -.0016862      1.031228   1.002031
        prenatal |  -.1927647   -.0887688      3.422968   1.948011
            male |   .0228178    .1038601      1.003106   .9970652
           first |   .1238746    .0217956      1.027565   1.009207
         preterm |    2.48568    1.312396      .9106964   .3954169
          momage |   .1467263    .1121158      3.477228   2.829429
        dayskidh |    1.18667     .867865      4.272903   2.632112
  ------------------------------------------------------------------
mat M = r(table)
coefplot matrix(M[,1]), noci xline(0) xline(-0.25 0.25, lpattern(dash)) title("Standardized differences")
graph export stdif.png, replace
coefplot matrix(M[,3]), noci xline(1) title("Variance ratios")
graph export var.png, replace
```

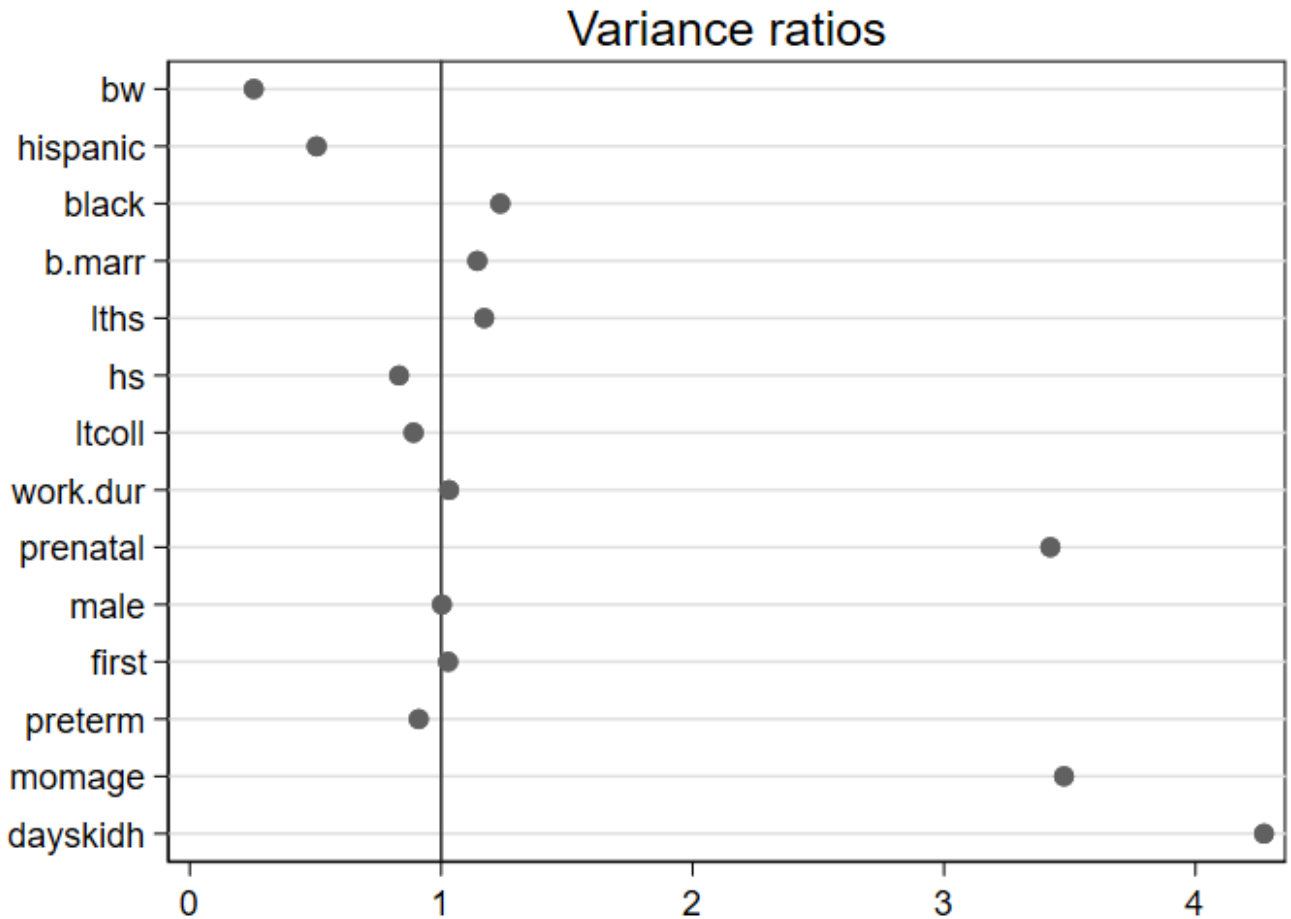
Balance

- Standardized differences



Balance

- Variance ratios



Balance vs overlap

- We have some balance problems between treated and controls in this dataset that would suggest regression adjustment would rely on extrapolation
- This likely translates into overlap problems, which can be due to one or more variables
- Next step is to check overlap using the propensity score since it's the definition of overlap

Overlap

- Using the propensity score to check overlap

```
. logit treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                 first preterm momage dayskidh, nolog

Logistic regression                             Number of obs     =      4,381
                                                LR chi2(14)       =    1406.34
                                                Prob > chi2       =     0.0000
Log likelihood =  -364.4063                     Pseudo R2         =     0.6587

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bw |  -.0044176   .0003203   -13.79   0.000    -.0050455   -.0037898
    hispanic |  -1.008019   .3283637    -3.07   0.002      -1.6516   -.3644382
       black |   .3852354   .2395514     1.61   0.108    -.0842768    .8547475
       bmarr |  -.6796435   .2310288    -2.94   0.003    -1.132452   -.2268353
        lths |  -.2753011   .4288705    -0.64   0.521    -1.115872    .5652696
          hs |  -1.233805   .4032258    -3.06   0.002    -2.024113   -.4434971
      ltcoll |  -1.008354   .4227533    -2.39   0.017    -1.836936   -.1797731
     workdur |   .2018587   .2149724     0.94   0.348    -.2194794    .6231969
    prenatal |  -.6206795   .5963921    -1.04   0.298    -1.789587    .5482276
        male |  -.0599874   .1943709    -0.31   0.758    -.4409474    .3209725
       first |   .5414952   .2146661     2.52   0.012     .1207574    .9622329
     preterm |   .3745637   .0495365     7.56   0.000     .2774739    .4716535
      momage |   .1053551   .0278844     3.78   0.000     .0507026    .1600076
    dayskidh |  -.0527067   .0101636    -5.19   0.000    -.0726271   -.0327864
       _cons |   6.591678   1.424636     4.63   0.000     3.799443    9.383913
------------------------------------------------------------------------------
Note: 4 failures and 0 successes completely determined.

predict double ps if e(sample)
(option pr assumed; Pr(treat))
```

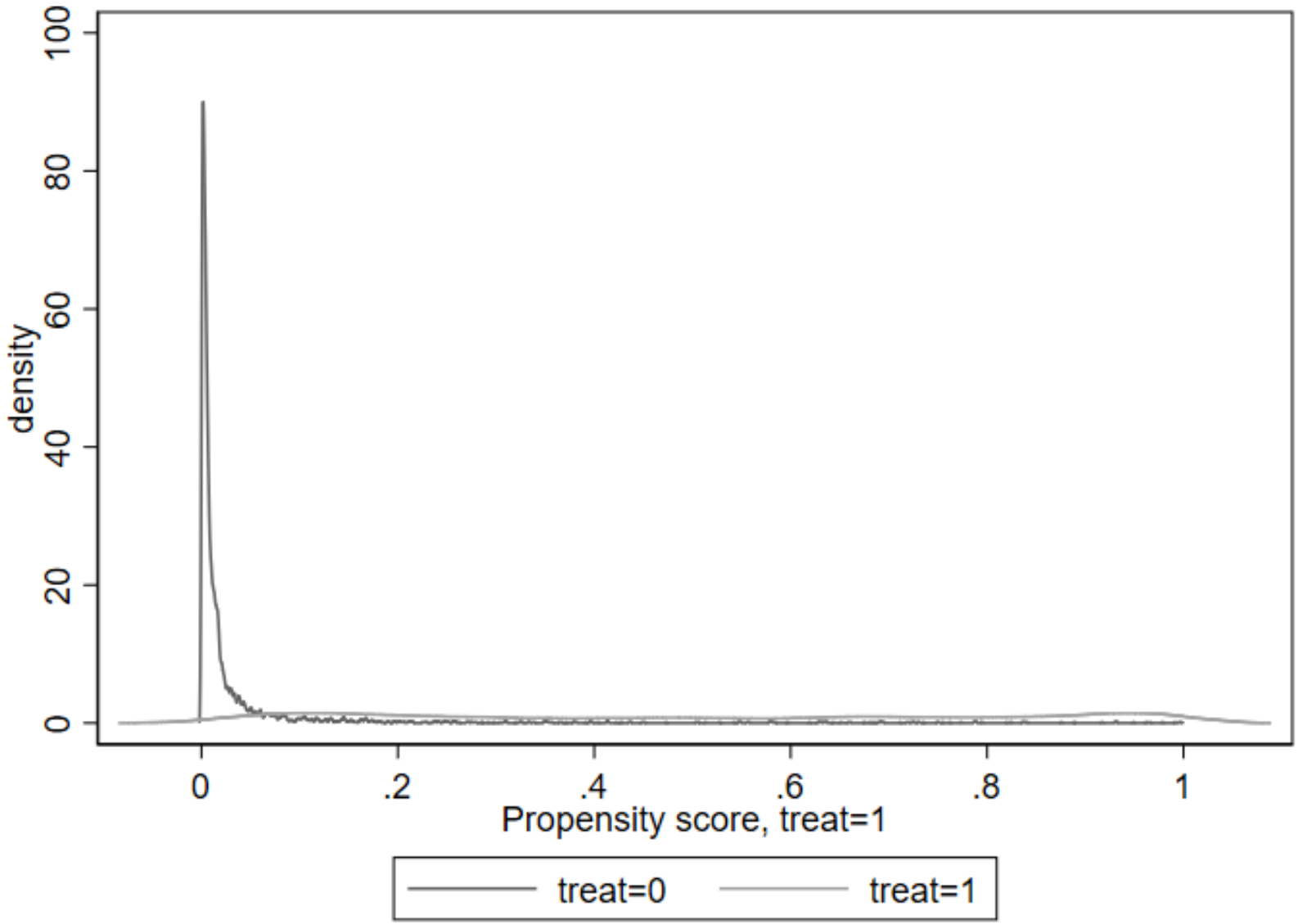
Overlap

- Remember that the propensity score is a summary score
- The region of overlap is [0.0181568, 0.9828839]. Note that in the control some scores are essentially zero – no changes of being treated

```
tabstat ps, by(treat) stats(N mean median min max)

Summary for variables: ps
     by categories of: treat

   treat |         N      mean       p50       min       max
---------+--------------------------------------------------
       0 |      4091  .0274727  .0007289  5.28e-12  .9828839
       1 |       290  .6124459  .6476379  .0181568   .995504
---------+--------------------------------------------------
   Total |      4381  .0661949  .0009125  5.28e-12   .995504
------------------------------------------------------------

qui teffects ipw (ppvtr36) ///
         (treat  bwg hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first preterm momage dayskidh), pstolerance(1e-50)
teffects overlap, ptl(1)
graph export overl.png, replace
```

Distribution of propensity scores

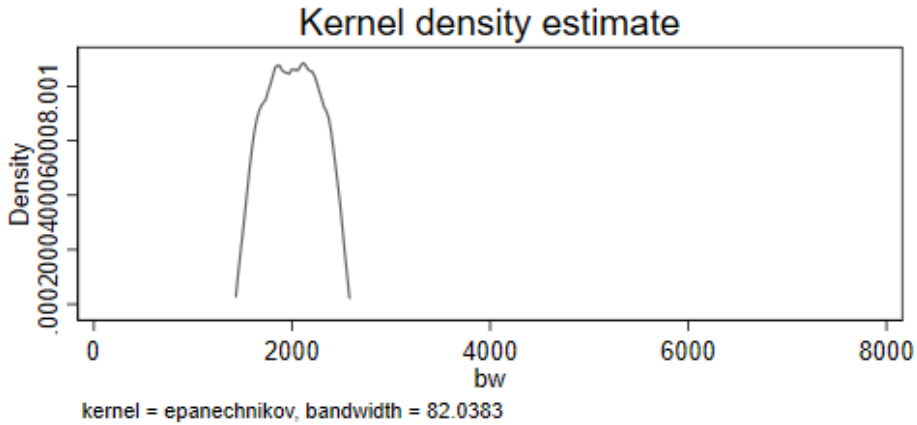
- Clearly, some controls have small changes of being treated

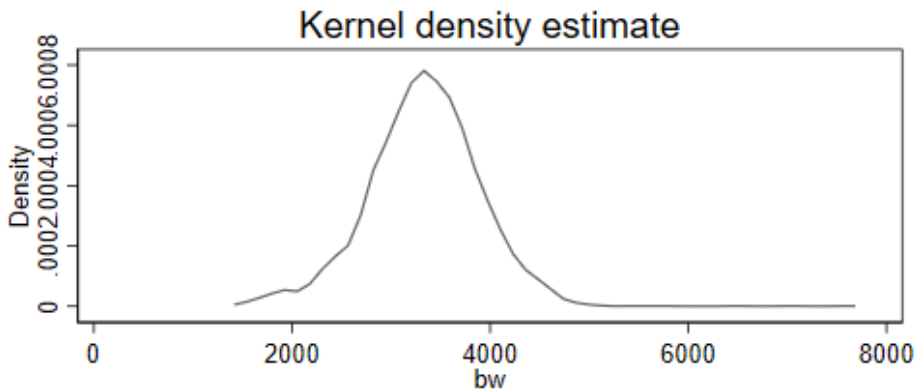


Check birth weight

- Check the distribution in birth weight for treated and control, but standardize difference suggest other variables are problematic, like not having prenatal care

```
kdensity bw if treat ==1, saving(tkden.gph, replace)
kdensity bw if treat ==0, saving(ckden.gph, replace)
graph combine tkden.gph ckden.gph, col(1) xcommon xsize(10) ysize(10)
graph export den.png, replace
```





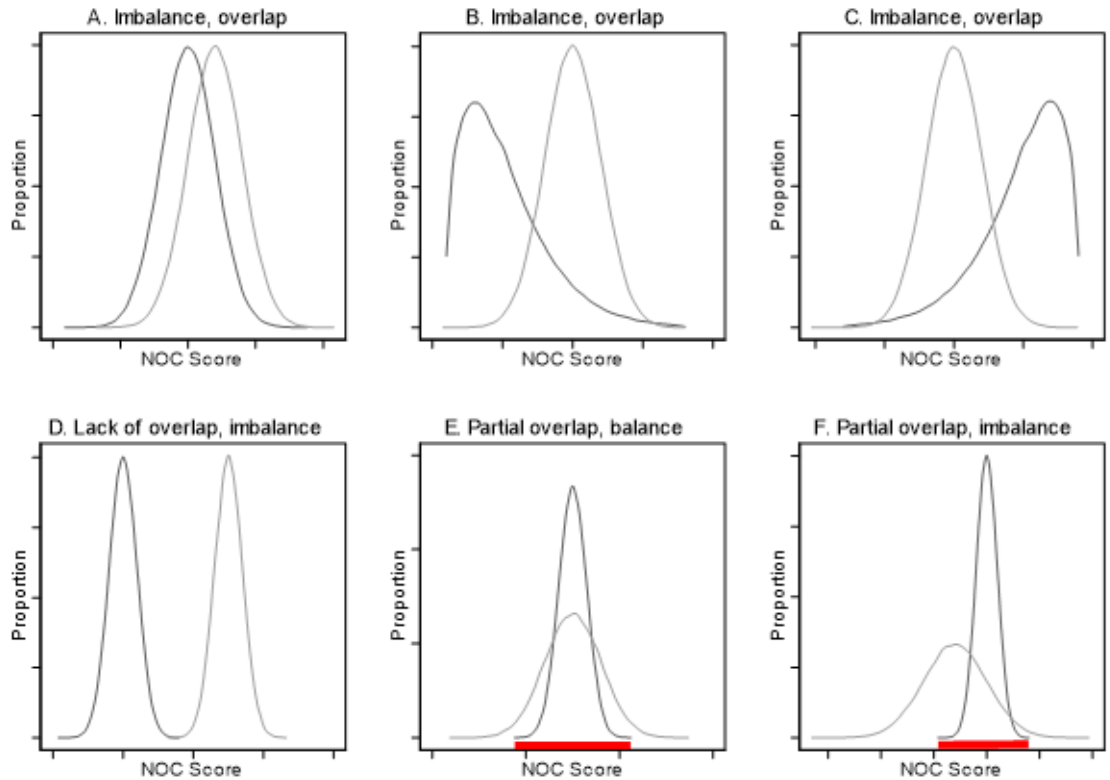
Balance versus overlap

- Lack of balance is not as serious unless lack of balance is serious enough – rule of thumb is 0.25 standardized difference
- Lack of overlap is more important. We could check one variable at a time of we could check the propensity score since the propensity score is a summary score
- We could also try to estimate a better model for the propensity score, say with interactions (more on this in a sec)

Balance versus overlap

- Standardized differences

Figure 1. Balance and overlap



Simulated distribution of NOC scores. Panels A, B, and C show situations in which there is complete overlap but lack of balance (both average and standard deviation are different). Panel D shows distributions with both lack of balance and overlap. Panel E shows distributions with balance in means but lack of complete overlap. Panel F shows both lack of balance and overlap. The thick lines on x-axis in panels E and F show regions of overlap. Propensity scores could be used to analyze E and F, but not D. Regression adjustment and propensity scores would yield similar estimates in panels A to C, since there is complete overlap. NOC: Nursing Outcome Classification.

Regression

- To make things more concrete. The issue is that the model below is probably not the best, and we haven't even dealt with model specification or residual analysis

```
. reg ppvtr36 treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                 first  preterm momage dayskidh, robust

Linear regression                               Number of obs     =      4,381
                                                F(15, 4365)       =     158.59
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3356
                                                Root MSE          =     16.428

------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   11.59116   1.227529     9.44   0.000     9.184584    13.99775
          bw |   .0005624   .0005151     1.09   0.275    -.0004475    .0015723
    hispanic |  -13.74123    .729361   -18.84   0.000    -15.17115   -12.31131
       black |   -17.2159    .640063   -26.90   0.000    -18.47075   -15.96106
       bmarr |    3.00947    .615206     4.89   0.000     1.803354    4.215586
        lths |  -14.59204   1.043248   -13.99   0.000    -16.63733   -12.54674
          hs |   -8.47883   .9122019    -9.29   0.000    -10.26721   -6.690451
      ltcoll |  -6.393914   .9666583    -6.61   0.000    -8.289055   -4.498773
     workdur |   2.820732   .5621512     5.02   0.000     1.718631    3.922834
    prenatal |   4.357118   2.219644     1.96   0.050     .0054882    8.708748
        male |   1.170581   .5042633     2.32   0.020     .1819689    2.159193
       first |   4.604955   .5528963     8.33   0.000     3.520998    5.688913
     preterm |   .0102207   .1408463     0.07   0.942    -.2659095    .2863509
      momage |    .167805   .0886327     1.89   0.058    -.0059601    .3415701
    dayskidh |  -.1446362   .0513661    -2.82   0.005    -.2453397   -.0439326
       _cons |   87.15683     3.8615    22.57   0.000     79.58633    94.72733
------------------------------------------------------------------------------
```

Matching

- One way to restrict the estimation to the region where there is overlap would be to find, for each treated unit, control units that are "similar" in their covariates
- If we do something like that, then the resulting sample would have good balance and overlap. The target of estimation will then be ATET. We are finding control units that are similar to the treated units to predict (impute) the counterfactual Y_{0i} for each unit i with $D_i = 1$
- We just need to find a way to measure **similar** using multiple variables (easier for few variables, like age and sex). It would make sense to use the propensity score as a measure of similarity
- Remember the main result of propensity scores: if a group of treated and control observations have the same propensity score then they have the same distribution of the covariates that entered into the estimation of the propensity score
- (We will see other ways of matching. The propensity score may not be the best, actually)

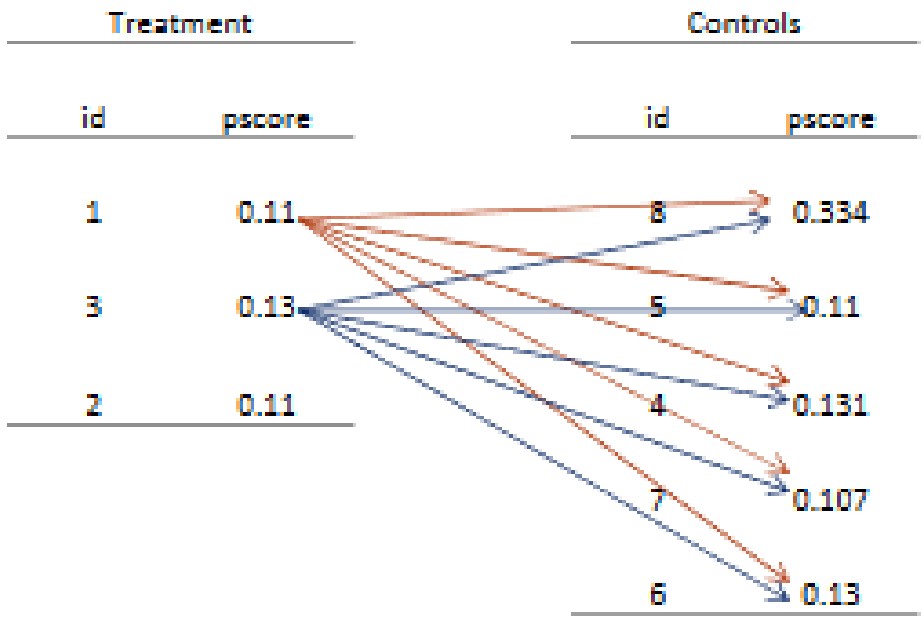
Many ways of matching, many ways of getting confused

- Over the years, many variants of matching have been proposed. And there are many decisions one can make with matching. Main issues:
  1. **Measure of similarity**: propensity score, Malahanobis, other metrics based on variance ("exact matching" could fit here)
  2. **Replacement or not**: Once a treated unit is matched with a control unit, can the control unit be a match for another treated unit? If no, then without replacement. If yes, with replacement
  3. **Number of matches**: 1 to 1, 1 to N or variable? If 1 to 1, usually called **pair matching**. Nearest-neighbor matching with replacement is common: For each treated unit, find the k closest observations (we define k a priori) in the control group. A control can be used multiple times. In case of ties, use all ties as matches
  4. **Caliper matching (radius)**: Use only controls with a distance smaller than a number c, the "caliper" (tries to avoid bad matches)
- Even more, we could also use different algorithms to perform the match: greedy, optimal, "genetic" algorithms
- We will focus on common ones and the ones that Stata implemented: commands -teffects psmatch- and -teffects nnmatch-

Classic: 1 to 1 matching without replacement, nearest neighbor (pair matching)

- Simple algorithm (sometimes called "greedy" algorithm)
  1. Sort treated units randomly
  2. For the first treated unit $i = 1$, calculate the absolute difference between i's propensity score and each of the control units' propensity scores
  3. Match $i = 1$ to the control unit with the smallest absolute difference
  4. Remove the matched control from the pool of potential controls
  5. Repeat for $i = 2$
- The result will be a dataset with $N^T \times 2$ observations
- Different implementations have different options for number of matches. With enough controls, a 1:1 match would discard too many observations

Matching



Treatment id = 1

		Abs Difference
0.11	0.334	0.22
0.11	0.11	0.00
0.11	0.131	0.02
0.11	0.107	0.00
0.11	0.13	0.02

Matches: 1 to 5; 3 to 6; 2 to 7

1:1 Matching, no replacement

- We will use the user-written command -psmatch2-
- Type ssc install psmatch2, replace
- The command performs different types of matching including some that are similar to the ones in -teffects psmatch- and -teffects nnmatch- but not exactly the same
- We will only use it for 1:1 matching without replacement
- Matched sample will be $290 \times 2 = 580$

1:1 Matching, no replacement

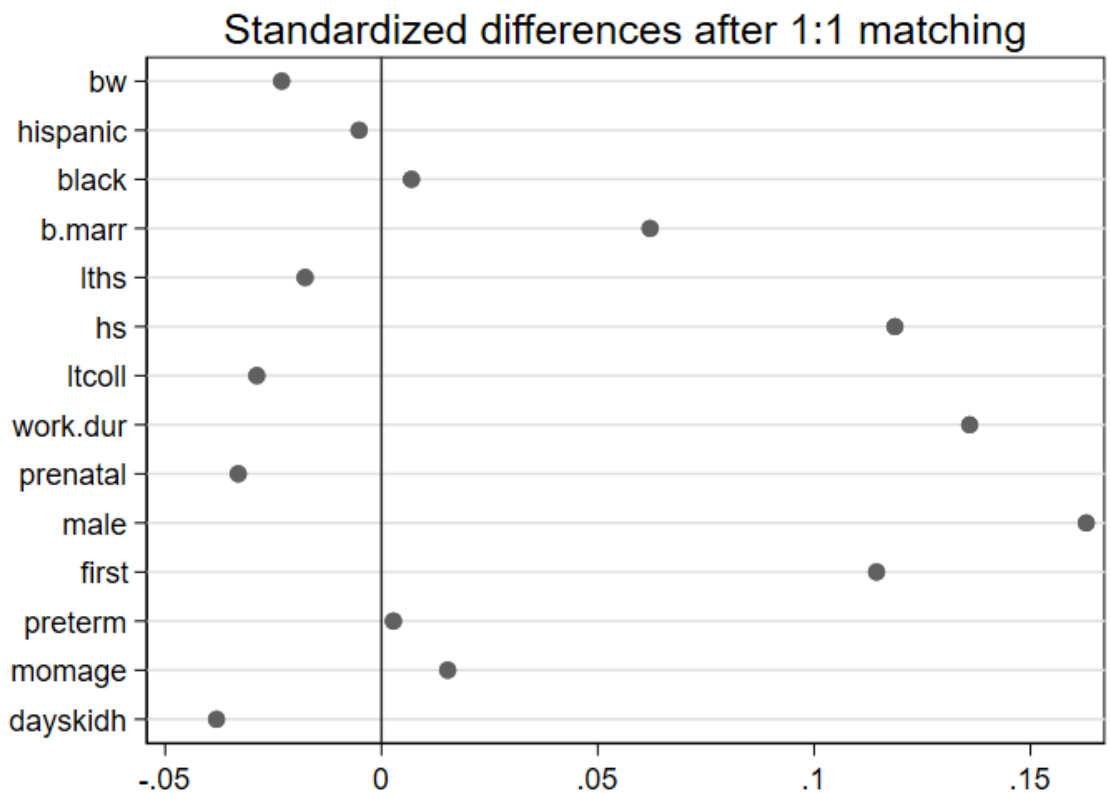
```
. psmatch2 treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                 first preterm momage dayskidh, n(1) logit out(ppvtr36) noreplacement

Logistic regression                             Number of obs     =      4,381
                                                LR chi2(14)       =    1406.34
                                                Prob > chi2       =     0.0000
Log likelihood =  -364.4063                     Pseudo R2         =     0.6587

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          bw |  -.0044176   .0003203   -13.79   0.000    -.0050455   -.0037898
    hispanic |  -1.008019   .3283637    -3.07   0.002      -1.6516   -.3644382
       black |   .3852354   .2395514     1.61   0.108    -.0842768    .8547475
...
...
 ------------------------------------------------------------------------------
Note: 4 failures and 0 successes completely determined.
-------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+-------------------------------------------------------
         ppvtr36  Unmatched | 92.1137901   86.0280498   6.08574029   1.21935202     4.99
                        ATT | 92.1137901   81.6837432   10.4300469    1.6297464     6.40
-----------------------------+-------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.
tab treat _weight

           | psmatch2:
           | weight of
           |  matched
           |  controls
     treat |         1 |     Total
-----------+-----------+----------
         0 |       290 |       290
         1 |       290 |       290
-----------+-----------+----------
     Total |       580 |       580
```

1:1 Matching, no replacement

```
* Replicate
* Raw, unmatched
reg ppvtr36 treat
      Source |       SS           df       MS      Number of obs   =     4,381
-------------+----------------------------------   F(1, 4379)      =     24.91
       Model |  10029.5409         1  10029.5409   Prob > F        =    0.0000
    Residual |  1763142.34     4,379   402.63584   R-squared       =    0.0057
-------------+----------------------------------   Adj R-squared   =    0.0054
       Total |  1773171.88     4,380  404.833763   Root MSE        =    20.066

------------------------------------------------------------------------------
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |    6.08574   1.219352     4.99   0.000     3.695193    8.476287
       _cons |   86.02805   .3137195   274.22   0.000       85.413     86.6431
------------------------------------------------------------------------------

* ATE, matched
reg ppvtr36 treat if _weight== 1

      Source |       SS           df       MS      Number of obs   =       580
-------------+----------------------------------   F(1, 578)       =     40.96
       Model |  15773.9524         1  15773.9524   Prob > F        =    0.0000
    Residual |  222605.505       578  385.130631   R-squared       =    0.0662
-------------+----------------------------------   Adj R-squared   =    0.0646
       Total |  238379.457       579  411.708907   Root MSE        =    19.625

------------------------------------------------------------------------------
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   10.43005   1.629746     6.40   0.000       7.2291    13.63099
       _cons |   81.68374   1.152405    70.88   0.000     79.42033    83.94715
------------------------------------------------------------------------------
```

Check balance

```
* Check balance (make sure you understand this; just using teffects to calculate
*  balance statistics)
qui teffects psmatch (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first  preterm momage dayskidh) if _weight ==1, nneighbor(1)
tebalance summarize
mat M = r(table)
coefplot matrix(M[,2]), noci xline(0) xline(-0.25 0.25, lpattern(dash)) title("Standardized differences after 1:1 matching")
graph export stdif_m.png, replace
```



Check balance

```
. sum bw hispanic black bmarr momage dayskidh if treat ==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          bw |        290    2008.648    283.3048       1515       2500
    hispanic |        290    .0931034    .2910796          0          1
       black |        290    .5034483    .5008524          0          1
       bmarr |        290    .4310345     .496077          0          1
      momage |        290    24.44483     5.87341         13         41
-------------+---------------------------------------------------------
    dayskidh |        290    14.68621    11.28376          1         71

. sum bw hispanic black bmarr momage dayskidh if treat ==0 & _weight ==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          bw |        290    2240.512    326.1465    1502.55    3033.45
    hispanic |        290    .1482759    .3559875          0          1
       black |        290    .4310345     .496077          0          1
       bmarr |        290    .5068966    .5008167          0          1
      momage |        290    23.62759    3.353343         17         31
-------------+---------------------------------------------------------
    dayskidh |        290    10.44443    13.77704          0        100
```

Matches are not identical

- This is important to understand the propensity scores. **Two matched units may have different covariate values**. On average, matched units are similar
- Below are two matches with their propensity score differing by only 0.00027

```
. list bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                 first preterm momage dayskidh _pdif if _id==4156 | _id == 3989

     +-------------------------------------------------------------------------+
  2. |      bw | hispanic | black | bmarr | lths | hs | ltcoll | workdur | prenatal |
     |    2240 |        0 |     1 |     0 |    1 |  0 |      0 |       0 |        1 |
     |---------+----------+-------------------------------------------------------|
     |  male   |  first   |   preterm   |   momage   |   dayskidh   |       _pdif   |
     |     1   |      0   |         3   |       22   |          4   |   .00027962   |
     +-------------------------------------------------------------------------+

     +-------------------------------------------------------------------------+
3659. |      bw | hispanic | black | bmarr | lths | hs | ltcoll | workdur | prenatal |
     | 2182.95 |        0 |     0 |     1 |    1 |  0 |      0 |       1 |        1 |
     |---------+----------+-------------------------------------------------------|
     |  male   |  first   |   preterm   |   momage   |   dayskidh   |       _pdif   |
     |     0   |      0   |         7   |       18   |         14   |           .   |
     +-------------------------------------------------------------------------+
```

Caveats

- There are probably thousands of studies that have used some version of the above analysis, but there are many problems with this strategy
  1. Standard errors do not take into account that the propensity score has been estimated (bootstrapping was the usual solution, but turns out that it doesn't quite work)
  2. Not very efficient since we discard thousand of potential controls (could do 1:N matching instead)
- The above issues are important and remember that this is an iterative process. Try different models, check balance. Choose the best approach that balances data
- Other important issues:
  1. As we saw, the propensity score balances on average, but other distance metrics could be better (i.e. Malahanobis)
  2. We could for example mimic conditional randomization by using other covariates to block
- There are strong arguments against PSM. For example, subtle papers like King and Nielsen (2019) "Why PSM Should Not Be Used for Matching"

The propensity score model

- Before we continue, we need to discuss the propensity score model itself
- So far, we have been estimating a simple one, but the specification of the propensity score matters. Usual suggestions:
  1. Include confounders. No need to include variables correlated with just the outcome or just the treatment
  2. Start with a simple (parsimonious model)
  3. If balance not acceptable, consider quadratic terms, categorizing continuous variables, interactions
- Iterative process. Repeat 3)
- Careful with empty cells in some cases (low sample sizes in some interactions)
- Many decisions: remember, you want balance and good overlap, not the decision that produces the result you want
- When lack of overlap is severe, may need to discard observations by restricting estimation to the region where there is overlap. Several options have been propose, like cardinality matching (Visconti and Zubizarreta, 2018)

Ways to restrict to overlap region

- From Visconti and Zubizarreta (2018)

Algorithm 1 Matching with standard matching methods.

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).
4. Assess covariate balance.

Until:

The matched sample satisfies the covariate balance requirements.

Algorithm 2 Matching with cardinality matching.

0. Specify the covariate balance requirements (e.g., mean balance).
1. Find the largest matched sample that satisfies the covariate balance requirements.
2. Rematch the matched sample to minimize covariate distances between matched units.

Stata's -teffects psmatch-

- Stata's -teffects psmatch- command implements a different version of matching
- It performs a k nearest neighbor matching in which treated units are matched with **at least** k controls with replacement
- There is no check on overlap region, so one must be careful
- Stata does check for propensity scores close to zero

-teffects psmatch-

- As we saw before, some propensity scores are too low. We can force teffects to continue by increasing the tolerance value (pstolerance option)

```
* Error
teffects psmatch (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first  preterm momage dayskidh) , nneighbor(1)
there are 232 propensity scores less than 1.00e-05
treatment overlap assumption has been violated; use the osample() option to identify the
observations
r(459);

teffects psmatch (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll workdur prenatal
>  male  ///
>                 first  preterm momage dayskidh) , nneighbor(1)  pstolerance(1e-50)

Treatment-effects estimation                   Number of obs      =      4,381
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: logit                                     max =          1
------------------------------------------------------------------------------
             |               AI Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   .3531638   1.428366     0.25   0.805    -2.446383     3.15271
------------------------------------------------------------------------------
```

-teffects psmatch-

- Note that the balance is not good at all. The lack of overlap problem is severe in this dataset

```
tebalance summarize
note: refitting the model using the generate() option

  Covariate balance summary
                                                 Raw      Matched
                          ----------------------------------------
                          Number of obs =       4,381        8,762
                          Treated obs   =         290        4,381
                          Control obs   =       4,091        4,381
                          ----------------------------------------

  ------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw     Matched           Raw    Matched
  ----------------+-------------------------------------------------
              bw |  -2.983154   -1.827003      .2545874    .054266
       hispanic |  -.3384636   -.6073668      .5046336   .1371829
          black |   .4620816   -.6451967      1.235165   .2753288
          bmarr |  -.5327536    .6834454      1.143919   .2889577
           lths |   .2789109   -.6155638      1.171242   .3297026
             hs |  -.3002187    1.148684      .8326499   .3932727
         ltcoll |  -.0712848   -.5759349      .8901424   .1280602
        workdur |  -.0638442    .7952702      1.031228    .291495
       prenatal |  -.1927647    .0930047      3.422968   .3300719
           male |   .0228178    1.035419      1.003106   .3176788
          first |   .1238746    -.926912      1.027565     .23562
        preterm |    2.48568    .9517999      .9106964   .2013297
         momage |   .1467263    1.478464      3.477228   .7966889
       dayskidh |    1.18667    3.104121      4.272903   1.269045
  ------------------------------------------------------------------
```

-teffects psmatch-

- Need to restrict to a region of overlap. Could use the propensity score, although we could use other strategies, including trimming based on bw for example
- We will use the propensity score for now. Note that we drop some treated units

```
capture drop ps
qui logit treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first preterm momage dayskidh, nolog
predict double ps if e(sample)
tabstat ps, by(treat) stats (N min max)

Summary for variables: ps
     by categories of: treat

   treat |         N       min       max
---------+------------------------------
       0 |      4091  5.28e-12  .9828839
       1 |       290  .0181568   .995504
---------+------------------------------
   Total |      4381  5.28e-12   .995504
----------------------------------------


gen keep = 1 if ps >= .0181568  & ps <= .9828839
tab treat keep

           |    keep
     treat |         1 |     Total
-----------+-----------+----------
         0 |       572 |       572
         1 |       283 |       283
-----------+-----------+----------
     Total |       855 |       855
```

-teffects psmatch-

- Need to restrict to a region of overlap. Could use the propensity score, although we could use other strategies, including trimming based on bw for example
- We will use the propensity score. Note that we drop some treated units

```
capture drop ps
qui logit treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first preterm momage dayskidh, nolog
predict double ps if e(sample)
tabstat ps, by(treat) stats (N min max)
gen keep = 1 if ps >= .0181568  & ps <= .9828839
tab treat keep
```

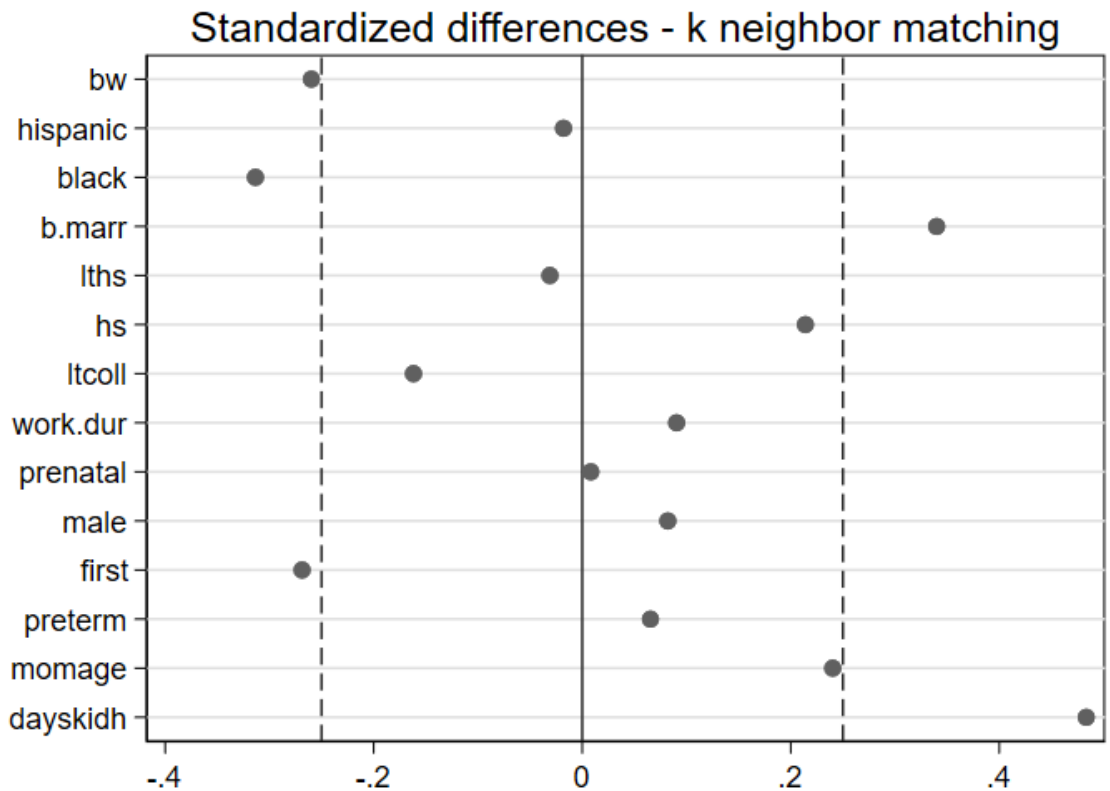
```
           |    keep
     treat |         1 |     Total
-----------+-----------+----------
         0 |       572 |       572
         1 |       283 |       283
-----------+-----------+----------
     Total |       855 |       855
```

-teffects psmatch-

```
teffects psmatch (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll workdur prenatal  male  ///
    first  preterm momage dayskidh) if keep ==1, nneighbor(1)
Treatment-effects estimation                   Number of obs      =        855
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: logit                                     max =          1
------------------------------------------------------------------------------
             |               AI Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   8.045777   1.259327     6.39   0.000     5.577541    10.51401
------------------------------------------------------------------------------

tebalance summarize
...
  ------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw     Matched           Raw    Matched
  ----------------+-------------------------------------------------
              bw |  -1.394722   -.2597216      .6006019   .4301904
        hispanic |  -.0650134   -.0179936      .8470438   .9587302
           black |   .0698415   -.3134383      1.008199   .8622675
           bmarr |   -.141656    .3399337      .9847759   .9090142
            lths |   .0908845   -.0310145      1.033483   .9874854
              hs |  -.1289964    .2140523       .904266   1.097288
          ltcoll |  -.0278215   -.1618443      .9542544   .7026459
         workdur |    .005303    .0905067       1.00003   .9662639
        prenatal |  -.1223482    .0082554      1.980514   .9455728
            male |   -.076909    .0820815        1.0083   .9888843
           first |   .0311325   -.2686938      1.004758   .8823296
         preterm |   1.061327    .0654548      .5034276   .3864448
          momage |    .100821    .2402564      2.852832   2.220045
        dayskidh |   .6435155     .483438      1.107973   .8175438
  ------------------------------------------------------------------

. mat M = r(table)
```

-teffects psmatch-

```
mat M = r(table)
coefplot matrix(M[,2]), noci xline(0) xline(-0.25 0.25, lpattern(dash)) title("Standardiz ed differences - k neighbor matching")
graph export stdifk.png, replace
```



-teffects psmatch-

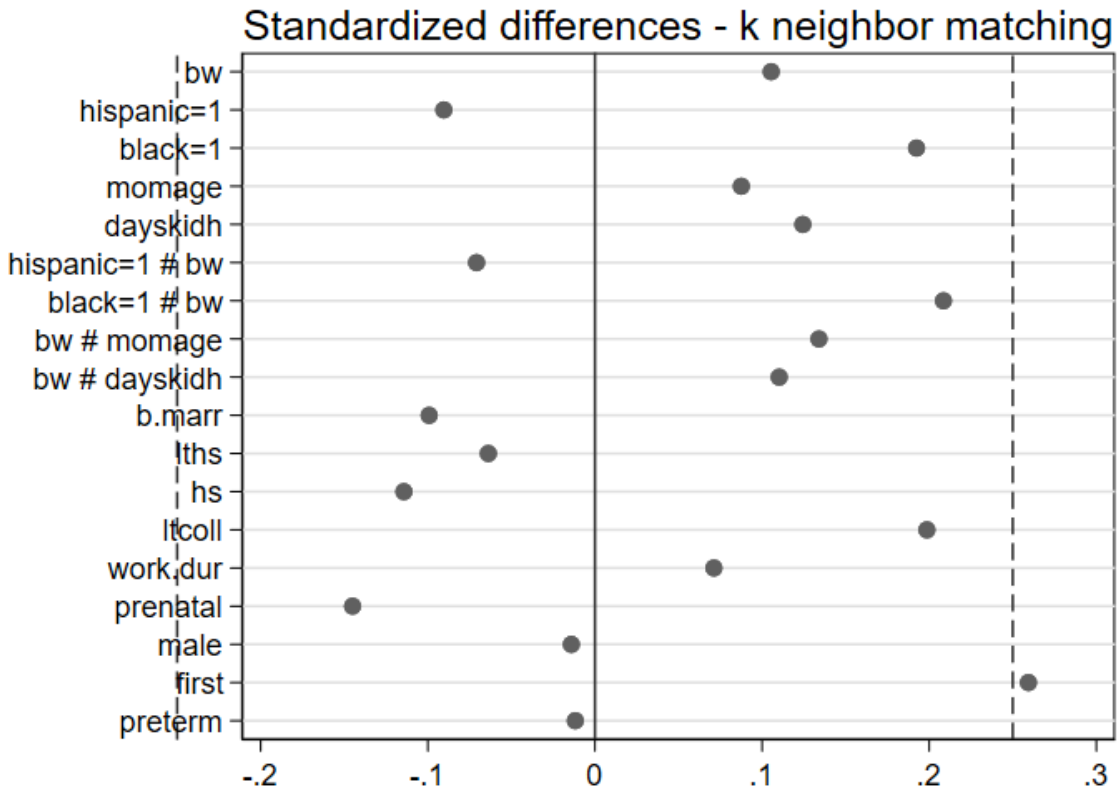
- We could try other specifications of the propensity score to see if balance improves
- Still some problems with days in hospital, but nothing extreme
- Interesting enough, similar results to regression adjustment
- As I said before, lack of overlap doesn't automatically means that regression adjustment is wrong

Different propensity score model

```
teffects psmatch (ppvtr36) (treat c.bw##(i.hispanic i.black c.momage c.dayskidh ) bmarr lths ///
         hs ltcoll workdur prenatal male first preterm ) if keep ==1, nneighbor(1)

Treatment-effects estimation                   Number of obs      =        855
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                     min =          1
Treatment model: logit                                        max =          1
------------------------------------------------------------------------------
             |               AI Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   7.564289   2.020666     3.74   0.000     3.603857    11.52472
------------------------------------------------------------------------------

qui tebalance summarize
mat M = r(table)
coefplot matrix(M[,2]), noci xline(0) xline(-0.25 0.25, lpattern(dash)) title("Standardized differences - k neighbor matching")
graph export stdifk_int.png, replace
```

Different propensity score model

- Much better balance. We should explore other models to detect overlap region as well
- As I said, this is an **iterative process**



Malahanobis

- Instead of the propensity score as a metric of similarity, we could use another metric. One is the Malahanobis distance (Rubin, 1980)
- Malahanobis is simply a measure of the distance between two vectors of data: $M(\mathbf{X_1}, \mathbf{X_2}) = \sqrt{(\mathbf{X_1} - \mathbf{X_2})\Sigma^{-1}(\mathbf{X_1} - \mathbf{X_2})}$
- Σ is the covariance matrix. If Σ is the identity matrix, then Malahanobis is the Euclidean distance
- Euclidean distance between (y_1, x_1) and (y_2, x_2) is $d(y, x) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ (Pythagorean theorem)
- If $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors of data, a smaller $M(\mathbf{X_1}, \mathbf{X_2})$ implies that observations are more similar in covariates values $\mathbf{X}$
- So as the propensity score, Malahanobis can be used as a measure of similarity, with the advantage that the matched observations are going to be more closely matched, **not just matched on average**

Malahanobis

```
teffects nnmatch (ppvtr36 bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first  preterm momage dayskidh ) (treat),  nneighbor(1)
tebalance summarize

Treatment-effects estimation                   Number of obs      =      4,381
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
------------------------------------------------------------------------------
             |               AI Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   8.886109   1.474341     6.03   0.000     5.996453    11.77576
------------------------------------------------------------------------------

tebalance summarize
  ------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw     Matched           Raw    Matched
  ----------------+-------------------------------------------------
              bw  |  -2.983154   -2.368282      .2545874   .1437058
         hispanic |  -.3384636   -.2499663      .5046336   .6221699
            black |   .4620816     .192254      1.235165   1.136648
            bmarr |  -.5327536   -.1851049      1.143919   1.100719
             lths |   .2789109    .0854561      1.171242   1.063502
               hs |  -.3002187   -.0685629      .8326499   .9719941
           ltcoll |  -.0712848    -.019315      .8901424   .9694677
          workdur |  -.0638442     .018842      1.031228   .9904492
         prenatal |  -.1927647     .001895      3.422968   .9848435
             male |   .0228178    .0744934      1.003106    .995244
            first |   .1238746    .0809175      1.027565   1.018515
          preterm |    2.48568    1.724164      .9106964   .5250509
           momage |   .1467263    .0970095      3.477228   1.684246
         dayskidh |    1.18667    .6236645      4.272903   1.185832
  ------------------------------------------------------------------
```

Malahanobis

1. As before, we need to restrict region of overlap. Let's use bw instead of the propensity score

```
tabstat bw, by(treat) stats(N mean sd min max)

Summary for variables: bw
     by categories of: treat

   treat |         N      mean        sd       min       max
---------+--------------------------------------------------
       0 |      4091  3335.268  561.4815   1502.55    7597.8
       1 |       290  2008.648  283.3048      1515      2500
---------+--------------------------------------------------
   Total |      4381  3247.453  639.1361   1502.55    7597.8
------------------------------------------------------------

gen keep1 = 1 if bw >= 1500 & bw <= 3000
tab treat keep1

           |   keep1
     treat |         1 |     Total
-----------+-----------+----------
         0 |     1,030 |     1,030
         1 |       290 |       290
-----------+-----------+----------
     Total |     1,320 |     1,320
```

Malahanobis

```
teffects nnmatch (ppvtr36 bw hispanic black bmarr lths hs ltcoll workdur prenatal male first  ///
preterm momage dayskidh ) (treat) if keep1 ==1,  nneighbor(1)

Treatment-effects estimation                   Number of obs      =      1,320
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
------------------------------------------------------------------------------
             |              AI Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   10.29988   1.368308     7.53   0.000     7.618042    12.98171
------------------------------------------------------------------------------
```

```
tebalance summarize
```

	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
bw	-2.008488	-1.272001	.724186	.5429634
hispanic	-.2688362	-.1046097	.5603778	.8118967
black	.257156	.0430653	1.067346	1.015228
bmarr	-.3324808	-.1080212	1.020362	1.017085
lths	.1930087	.0563812	1.096464	1.029957
hs	-.2949933	-.0561583	.8333505	.9728451
ltcoll	-.0573106	-.0039269	.909349	.9935394
workdur	.0242883	.0092262	.9942454	.9967681
prenatal	-.112584	.0045582	1.812482	.9744437
male	-.0736482	-.0424846	1.010007	1.004785
first	.0705086	.0091229	1.012425	1.001612
preterm	1.649226	1.112493	.5965296	.465296
momage	.1905657	.0955412	3.280781	1.815981
dayskidh	.8504919	.4556944	1.585644	.9601355

Where are we?

- Many different ways of matching, many decisions that can affect results. No clear answers on the best strategy
- If you this about it, we could have restricted the estimation to the region of overlap and then run a regression model
- Knowledge about the subject is important when deciding what should be carefully balanced. And all depends on the dataset. In this dataset, there is a **severe** overlap problem, mostly birth weight
- Many approaches are reasonable – we found similar results
- With -teffects nnmatch- we could force an exact match with the ematch(varlist) option on some variables (or by creating categorical variables)
- But don't lose track of big picture: **the goal is to restrict estimation to a region where comparisons are possible, and then make those comparisons**
- Careful that R, SAS, Stata implement different versions of matching

Matching as an imputation and weighting scheme

- One way to frame matching is that we are **imputing (predicting)** the conterfactual by assigning weights to units
- In a general way, we can write:
  $ATE_{matched} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_{1i} - \hat{Y}_{0i})$
  $ATET_{matched} = \frac{1}{N_T} \sum_{i=1}^{N} w_i (Y_i - \hat{Y}_{0i})$
- If $D_i = 1$ then $\hat{Y}_{1i} = Y_i$. If $D_i = 0$ then $\hat{Y}_{0i} = Y_i$
- For ATET, we don't need to impute Y_i
- Matching uses different ways of imputing $Y_0 i$ for treated units (or $Y_1 i$ for control units)
- The other way of understanding matching is that it is a weighting scheme. With exact matching, for example, $w_{ij} = 1/N_{matched}$ if $\mathbf{X_1} = \mathbf{X_0}$ and $w_{ij} = 0$ if $\mathbf{X_1} \neq \mathbf{X_0}$ (i indexes treated units and j indexes controls)

Inverse Probability Weighting

- We saw this in the intro class and homework. We use the propensity score as an inverse weight
- Assuming $\hat{p}(\mathbf{x}_i)$ is the predicted propensity score, for ATE $ipw_i = \frac{1}{\hat{p}(\mathbf{x}_i)}$ if $D_i = 1$ and $ipw_i = \frac{1}{1-\hat{p}(\mathbf{x}_i)}$ if $D_i = 0$
- We can also define weights to get ATET and ATEC
- ATET: $ipw_i = 1$ if $D_i = 1$ and $ipw_i = \frac{\hat{p}(\mathbf{x}_i)}{1-\hat{p}(\mathbf{x}_i)}$ if $D_i = 0$
- ATEC: $ipw_i = \frac{1-\hat{p}(\mathbf{x}_i)}{\hat{p}(\mathbf{x}_i)}$ if $D_i = 1$ and $ipw_i = 1$ if $D_i = 0$
- We did IPW by hand, but we can use -teffects ipw- or -teffects ipwra-, although teffects runs stratified models
- Again, we need to restrict the region of overlap somehow. For simplicity, we will restrict weights between 1500 and 3000, although we may get large IPW weights

IPW

```
qui logit treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first preterm momage dayskidh if keep1==1, nolog
predict double ps1 if e(sample)
gen     ipw = 1/ps1 if treat==1
replace ipw = 1/(1-ps1) if treat==0

* Outcome model not controlling for covariates
reg ppvtr36 treat [pw=ipw], robust

. reg ppvtr36 treat [pw=ipw], robust
(sum of wgt is 2,191.88404154778)

Linear regression                               Number of obs     =      1,320
                                                F(1, 1318)        =      36.77
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0462
                                                Root MSE          =     19.217

------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   8.996997   1.483815     6.06   0.000     6.086099    11.90789
       _cons |   82.55631   1.013474    81.46   0.000     80.56811    84.54451
------------------------------------------------------------------------------
```

IPW - teffects

```
teffects ipw (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll ///
  workdur prenatal male first preterm momage dayskidh) if keep1==1

Iteration 0:   EE criterion =  2.574e-16
Iteration 1:   EE criterion =  1.478e-26

Treatment-effects estimation                    Number of obs     =      1,320
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   8.996997   1.277541     7.04   0.000     6.493063    11.50093
-------------+----------------------------------------------------------------
POmean       |
       treat |
          0  |   82.55631    .936375    88.17   0.000     80.72105    84.39157
------------------------------------------------------------------------------
```

IPW - teffects

- Still not optimal, we could further restrict to overlap region based on ps or try other models

```
. tebalance summarize

  Covariate balance summary
                                                 Raw     Weighted
                          ----------------------------------------
                          Number of obs =        1,320      1,320.0
                          Treated obs   =          290        434.0
                          Control obs   =        1,030        886.0
                          ----------------------------------------

  ------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw    Weighted           Raw   Weighted
  ----------------+-------------------------------------------------
              bw  |  -2.008488   -.6226793       .724186   .3448521
         hispanic |  -.2688362   -.0461161      .5603778   .9176752
            black |    .257156   -.0656871      1.067346   .9807627
            bmarr |  -.3324808   -.0437763      1.020362   1.009212
             lths |   .1930087    .0730392      1.096464   1.041372
               hs |  -.2949933    .0096082      .8333505   1.005327
           ltcoll |  -.0573106   -.1439304       .909349   .7750979
          workdur |   .0242883     .117083      .9942454   .9718457
         prenatal |   -.112584   -.0381501      1.812482   1.281361
             male |  -.0736482     .071718      1.010007   .9965046
            first |   .0705086    .0366211      1.012425   1.014521
          preterm |   1.649226    .4140395      .5965296   .2586534
           momage |   .1905657    .1157435      3.280781   2.578142
         dayskidh |   .8504919    .4614788      1.585644   1.137282
  ------------------------------------------------------------------
```

IPW - teffects - ATET

- ATE restricting to the previous overlap region

```
. teffects ipw (ppvtr36) (treat bw hispanic black bmarr lths hs ltcoll ///
>               workdur prenatal male first preterm momage dayskidh) if keep ==1, atet

Iteration 0:   EE criterion =  1.737e-24
Iteration 1:   EE criterion =  1.818e-29

Treatment-effects estimation                    Number of obs     =        855
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATET         |
       treat |
   (1 vs 0)  |   10.53538   2.411306     4.37   0.000     5.809307    15.26145
-------------+----------------------------------------------------------------
POmean       |
       treat |
          0  |   81.58247   2.366822    34.47   0.000     76.94358    86.22135
------------------------------------------------------------------------------
```

IPW - teffects - ATET

- ATE restricting to the previous overlap region

```
. tebalance summarize

  Covariate balance summary
                                                      Raw     Weighted
                          ----------------------------------------------
                          Number of obs =             855        855.0
                          Treated obs   =             283        362.7
                          Control obs   =             572        492.3
                          ----------------------------------------------

  ------------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw    Weighted           Raw   Weighted
  ----------------+-------------------------------------------------------
              bw  |  -1.394722    .3778824      .6006019   .7275763
         hispanic |  -.0650134   -.0841342      .8470438   .8086912
            black |   .0698415    -.119931      1.008199   1.012672
            bmarr |   -.141656    -.055692      .9847759   .9892314
             lths |   .0908845    .0970404      1.033483   1.035198
               hs |  -.1289964   -.0933433       .904266   .9256661
           ltcoll |  -.0278215    -.079925      .9542544   .8764902
          workdur |    .005303     .246544       1.00003   .9791861
         prenatal |  -.1223482   -.1930732      1.980514   3.637795
             male |   -.076909    .1539902        1.0083   1.023679
            first |   .0311325    .3258814      1.004758   1.137434
          preterm |   1.061327   -.4482802      .5034276   .2973894
           momage |    .100821   -.0030976      2.852832   2.613683
         dayskidh |   .6435155    .0642314      1.107973     .76524
  ------------------------------------------------------------------------
```

-teffects ipwra-

- We didn't control for covariates in the outcome model but we could, and **that would help us deal with the remaining imbalance** (!)
- That is, the outcome model would be

```
. reg ppvtr36 treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                  first preterm momage dayskidh [pw=ipw], robust
(sum of wgt is 2,191.88404154778)

Linear regression                               Number of obs     =      1,320
                                                F(15, 1304)       =      25.60
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3512
                                                Root MSE          =     15.934

------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   8.380397   1.870968     4.48   0.000     4.709961    12.05083
          bw |  -.0019433    .003049    -0.64   0.524    -.0079248    .0040381
..
```

- This won't exactly match -teffects ipwra- since Stata runs stratified models
- Please check the code using the link below to see how you can match teffects by hand. teffects has the correct SEs, though (estimates the propensity score models and the oucome model simultaneously with GMM)

  https://clas.ucdenver.edu/marcelo-perraillon/sites/default/files/attached-files/matching_teffects_code_perraillon_0.do

-teffects ipwra-

- We could and should try different model specifications as well (interactions, etc)
- Note that with IPWRA we could tolerate some imbalance because we also control for covariates in the outcome model

```
. teffects ipwra (ppvtr36 bw hispanic black bmarr lths hs ltcoll  ///
>               workdur prenatal male first preterm momage dayskidh) ///
>                           (treat bw hispanic black bmarr lths hs ltcoll workdur prenatal
> ///
>                           male first preterm momage dayskidh) if keep1 ==1, ate

Iteration 0:   EE criterion =  2.574e-16
Iteration 1:   EE criterion =  7.165e-26

Treatment-effects estimation                    Number of obs     =      1,320
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
       treat |
   (1 vs 0)  |   8.316608   1.730835     4.80   0.000     4.924233    11.70898
-------------+----------------------------------------------------------------
POmean       |
       treat |
          0  |   83.15841   .8359674    99.48   0.000     81.51995    84.79688
------------------------------------------------------------------------------
```

Strong ignorability

- If ignorability and overlap hold (strong ignorability), it turns out that IPW is just another way of estimating $E[Y_i|Di = 1, \mathbf{X}_i]$ and $E[Y_i|Di = 0, \mathbf{X}_i]$
- One can show that:

  $E[\frac{Y_i D_i}{\hat{p}(\mathbf{X}_i)}] = E[Y_i|D_i = 1, \mathbf{X}_i]$ and

  $E[\frac{Y_i(1-D_i)}{1-\hat{p}(\mathbf{X}_i)}] = E[Y_i|D_i = 0, \mathbf{X}_i]$
- IPW is equivalent to the Horvitz and Thompson (1952) estimator for handling nonrandom sampling in surveys, in which the weight is the inverse probability of being in the sample
- Note that for the above to work, $\hat{p}(\mathbf{X}_i)$ cannot be 0 or 1, which means that overlap must hold
- A similar approach can be used for ATET
- So when overlap holds and assuming that model specification in regression adjustment is correct, we shouldn't expect to find much different between IPW and regression adjustment, with bonus that IPW is can be doubly robust
- Stata also has the command -teffects aipw- for "augmented" IPW that has the doubly robust property (-teffects aipw-)

Stratification

- Create groups based on the propensity score, say quintiles
- Make comparisons within quintiles defined by the propensity score. Could combine estimation using sample sizes
- The problem is that in some quintiles balance could be bad, or in extreme cases there could be no treated or control observations
- Stratification by the propensity score has a deep connection with an alternative to regression adjustment when only one variable determines treatment (Rubin, 1977). Robust to treatment heterogeneity

Rubin (1977)

Journal of Educational Statistics
Spring 1977, Volume 2, Number 1, Pp. 1-26

ASSIGNMENT TO TREATMENT GROUP ON THE BASIS OF A COVARIATE

Donald B. Rubin

Educational Testing Service

Key words: Non-Randomized Studies; Observational Studies; Covariance Adjustment; Causal Inference; Experimental Design; Treatment Assignment; Average Treatment Effects

ABSTRACT

When assignment to treatment group is made solely on the basis of the value of a covariate, X , effort should be concentrated on estimating the conditional expectations of the dependent variable Y given X in the treatment and control groups. One then averages the difference between these conditional expectations over the distribution of X in the relevant population. There is no need for concern about "other" sources of bias, e.g., unreliability of X , unmeasured background variables. If the conditional expectations are parallel and linear, the proper regression adjustment is the simple covariance adjustment. However, since the quality of the resulting estimates may be sensitive to the adequacy of the underlying model, it is wise to search for nonparallelism and nonlinearity in these conditional expectations. Blocking on the values of X is also appropriate, although the quality of the resulting estimates may be sensitive to the coarseness of the blocking employed. In order for these techniques to be useful in practice, there must be either substantial overlap in the distribution of X in the treatment groups or strong prior information.

Digression

- Stratification same as interacted model

```
bcuse bwght, clear
gen     smoked = 0
replace smoked = 1 if cigs ==0

qui reg bwght i.smoked##i.white
margins, dydx(smoked)
Average marginal effects                        Number of obs     =      1,388
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.smoked

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smoked |   8.889065   1.488571     5.97   0.000     5.968966    11.80917
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

quietly {
   reg bwght i.smoked if white ==1
   scalar beta1 = _b[1.smoked]
   scalar N1 = e(N)

   reg bwght i.smoked if white ==0
   scalar beta2 = _b[1.smoked]
   scalar N2 = e(N)
}
di (N1*beta1 + N2*beta2)/(N1+N2)
8.8890654
```

Stratification

- Overlap only exists in one region, the same we found before!

```
capture drop ps
qui logit treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
 first preterm momage dayskidh, nolog
predict double ps if e(sample)

xtile pscats = ps, n(5)
tab pscats treat
         5 |
 quantiles |         treat
     of ps |         0          1 |     Total
-----------+----------------------+----------
         1 |       877          0 |       877
         2 |       876          0 |       876
         3 |       876          0 |       876
         4 |       876          0 |       876
         5 |       586        290 |       876
-----------+----------------------+----------
     Total |     4,091        290 |     4,381
tabstat ps if pscats==5, by(treat) stats(N mean max min)
Summary for variables: ps
     by categories of: treat

   treat |         N      mean       max       min
---------+--------------------------------------------
       0 |       586  .1802457  .9828839   .016451
       1 |       290  .6124459   .995504  .0181568
---------+--------------------------------------------
   Total |       876  .3233256   .995504   .016451
-------------------------------------------------

reg ppvtr36 treat if pscats ==5, robust
...
------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   9.866971    1.32997     7.42   0.000     7.256663    12.47728
```

Stratification

- Could control for variables

```
. reg ppvtr36 treat bw hispanic black bmarr lths hs ltcoll workdur prenatal male  ///
>                 first preterm momage dayskidh if pscats ==5, robust

Linear regression                               Number of obs     =        876
                                                F(15, 860)        =      37.38
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3928
                                                Root MSE          =     16.349

------------------------------------------------------------------------------
             |               Robust
     ppvtr36 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   10.45539   1.477092     7.08   0.000     7.556261    13.35452
          bw |  -.0012955    .002095    -0.62   0.537    -.0054075    .0028165
    hispanic |  -15.29031   2.348434    -6.51   0.000    -19.89965   -10.68098
       black |  -18.06875   1.321689   -13.67   0.000    -20.66286   -15.47463
       bmarr |   1.456162   1.353218     1.08   0.282    -1.199835    4.112159
        lths |  -12.13782   2.308979    -5.26   0.000    -16.66971   -7.605922
          hs |  -7.425949    2.09373    -3.55   0.000    -11.53537   -3.316531
      ltcoll |  -5.144063   2.107533    -2.44   0.015    -9.280573   -1.007553
     workdur |    4.13442   1.225533     3.37   0.001     1.729034    6.539806
    prenatal |   .1939434   3.359773     0.06   0.954    -6.400372    6.788259
        male |    1.24003   1.122819     1.10   0.270    -.9637574    3.443816
       first |   4.195408   1.207222     3.48   0.001     1.825962    6.564853
     preterm |   .5278955   .2720722     1.94   0.053    -.0061078    1.061899
      momage |   -.119995   .1597463    -0.75   0.453    -.4335333    .1935433
    dayskidh |  -.2562127   .0816998    -3.14   0.002    -.4165671   -.0958583
       _cons |   100.5914   8.450377    11.90   0.000      84.0056    117.1772
------------------------------------------------------------------------------
```

Final comment

- Perhaps the most important question when thinking about matching, IPW, stratification is: why not regression adjustment? What's the problem with it?
- Lack of overlap is the problem if in fact it exists - so check for it with the tools you learned in these lectures
- If strong ignorability holds (ignorability plus overlap or selection on observables plus overlap), then think about matching, IPW, stratification as alternatives to regression adjustment
- Alternatives that help you understand your data better, and alternatives that have interesting properties – like more robust to model mispecification