Propensity Scores II

Marcelo Coca Perraillon
HSR - 2012

Reminders

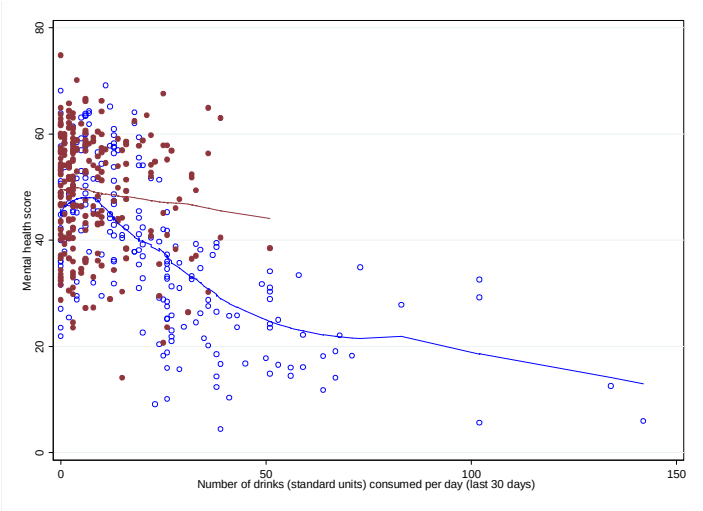
- Read Gelman & Hill's chapter 10
- Really, read Gelman & Hill…
- Homework will be based on Dehejia and Wahba (1999). You'll replicate some of the analyses in their paper

HELP data

```
. bysort homeless: sum age female i1 pcs drugrisk

-------------------------------------------------------------------------------------------------
-> homeless = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       244    35.04098    7.165759         21         58
      female |       244    .2745902    .4472249          0          1
          i1 |       244    8.061475    9.743221          0         51
         pcs |       244    49.00083    10.82878   14.07429   74.80633
    drugrisk |       243    1.728395    3.975168          0         21

-------------------------------------------------------------------------------------------------
-> homeless = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       209    36.36842    8.260958         19         60
      female |       209    .1913876    .3943379          0          1
          i1 |       209    23.03828    23.47315          0        142
         pcs |       209    38.42003    15.07664   4.435177   69.17161
    drugrisk |       209     2.07177    4.725098          0         21
```

Note: some people prefer to use the standardized mean difference instead. Always check distributions.

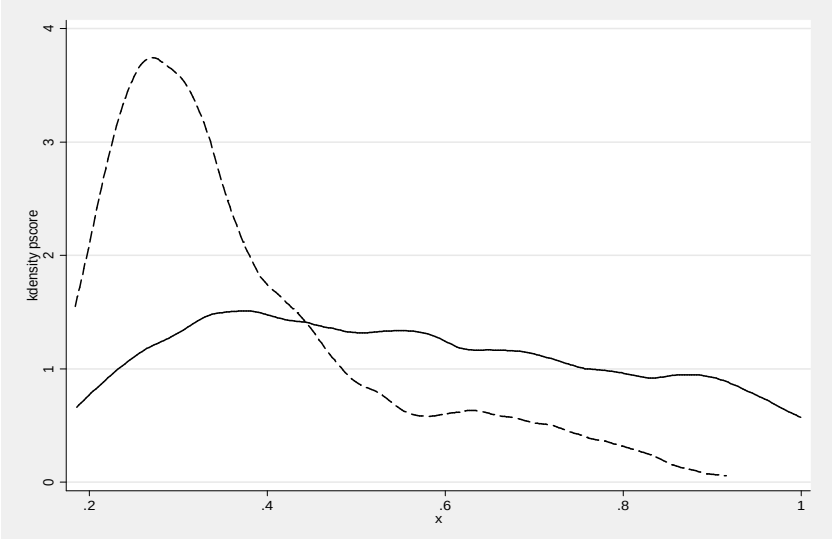
Lowess



Estimate the propensity score

```
. logit homeless age female i1 drugrisk, nolog

Logistic regression                               Number of obs   =        452
                                                  LR chi2(4)      =      95.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -264.17304                       Pseudo R2       =     0.1534

------------------------------------------------------------------------------
    homeless |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0048899   .0138751     0.35   0.725    -.0223048    .0320846
      female |  -.3696801   .2572309    -1.44   0.151    -.8738435    .1344833
          i1 |   .0677485   .0089214     7.59   0.000     .0502629     .085234
    drugrisk |   .0483531   .0244174     1.98   0.048     .0004959    .0962103
       _cons |  -1.247658   .5101496    -2.45   0.014    -2.247532   -.2477826
------------------------------------------------------------------------------

predict pscore
```

Important: the outcome is NOT in the regression. We want to mimic randomization.



What to do with the propensity score?

Typical uses:

1. Stratification
2. Weighting
3. Matching
4. ~~Adding the propensity score as a covariate~~

Stratification

```
xtile quints = pscore, nq(5)
reg pcs homeless age female i1 drugrisk
est sto model_all
forvalues i= 1(1)5 {
   reg pcs homeless age female i1 drugrisk if quints == `i'
    est sto model_q`i'
}
est table model_all model_q1 model_q2 model_q3 model_q4 model_q5, star  b(%7.2f)
```

Variable	model_all	model_q1	model_q2	model_q3	model_q4	model_q5
homeless	-5.38***	-1.82	-1.46	-0.13	-5.58*	-18.37***
age	-0.19**	-0.26	-0.23	-0.09	-0.25	-0.29*
female	-5.62***	-3.77	-5.33	-1.99	-7.40	-4.88
i1	-0.36***	0.96	0.20	-0.59	-0.85*	-0.12*
drugrisk	-0.33**	0.21	-0.50	-0.42	-1.01*	-0.03
_cons	60.83***	57.38***	60.25***	59.50***	74.42***	60.88***

legend: * p<0.05; ** p<0.01; *** p<0.001

Are the strata comparable?

```
. bysort homeless: sum age female i1 drugrisk if quints ==1
-------------------------------------------------------------------------------------------
-> homeless = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        70        34.6    7.645023         21         58
      female |        70    .6714286    .4730851          0          1
          i1 |        70    1.257143    1.699836          0          6
    drugrisk |        70    .9285714    2.379389          0          9

-------------------------------------------------------------------------------------------
-> homeless = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        21    33.52381    7.737048         23         48
      female |        21    .5714286    .5070926          0          1
          i1 |        21    1.285714    2.305273          0          7
    drugrisk |        21    .0952381    .3007926          0          1
```

What about stratum 4?

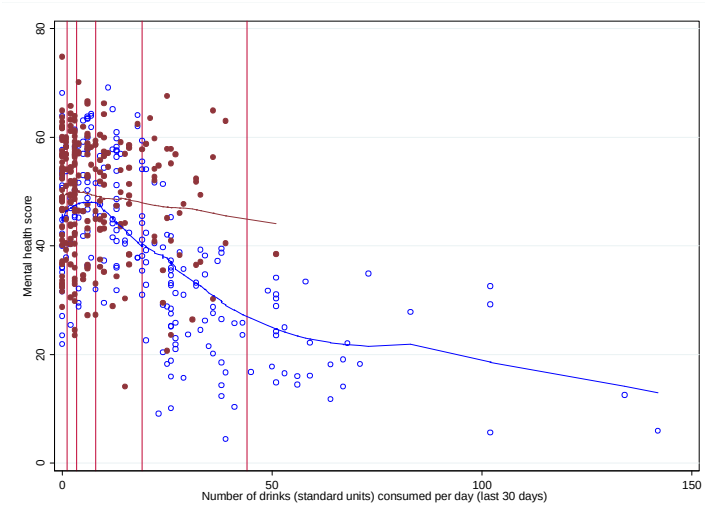
```
. bysort homeless: sum age female i1 drugrisk if quints ==4

-------------------------------------------------------------------------------
-> homeless = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        37    35.32432    4.904952         27         48
      female |        37    .0540541    .2292434          0          1
          i1 |        37    18.62162    4.917945          9         26
    drugrisk |        37    1.513514    3.746269          0         14

-------------------------------------------------------------------------------
-> homeless = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        53    37.71698    9.810468         19         60
      female |        53    .1698113       .37906          0          1
          i1 |        53    18.84906    6.794784          2         29
    drugrisk |        53    2.396226    5.248955          0         21
```

Stratum 5

```
. bysort homeless: sum age female i1 drugrisk if quints ==5

-------------------------------------------------------------------------------------------------
-> homeless = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        18    40.11111    7.583953         27         55
      female |        18    .1666667    .3834825          0          1
          i1 |        18          33    6.087596         26         51
    drugrisk |        18    1.777778    5.341905          0         21

-------------------------------------------------------------------------------------------------
-> homeless = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        72    38.15278    7.328624         20         57
      female |        72    .1944444    .3985498          0          1
          i1 |        72    46.77778    24.47604         15        142
    drugrisk |        72    2.277778    5.286769          0         20
```

Stats for i1 by quintile

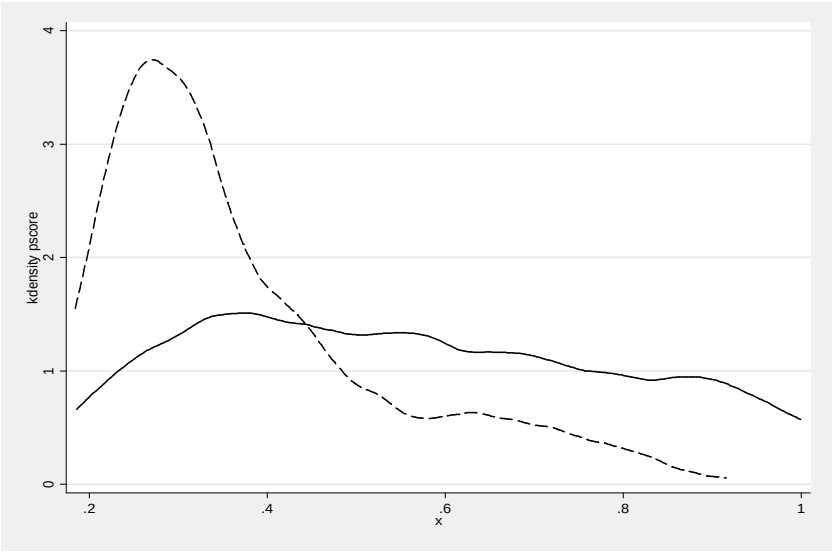
```
. table quints, c(mean i1 sd i1 min i1 max i1)

----------------------------------------------------------
5          |
quantiles |
of pscore |   mean(i1)      sd(i1)     min(i1)     max(i1)
----------+-----------------------------------------------
        1 |    1.26374    1.842915           0           7
        2 |    3.36264    2.382886           0          11
        3 |        7.9    4.248066           0          19
        4 |    18.7556    6.063912           2          29
        5 |    44.0222    22.70919          15         142
----------------------------------------------------------
. tab quints homeless

           5 |
 quantiles |     1 if Homeless
 of pscore |         0          1 |     Total
-----------+----------------------+----------
         1 |        70         21 |        91
         2 |        66         25 |        91
         3 |        52         38 |        90
         4 |        37         53 |        90
         5 |        18         72 |        90
-----------+----------------------+----------
     Total |       243        209 |       452
```



Weighting

- Use the propensity score as a weight in a regression
- The main idea is that the weighted data will be balanced
- For treated units: 1/pscore. For controls: 1/(1-pscore).
- You may see this written in a different way: $w_i = Z_i/e_i + (1-Z_i)/(1-e_i)$

Warning: Unstable weights in some cases (propensity score close to zero or one)



Weighting example

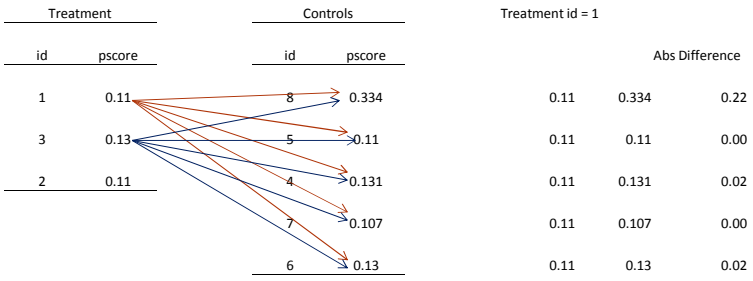
```
gen     w = 1/pscore if homeless == 1
replace w = 1/(1-pscore) if homeless == 0
* now, compare the groups again weighting
. bysort homeless: sum age female i1 drugrisk [aweight=w]

-------------------------------------------------------------------------------------------------
-> homeless = 0
    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         age |     243  441.817611    35.58435   7.058303         21         58
      female |     243  441.817611    .2210831   .4158329          0          1
          i1 |     243  441.817611    12.82159   13.24991          0         51
    drugrisk |     243  441.817611    1.925442   4.516877          0         21

-------------------------------------------------------------------------------------------------
-> homeless = 1
    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         age |     209  450.540094    35.67488   8.281236         19         60
      female |     209  450.540094    .2286887   .4209967          0          1
          i1 |     209  450.540094    15.10925   19.02263          0        142
    drugrisk |     209  450.540094    1.810754   4.315663          0         21
```

Run a weighted regression

```
. reg pcs homeless age female i1 drugrisk [pweight = w]
(sum of wgt is   8.9236e+02)

Linear regression                               Number of obs =     452
                                                F(  5,   446) =   53.12
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2931
                                                Root MSE      =  11.111

------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homeless |  -5.198854   1.201281    -4.33   0.000    -7.559728   -2.837981
         age |  -.2360235   .0767722    -3.07   0.002    -.3869037   -.0851433
      female |  -5.687987   1.373758    -4.14   0.000     -8.38783   -2.988144
          i1 |  -.3369473   .0346501    -9.72   0.000    -.4050452   -.2688495
    drugrisk |  -.4088917   .1090059    -3.75   0.000    -.6231207   -.1946627
       _cons |   63.16712   2.798783    22.57   0.000     57.66668    68.66756
------------------------------------------------------------------------------
```

Matching

- Most common: 1 to 1, nearest-neighbor matching without replacement



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

In Stata

Prepare control dataset, save it

```
preserve
keep pscore id homeless
keep if homeless ==0
set seed 12345
gen ru = runiform()
sort ru
drop ru
gen matched = .
gen mdiff   = .
save controls, replace
restore

gsort -homeless

forvalues i=1(1)209 {
   local ps = pscore[`i']
   preserve
   use controls, clear
   gen diff = abs(`ps' - pscore)
   sort diff
   replace matched = 1 if _n==1
   replace mdiff   = diff if _n==1
   replace pscore = .  if _n==1
   drop diff
   save controls, replace
   restore
}

merge 1:1 id using controls
replace matched = 1 if homeless == 1
drop _merge
tab matched homeless
```

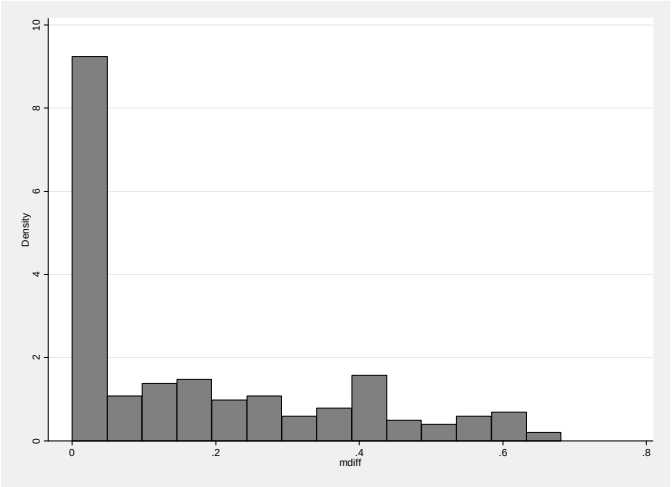
Do the matching

```
. tab matched homeless

           |     1 if Homeless
   matched |         0          1 |     Total
-----------+----------------------+----------
         1 |       209        209 |       418
-----------+----------------------+----------
     Total |       209        209 |       418
```

Is it a good match?

```
. sum age female i1 drugrisk if (matched == 1 & homeless == 0)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       209    34.89474    6.997144         21         58
      female |       209    .1866029    .3905278          0          1
          i1 |       209    9.177033     10.0877          0         51
    drugrisk |       209    1.933014     4.21254          0         21

. sum age female i1 drugrisk if homeless == 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       209    36.36842    8.260958         19         60
      female |       209    .1913876    .3943379          0          1
          i1 |       209    23.03828    23.47315          0        142
    drugrisk |       209     2.07177    4.725098          0         21
```

Check the difference in propensity scores




```
. sum mdiff, det

                          mdiff
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%      .000043             0
10%     .0001546             0       Obs                 209
25%     .0015045             0       Sum of Wgt.         209

50%     .0574722                     Mean           .1688572
                        Largest      Std. Dev.      .2070842
75%     .2763925       .7031915
90%     .5172513       .7037406      Variance       .0428839
95%     .5703443       .7057423      Skewness       1.013636
99%     .7037406       .7265043      Kurtosis       2.769244
```

It's a jungle out there…

Way too many flavors and combinations

- With replacement
- 1 to many
- Caliper
- Change metric of distance: Mahalanobis distance. You can then include other covariates in the distance metric.
- Global matching: you want to minimize the total distance, not just the local distance (can't do with large datasets)
- Full matching: a hybrid of stratification and matching. Minimizes the average distance
- Genetic algorithms (for large datasets)

Solutions?

- Could try caliper matching. Treatment units would be dropped
- Even harder to interpret results. It is actually a very common way to implement propensity scores
- In your homework you'll have thousands of potential controls. The matching will be very good

Any difference in the regression?

```
. reg pcs homeless age female i1 drugrisk if matched == 1

      Source |       SS       df       MS              Number of obs =     418
-------------+------------------------------           F(  5,   412) =   57.49
       Model |  34490.3348     5  6898.06697            Prob > F      =  0.0000
    Residual |  49430.4831   412  119.976901            R-squared     =  0.4110
-------------+------------------------------           Adj R-squared =  0.4038
       Total |   83920.818   417  201.248964            Root MSE      =  10.953

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homeless |  -5.803444   1.150199    -5.05   0.000    -8.064434   -3.542453
         age |  -.1727124   .0715858    -2.41   0.016    -.3134314   -.0319935
      female |  -4.508204   1.368994    -3.29   0.001    -7.199288   -1.817121
          i1 |   -.363829   .0304983   -11.93   0.000    -.4237808   -.3038773
    drugrisk |  -.3475819   .1209357    -2.87   0.004    -.5853099   -.1098539
       _cons |   60.46967   2.602416    23.24   0.000       55.354    65.58534
------------------------------------------------------------------------------
```

Note: No analysis for paired data here. See Gelman and Hill (2007)

- Coef for homessless was 5.38

Conclusions I

- Propensity scores are awesome to check the balance and overlap of covariates
- There are extensions for more than two treatments (multinomial models)
- The propensity score can make groups comparable but only on the variables used to estimate the propensity score. There is NO guarantee that you are balancing unobserved covariates
- If you know that there are important unobservable variables, you may need another tool
- Randomization ensures that both observable and unobservable variables are balanced

Conclusions II

- Think hard about this question: ***Do you really need a propensity score analysis?***
- Why not adjusting for confounders?
- There is something about IPW
- If you do propensity scores, what is the external validity? What are you estimating?