Propensity Scores I

Marcelo Coca Perraillon

HSR - 2012

Example 1: HELP data

- Data are from the Health Evaluation and Linkage to Primary Care (HELP).
- WARNING: The qualitative results are valid but the magnitudes are not. Part of the data for this example is simulated
- Details and complete data are on http://sas-and-r.blogspot.com/
- Code and simulated data will be posted on Chalk

Variables

```
. describe

Contains data from help_1.dta
  obs:           453
 vars:             6                          10 Apr 2012 11:34
 size:         5,889
------------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------------------
i1              int    %8.0g                  Number of drinks (standard units) consumed
   per day (last 30 days)
age             byte   %8.0g                  Age (years)
homeless        byte   %8.0g                  1 if Homeless
pcs             float  %9.0g                  SF-36 Mental Composite Score
drugrisk        byte   %8.0g                  Risk assesment battery (RAB) drug risk score
female          float  %9.0g                  1 if Female
------------------------------------------------------------------------------------------
Sorted by:  homeless
```

Mental score and homelessness

```
. reg pcs homeless

      Source |       SS       df       MS              Number of obs =     453
-------------+------------------------------           F(  1,   451) =   75.01
       Model |  12603.0176     1  12603.0176           Prob > F      =  0.0000
    Residual |  75774.2404   451  168.013837           R-squared     =  0.1426
-------------+------------------------------           Adj R-squared =  0.1407
       Total |   88377.258   452  195.524907           Root MSE      =  12.962

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homeless |   -10.5808   1.221669    -8.66   0.000    -12.98167   -8.179925
       _cons |   49.00083    .829808    59.05   0.000     47.37006     50.6316
------------------------------------------------------------------------------
```

Those who are homeless have on average 10.6 fewer points in the SF-36 mental health score. But we suspect drinking is a confounder…

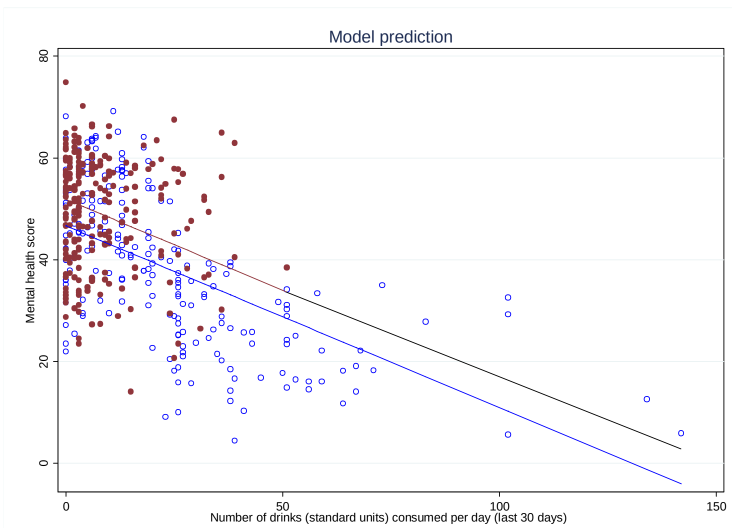
Adjust for the number of drinks

```
. reg pcs homeless i1

      Source |       SS       df       MS              Number of obs =     453
-------------+------------------------------           F(  2,   450) =  116.59
       Model |  30163.8819     2  15081.941            Prob > F      =  0.0000
    Residual |  58213.3761   450  129.363058           R-squared     =  0.3413
-------------+------------------------------           Adj R-squared =  0.3384
       Total |   88377.258   452  195.524907           Root MSE      =  11.374

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homeless |  -5.231871   1.166149    -4.49   0.000    -7.523645   -2.940098
          i1 |  -.3571473   .0306535   -11.65   0.000     -.417389   -.2969056
       _cons |   51.87996   .7689219    67.47   0.000     50.36884    53.39109
------------------------------------------------------------------------------
```

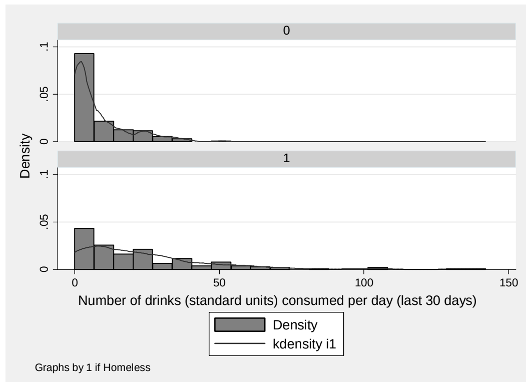
THIS CONCLUSION MAY BE (VERY) WRONG…

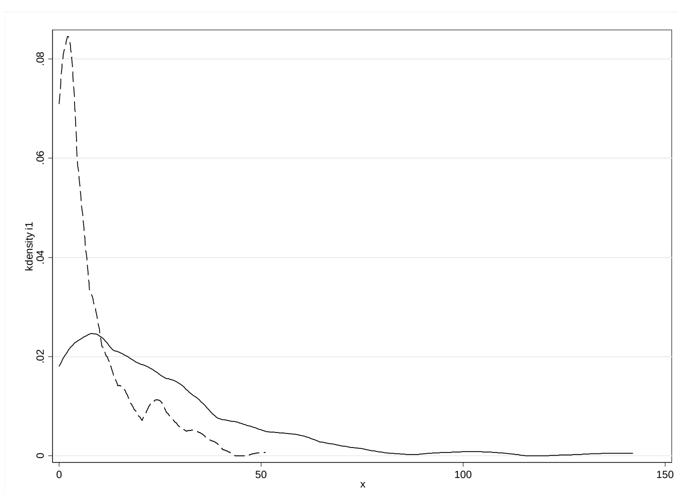
So it seems that the average number of drinks (per day) in the last 30 days is indeed a confounder. The model says that a homeless person has a mental health score that is 5.23 points lower than a person who drinks the same amount but is not homeless.

Check the model graphically



Check the distribution of i1

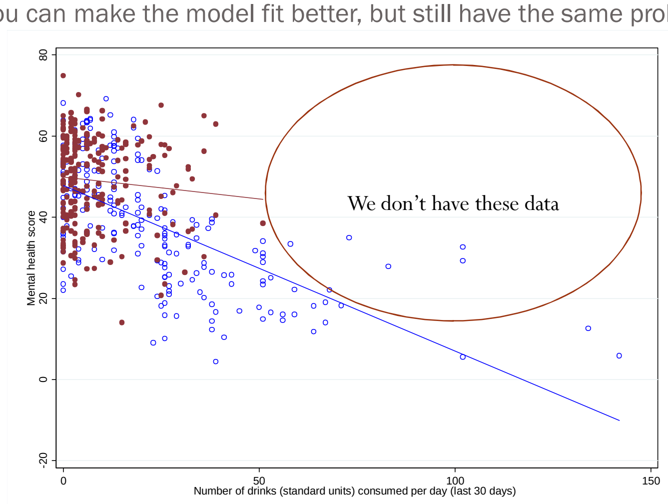




What's the problem?

- Those who consumed large amounts of alcohol are homeless. No person who is not homeless consumed more than 50 drinks (on average) per day in the last 30 days
- Alcohol consumption is a strong predictor of mental health status
- In other words, there is no counterfactual at high values of drinking
- We are relying on model extrapolation
- Graph the data again

You can make the model fit better, but still have the same problem

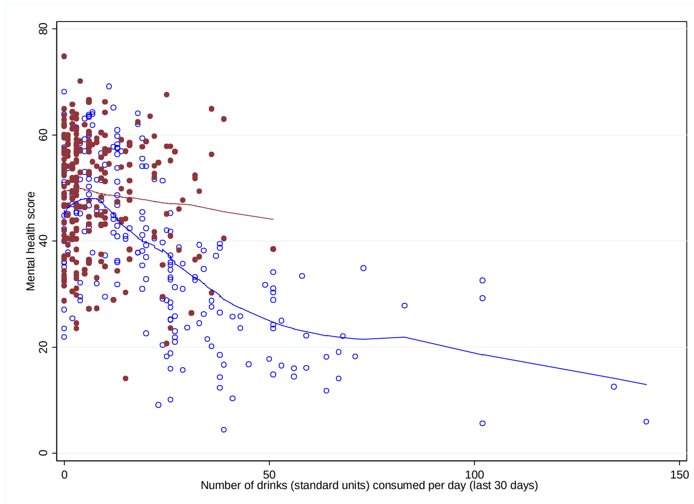


- Could add an interaction term
- Non-linearity (to avoid negative mental health score)

```
di (_b[_cons] + _b[1.homeless]*1 + _b[i1]*55 + _b[1.homeless#c.i1]*1*55) - ///
   (_b[_cons] + _b[1.homeless]*0 + _b[i1]*55 + _b[1.homeless#c.i1]*0*55)
-18.60
```

(Could use margins, dydx(homeless) at(i1=55))

Better yet, do lowess



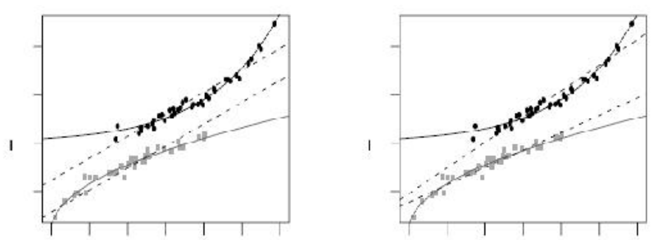
What could we do?

- Collect more data or design new experiment
- Ignore the problem and rely on model extrapolation. This may work if substantive knowledge tells you that the model is correct
- Limit analysis to region of overlap. In this example, run the same regression for i1 <= 37

Regression restricted to region of overlap (ignore interaction for now)

```
reg pcs homeless i1 if i1 <= 37

      Source |       SS       df       MS              Number of obs =     408
-------------+------------------------------           F(  2,   405) =   36.98
       Model |   9444.6708     2   4722.3354           Prob > F      =  0.0000
    Residual |  51712.1483   405  127.684317           R-squared     =  0.1544
-------------+------------------------------           Adj R-squared =  0.1503
       Total |  61156.8191   407  150.262455           Root MSE      =    11.3

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    homeless |  -4.031084   1.192311    -3.38   0.001    -6.374975   -1.687194
          i1 |  -.3766648   .0576415    -6.53   0.000    -.4899788   -.2633508
       _cons |   51.89442   .8503319    61.03   0.000      50.2228    53.56603
------------------------------------------------------------------------------

.
```

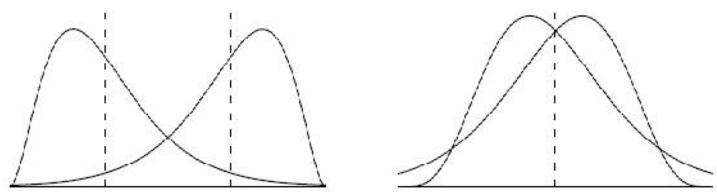
Example 2

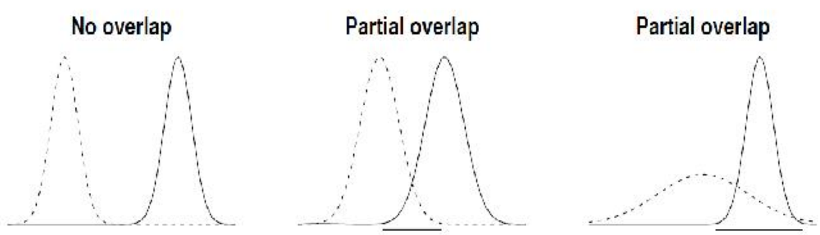


From Gelman & Hill (2007, Chapter 9). Y-axis is post-treatment score and x-axis is pre-treatment score of an educational intervention. The outcome (post-treatment score) depends on pre-treatment scores (the confounder). The second panel allows for interactions with treatment. Dark circles are treated units.

Lack of balance vs lack of overlap

Unbalanced but good overlap, not a big problem. Lack of overlap or partial overlap is a problem





The propensity score

- In more realistic situations, the treatment and control units may not be comparable in many covariates, not just one
- The propensity score is a very useful tool to determine if there is overlap and balance on multiple variables
- It is also useful to select comparable units (matching)
- It is less clear how much you can say about the estimate of interest after choosing comparable groups. To whom the estimate applies?
- Depends on many subjective choices

The propensity score

- Think of the propensity score as a one-number summary of all covariates. If treated and a control units have the same propensity score, then they have the same distribution of all the covariates that were used to estimate the propensity score. See Rosenbaum & Rubin (1983) for technical details.
- Assumes ignorability of treatment assignment (stats) or selection on the observables (economics).
- Definition (from Gelman and Hill, 2007): the propensity score for the ith individual is defined as the probability that he or she receives the treatment given everything we observe before the treatment (all the confounding covariates we want to control).

HELP data, again

```
. 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
```

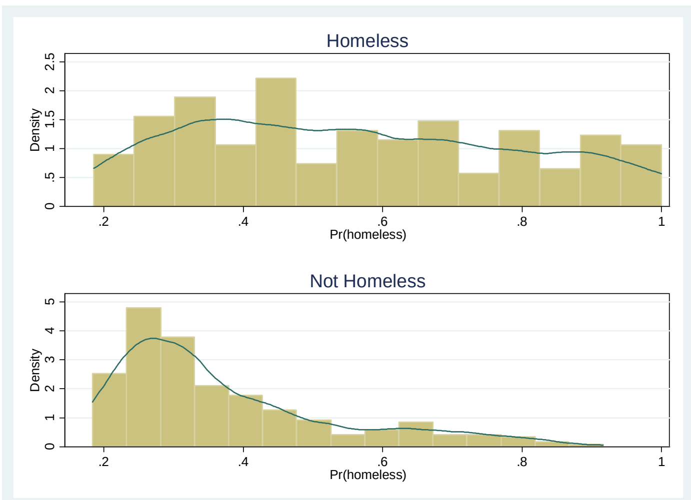
Note: some people prefer to use the standardized mean difference instead. Always check distributions.

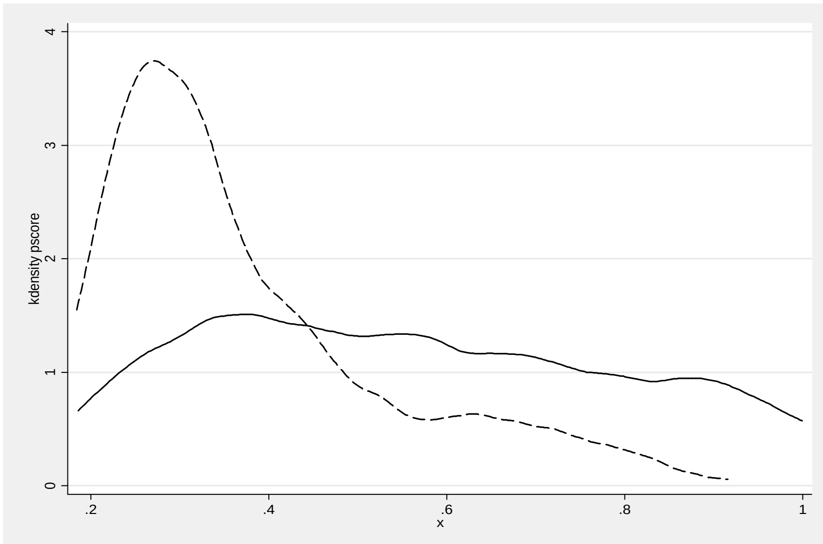
Estimate the propensity score

```
. 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.

Always check the propensity score





A silly but illustrative example

- Suppose you run a ***randomized*** experiment assigning 25% of your sample to the treatment group and 75% to the control group
- Suppose you then estimate the propensity score, that is, the probability of treatment assignment given some baseline covariates
- **Question 1**: What is the average value of the estimated propensity score for the treated? And for the control?

Simulated silliness, part I

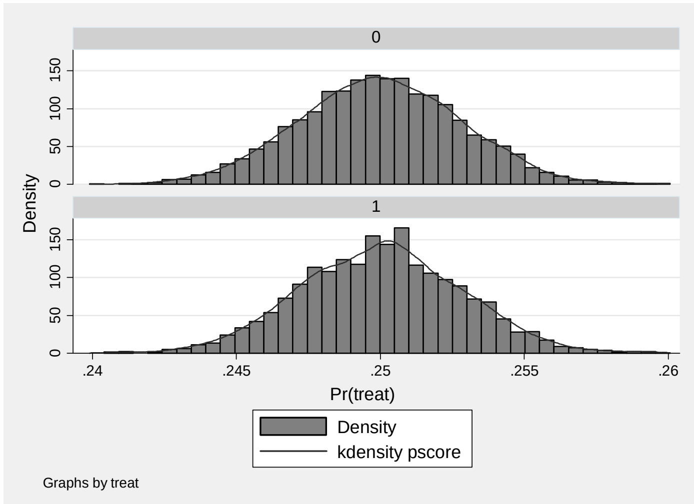
```
. bysort treat: sum ngirls age income

--------------------------------------------------------------------------------
-> treat = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      ngirls |      7500    195.0243    9.904098        160        229
         age |      7500    39.81803    10.03108   6.635278   74.93462
      income |      7500    1078.771    100.2127   732.0704    1455.97

--------------------------------------------------------------------------------
-> treat = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      ngirls |      2500    195.0604    9.882096        163        231
         age |      2500    39.96299    9.966808   5.043625   74.80503
      income |      2500    1078.867    102.2263   761.0687   1437.856
```

Simulated silliness, part II

```
. logit treat ngirls age income, nolog

Logistic regression                               Number of obs   =      10000
                                                  LR chi2(3)      =       0.42
                                                  Prob > chi2     =     0.9354
Log likelihood = -5623.1397                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ngirls |   .0003626   .0023338     0.16   0.877    -.0042116    .0049367
         age |   .0014799   .0023501     0.63   0.529    -.0031261    .0060859
      income |  -.0000183   .0002337    -0.08   0.938    -.0004764    .0004398
       _cons |  -1.208605   .5233735    -2.31   0.021    -2.234398   -.1828116
------------------------------------------------------------------------------

predict pscore
```

Question 2: If you plotted the distribution of the propensity scores by treatment group, what would you expect to see?

Simulated silliness, part III



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?

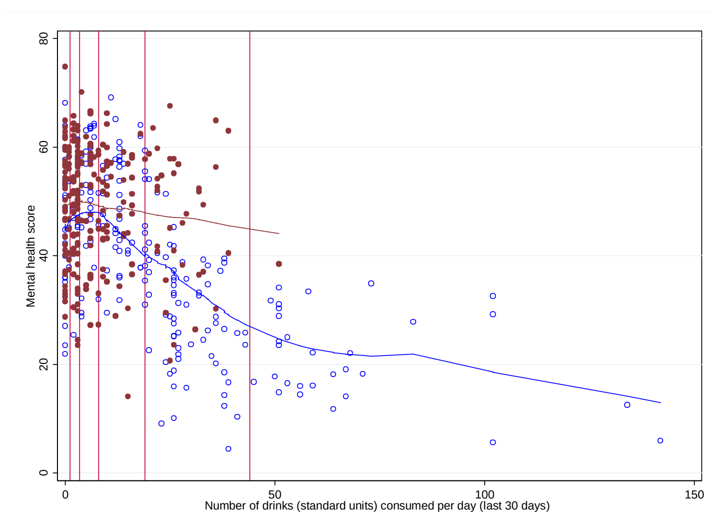
```
. bysort homeless: sum age female i1 drugrisk if quints ==4

--------------------------------------------------------------------------------
-> 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
```

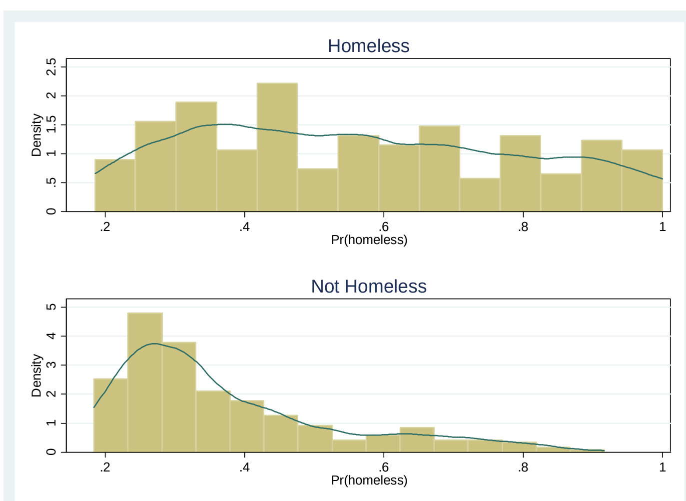
Stats for i1 by quintile

```
. 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
- Using weights will be clearer once we cover survey analysis
- 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: $w_i = Z_i/e_i + (1-Z_i)/(1-e_i)$



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
------------------------------------------------------------------------------
```