Week 4: Regression adjustment and propensity scores

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- Review of using regression for treatment effects
- Regression adjustment facelift: following the definition of causal effects
- Estimating ATE and ATET
- Checking for overlap (informally)
- The propensity score
- Checking for overlap and common support (formally)
- Applications
  1. Matching
  2. Stratification
  3. Inverse probability weighting
- teffects command
- Next steps

Regression adjustment: Main assumptions for causal inference

- We saw that we needed two assumptions to use regression adjustment for causal inference
  1. Ignorability or unconfoundness or CIA: $(Y_{1i}, Y_{0i}) \perp D_i | \boldsymbol{X}_i$
  2. Overlap (aka common support): For all $\boldsymbol{X}_i \in \varphi$, where φ is the support (domain) of the covariates $\boldsymbol{X}_i$, $0 < P(D = 1|\boldsymbol{X}_i) < 1$
- Rosenbaum and Rubin (1983) called the two assumptions together **strong ignorability**
- The other, of course, is SUTVA, which is **always** needed
- We also saw that a weaker version of 1) is Ignorability of Means: $E[Y_{0i}|D_i, \boldsymbol{X}_i] = E[Y_{0i}|\boldsymbol{X}_i]$ (same for Y_{1i})
- Randomization (conditional randomization) guarantees both are satisfied and we must argue SUTVA (a type of exclusion restriction)

Parametric, nonparametric, semiparametric

- With regression adjustment we can obtain, using observed data, $E[Y_i|D_i, \boldsymbol{X}_i]$
- Remember too that in the class on causal inference I said that we don't need to assume anywhere that $E[Y_i|D_i, \boldsymbol{X}_i]$ must be estimated with linear/OLS models or any parametric model. The estimation could be non-parametric or semiparametric – causal effects are identified either way
- Example of parametric model: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 Z + \epsilon$. In this model, we the obtain $E[Y_i|D_i, Z_i]$ as a function of parameters $\beta_1, \beta_2, \beta_3$
- A nonparametric model could be $Y_i = g(D_i, Z_i) + u_i$, where $g(.)$ is an unknown function (of an infinite set of functions). We don't estimate parameters, but we get a series of $\hat{Y}_i$ from which we can calculate $E[Y_i|D_i, Z_i]$
- Semiparametric is a combination of both, but there is confusion on what is called nonparametric vs semiparametric in the literature
- Nonparametric methods are not a panacea either. You trade one set of assumptions for another: bandwidth choice, weighting schemes, dimensionality issues

Data

- We will use a dataset to explore the impact of an intervention on mental health status score from the SF-36
- The dataset started as a real dataset but over time I made some changes to illustrate some points so by now it's simulated data. See do file

```
webuse set "https://perraillon.com/s/"
webuse "help_1_stata12.dta", clear

<..code omitted...>

Contains data from https://perraillon.com/s/help_1_stata12.dta
  obs:           452
 vars:             6                          15 Apr 2012 11:34
-----------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------------------
ndrinks         int     %8.0g                 Number of drinks (standard units) consumed per
                                                day (last 30 days)
age             byte    %8.0g                 Age (years)
intervention    byte    %8.0g                 1 if received intervention
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:
     Note: Dataset has changed since last saved.
```

Regression adjustment

- We are going to **pretend** that ignorability holds. Let's run our trusty, old fashioned linear/OLS model. What is the coefficient for intervention (5.38) telling us? (Higher PSC score is better outcome)

```
reg pcs intervention age female ndrinks drugrisk

      Source |       SS           df       MS      Number of obs   =       452
-------------+----------------------------------   F(5, 446)       =     57.54
       Model |  34647.4528         5  6929.49057   Prob > F        =    0.0000
    Residual |  53713.7962       446  120.434521   R-squared       =    0.3921
-------------+----------------------------------   Adj R-squared   =    0.3853
       Total |   88361.249       451  195.922947   Root MSE        =    10.974

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
intervention |   5.383645   1.132658     4.75   0.000     3.157635    7.609655
         age |  -.1944413   .0687731    -2.83   0.005    -.3296009   -.0592817
      female |  -5.617188   1.223214    -4.59   0.000    -8.021167   -3.213209
     ndrinks |  -.3554573   .0302739   -11.74   0.000    -.4149546     -.29596
    drugrisk |   -.334938   .1201294    -2.79   0.006    -.5710279   -.0988481
       _cons |   55.44966   2.617823    21.18   0.000     50.30486    60.59446
------------------------------------------------------------------------------
```

We can try other specifications

- We could interact intervention with number of drinks, for example. Effect of intervention non-constant (non-linear)

```
. reg pcs i.intervention##c.ndrinks age i.female  drugrisk

      Source |       SS           df       MS      Number of obs   =       452
-------------+----------------------------------   F(6, 445)       =     50.54
       Model |  35811.3652         6  5968.56086   Prob > F        =    0.0000
    Residual |  52549.8838       445  118.089627   R-squared       =    0.4053
-------------+----------------------------------   Adj R-squared   =    0.3973
       Total |   88361.249       451  195.922947   Root MSE        =    10.867

---------------------------------------------------------------------------------------
                  pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------+----------------------------------------------------------------
       1.intervention |   2.705919    1.40905     1.92   0.055    -.0632991    5.475137
              ndrinks |  -.3954348   .0325702   -12.14   0.000    -.4594454   -.3314243
                      |
intervention#c.ndrinks |
                    1  |   .2499904   .0796286     3.14   0.002     .0934956    .4064852
                      |
                  age |  -.2043996   .0681742    -3.00   0.003    -.3383829   -.0704163
             1.female |  -5.018761   1.226154    -4.09   0.000    -7.428534   -2.608989
             drugrisk |  -.3162415   .1191031    -2.66   0.008     -.550316    -.082167
                _cons |   56.57957   2.617078    21.62   0.000      51.4362    61.72294
---------------------------------------------------------------------------------------
margins, dydx(intervention)
<... output omitted ..>
+-----------------------------------------------------------------
1.intervention |   6.456881   1.172519     5.51   0.000     4.152519    8.761243
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

We can try other specifications

- Number of drinks could be quadratic. Again, the effect of intervention is non-constant (non-linear)

```
reg pcs i.intervention##(c.ndrinks##c.ndrinks) i.female  drugrisk
------------------------------------------------------------------------------------------
                   pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------------
        1.intervention |  -.5992266   1.687401    -0.36   0.723    -3.915512    2.717059
               ndrinks |  -.7272449   .0731894    -9.94   0.000    -.8710857   -.5834042
                       |
   c.ndrinks#c.ndrinks |   .0034224    .000711     4.81   0.000     .0020251    .0048197
                       |
 intervention#c.ndrinks |
                     1  |   .4953355   .2128372     2.33   0.020      .077042    .9136289
                       |
 intervention#c.ndrinks#|
             c.ndrinks |
                     1  |  -.0015199   .0058563    -0.26   0.795    -.0130294    .0099895
                       |
              1.female |   -5.02411   1.213554    -4.14   0.000    -7.409133   -2.639086
              drugrisk |  -.3451062   .1178018    -2.93   0.004    -.5766246   -.1135877
                 _cons |   53.15791   1.340412    39.66   0.000     50.52357    55.79225
------------------------------------------------------------------------------------------
margins, dydx(intervention)
<... output omitted ..>
--------------------------------------------------------------------------------
              |            Delta-method
              |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+-----------------------------------------------------------------
1.intervention |   5.943705   1.668067     3.56   0.000     2.665418    9.221992
--------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Regression adjustment following definition of causal effects

- Stata implemented a treatment effects group of commands
- The command **teffects ra** performs another way of doing regression adjustment
- The conceptual idea follows Wooldridge (2010), Chapter 21, overview of causal effects, but in essence follows basic principles that suggest nonparametric (or semiparametric) identification: Remember, under ignorability comparing $E[Y_i|\boldsymbol{X}_i, D_i = 1]$ to $E[Y_i|\boldsymbol{X}_i, D_i = 0]$ provides an estimate of causal effects
- We just did that using a linear/OLS model, but we could do it using a series of steps, which has didactical advantages and we can get ATE and ATET
- teffects ra estimates the steps, but **estimates all steps simultaneously using generalized methods of moments estimation** (GMM) (See Stata's PDF help on command gmm for a nice intro)

Regression adjustment **teffects ra** style, ATE

Step 1: Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 1]$ with a linear/OLS model using only treated observations

Step 2: Using estimates from 1), predict $\hat{Y}_{treated}$ in the **entire sample**

Step 3: Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 0]$ with a linear/OLS model using only control observations

Step 4: Using estimates from 3), predict $\hat{Y}_{control}$ in the **entire sample**

Step 5: The difference (contrast) between $E[\hat{Y}_{treated}] - E[\hat{Y}_{control}]$ is the ATE

- Note the logic. We use the experience of the treated to estimate how covariates $\boldsymbol{X}$ affect the outcome Y. We use the estimated model to make predictions about the counterfactual for the control $E[Y_{0i}|D = 1]$ (and the treated). Same logic for control group. See, causal inference is a **PREDICTION** problem

Estimating the five steps

```
* Steps 1 and 2
qui reg pcs age female ndrinks drugrisk if intervention == 1
predict double yhat_t
* Steps 3 and 4
qui reg pcs age female ndrinks drugrisk if intervention == 0
predict double yhat_c

sum yhat_t
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      yhat_t |        452    48.14365    3.929616   27.77863   55.38451
local pom_t = r(mean)

sum yhat_c
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      yhat_c |        452    41.55624     8.21028  -11.22789   51.53394

local pom_c = r(mean)
di ‘pom_t’ - ‘pom_c’
6.5874079
```

- We find that the treatment effect is 6.58. This approach can be called semiparametric

Using **teffects ra**

```
* Using teffects
teffects ra (pcs age female ndrinks drugrisk) (intervention), ate

Iteration 0:   EE criterion =  1.247e-28
Iteration 1:   EE criterion =  1.696e-29

Treatment-effects estimation                    Number of obs     =        452
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
intervention |
   (1 vs 0)  |   6.587408    1.24669     5.28   0.000      4.14394    9.030876
-------------+----------------------------------------------------------------
POmean       |
intervention |
          0  |   41.55624   .9719151    42.76   0.000     39.65133    43.46116
------------------------------------------------------------------------------

*teffects ra (pcs age female ndrinks drugrisk) (intervention), ate aeq
```

- The ate option is the default. You can get more info with aeq option. PO means "population outcome"

teffect ra for ATE is really a fully interacted parametric model

- We are interacting intervention with all the other covariates

```
. qui reg pcs i.intervention##(c.age i.female c.ndrinks c.drugrisk)

. margins, dydx(intervention)

Average marginal effects                        Number of obs     =        452
Model VCE    : OLS

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

--------------------------------------------------------------------------------
               |            Delta-method
               |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
1.intervention |   6.587408   1.176083     5.60   0.000     4.275999    8.898817
--------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Average Treatment Effect on the Treated (ATET)

- Here is where things get interesting. Following this logic, we can estimate ATET

  Step 1: Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 1]$ with a linear/OLS model using only treated observations

  Step 2: Using estimates from 1), predict $\hat{Y}_{treated}$ **only using the treated sample**

  Step 3: Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 0]$ with a linear/OLS model using only control observations

  Step 4: Using estimates from 3), predict $\hat{Y}_{treated_c}$ using only the **treated** sample. Essentially, this is the counterfactual for the treated
- The difference (contrast) between $E[\hat{Y}_{treated}]$ and $E[\hat{Y}_{treated_c}]$ is ATET
- Steps 1 and 2 are actually not necessary. We know that the average of the predictions will be the same as the average of observed Y since $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$, so $E[\hat{Y}_{treated}] = E[Y]$

ATET “by hand”

- Pay attention to the “if” operator in all steps

```
/// --- ATET
* Steps 1 and 2
qui reg pcs age female ndrinks drugrisk if intervention == 1
predict yhat_t1 if intervention == 1
* Steps 3 and 4
qui reg pcs age female ndrinks drugrisk if intervention == 0
predict yhat_t11 if intervention == 1

sum yhat_t1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     yhat_t1 |        243    49.00447    3.212145   39.70965   55.36874
local pom_t1 = r(mean)
* same as
sum pcs if intervention ==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pcs |        243    49.00447    10.85098   14.07429   74.80633

sum yhat_t11
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    yhat_t11 |        243    44.25364    4.726526    28.4412   51.53394
local pom_t11 = r(mean)
di ‘pom_t1’ - ‘pom_t11’
4.7508282
```

ATET using teffects ra

- The GMM estimation does need to estimate model 1

```
teffects ra (pcs age female ndrinks drugrisk) (intervention), atet aeq
Treatment-effects estimation                    Number of obs     =        452
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATET         |
intervention |
   (1 vs 0)  |   4.750828   1.200882     3.96   0.000     2.397142    7.104515
-------------+----------------------------------------------------------------
POmean       |
intervention |
          0  |   44.25364   1.028533    43.03   0.000     42.23775    46.26953
-------------+----------------------------------------------------------------
OME0         |
         age |  -.1194283   .0930727    -1.28   0.199    -.3018474    .0629908
      female |  -6.031054   1.930069    -3.12   0.002    -9.813919    -2.24819
     ndrinks |  -.4037942   .0422099    -9.57   0.000    -.4865241   -.3210644
    drugrisk |  -.4541629   .1403639    -3.24   0.001     -.729271   -.1790547
       _cons |   54.16137   3.475175    15.59   0.000     47.35015    60.97258
-------------+----------------------------------------------------------------
OME1         |
         age |  -.3030211   .0905716    -3.35   0.001    -.4805381    -.125504
      female |  -4.131283   1.431898    -2.89   0.004     -6.93775   -1.324815
     ndrinks |  -.1180554   .0702388    -1.68   0.093     -.255721    .0196102
    drugrisk |  -.1745113   .1817243    -0.96   0.337    -.5306844    .1816617
       _cons |   62.03521   3.253079    19.07   0.000     55.65929    68.41113
------------------------------------------------------------------------------
```

ATET with regression

- It will become clearer when we cover marginal effects

```
qui reg pcs i.intervention##(c.age i.female c.ndrinks c.drugrisk)
margins r.intervention, subpop(intervention)

Contrasts of predictive margins                 Number of obs     =        452
Model VCE    : OLS                              Subpop. no. obs   =        243

Expression   : Linear prediction, predict()

------------------------------------------------
             |         df           F        P>F
-------------+----------------------------------
intervention |          1       17.08     0.0000
             |
 Denominator |        442
------------------------------------------------

--------------------------------------------------------------
             |            Delta-method
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
intervention |
   (1 vs 0)  |   4.750828   1.149684      2.491301    7.010356
--------------------------------------------------------------
```

OLS and identification of ATE and ATET

- There is a subtle point in the previous discussion
- The treatment effects using the linear/OLS model only identifies ATE if there is no treatment heterogeneity
- If there is no treatment heterogeneity, then the usual way of doing regression adjustment would recover ATE
- We had to interact treatment with all the covariates to obtain ATE

Big picture

- We went straight from the definition of causal effects to ways to estimate ATE and ATE using different but related approaches
- ATET is 4.75 while ATE is 6.58, both statistically significant (trust teffects for SEs)
- That tells you something: the covariates may not be balanced between treatment and control and/or the effects of covariates on outcome could be different between treatment and control (heterogenous effects) – or something else could be going on
- As we will soon see, this makes substantive sense – the intervention group is different
- Remember that under randomization ATE = ATET. The treated and the control are similar (i.e. same distribution) in all observed characteristics $\boldsymbol{X}$ and all **unobserved** characteristics
- Remember too that we are **assuming** ignorability or conditional randomization
- **But what about overlap?**

Notice something odd?

- Below is the usual regression adjustment model you would use under ignorability
- There is nothing odd in the regression output, but in fact we have a problem in the regression below: **overlap doesn't hold**

```
reg pcs intervention age female ndrinks drugrisk

      Source |       SS           df       MS      Number of obs   =       452
-------------+----------------------------------   F(5, 446)       =     57.54
       Model |  34647.4528         5  6929.49057   Prob > F        =    0.0000
    Residual |  53713.7962       446  120.434521   R-squared       =    0.3921
-------------+----------------------------------   Adj R-squared   =    0.3853
       Total |   88361.249       451  195.922947   Root MSE        =    10.974

------------------------------------------------------------------------------
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
intervention |   5.383645   1.132658     4.75   0.000     3.157635    7.609655
         age |  -.1944413   .0687731    -2.83   0.005    -.3296009   -.0592817
      female |  -5.617188   1.223214    -4.59   0.000    -8.021167   -3.213209
     ndrinks |  -.3554573   .0302739   -11.74   0.000    -.4149546     -.29596
    drugrisk |   -.334938   .1201294    -2.79   0.006    -.5710279   -.0988481
       _cons |   55.44966   2.617823    21.18   0.000     50.30486    60.59446
------------------------------------------------------------------------------
```

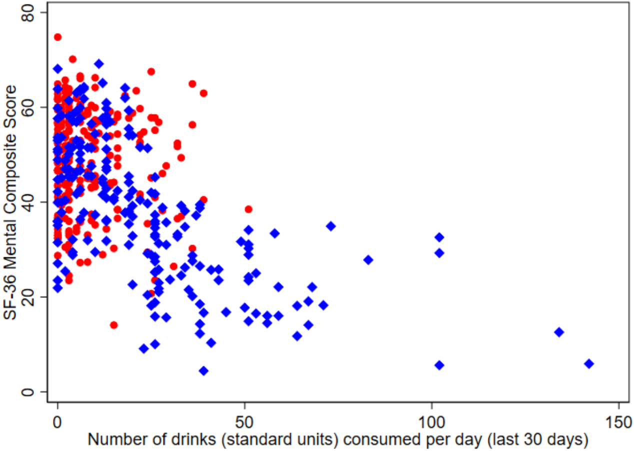
Checking overlap (informally)

- Number of drinks is a confounder and notice that in the control group there are more people who drank much more

```
sum  age female ndrinks drugrisk if intervention ==1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |        243    35.09465    7.131244         21         58
      female |        243    .2757202    .4477988          0          1
     ndrinks |        243     8.09465    9.749512          0         51
    drugrisk |        243    1.728395    3.975168          0         21

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

corr ndrinks pcs
(obs=452)
             | ndrinks      pcs
-------------+------------------
     ndrinks |   1.0000
         pcs |  -0.5584   1.0000

scatter pcs ndrinks if intervention ==1, color(red) msize(small) || ///
      scatter pcs ndrinks if intervention ==0, color(blue) msize(small) ///
  legend(off)
graph export pcs_drinks.png, replace
```

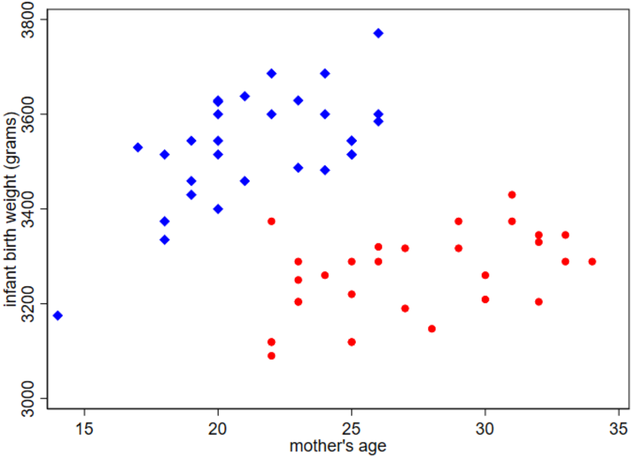
Picture worth a thousand words, etc

- Blue are controls. There is not a single treated unit with more than 51 drinks, which means that the **probability of receiving treatment** is zero for those who drink more than 51 drinks. There are fewer controls who a few drinks



Overlap

- The definition of overlap is broad and could go in either direction. See similar problem using sample data from Stata (see do file for code)



But what is the problem?

- The problem is that **implicitly** we are extrapolating information
- We are using the information from those in the control group who drank more than 51 drinks to make predictions about the treated group, but nobody in the treated group drank more than 51 drinks. You can frame the problem the other way, too
- So $E[Y_i|\boldsymbol{X}_i, D_i = 0] \neq E[Y_{0i}|\boldsymbol{X}_i, D_i = 1]$, which is equivalent to $E[Y_{0i}|\boldsymbol{X}_i, D_i = 0] \neq E[Y_{0i}|\boldsymbol{X}_i, D_i = 1]$
- It's a subtle problem that is easy to overlook if you don't carefully explore the data
- Whether the problem matters or not depends on how covariates affect treatment and outcomes
- It also depends on **functional form**: if we model correctly the relationship between drinks and pcs, then our predictions will be better. But we never know the true model

Implicit, explicit extrapolation

- I wrote above that when we use regression, the extrapolation is **implicit**
- Compare the usual regression adjustment with the new approach we covered at the beginning of the class (teffects ra)
- With that approach, the extrapolation is **explicit**. For example, in Step 1 for ATE, the estimates from a model using only the treated observations are used to make predictions in both treated and *controls*
- In other words, it's explicit that we use the information of the treated group –who never drank more than 51 drinks – to predict what would have happened to those in the control group when they drink a lot more
- Again, how big is the problem depends on the relationship between the number of drinks consumed and the outcome. Intuitively, modeling that relationship (functional form) correctly is important

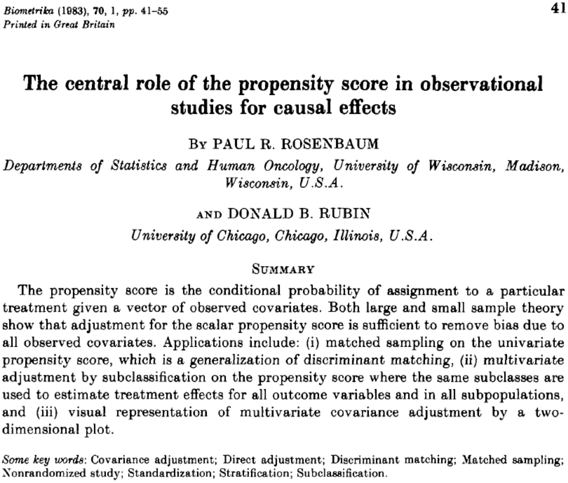
What could we do?

- Here is some intuition for the methods that we will cover. It's easier to intuitively think about solutions when the problem is with one variable, number of drinks here
  1. We could restrict estimation to the region where there is overlap – the region where we have information to make extrapolations (drinks ≤ 51)
  2. We could use the entire sample, but we could give more importance (weight) to the observations where overlap is good
  3. We could stratify the analysis instead comparing different regions. Say, 0 to 15 drinks, 16 to 20, 30+. This partially solves the problem. The comparison of 30+ now has pretty bad overlap
- The solutions above correspond to 1) **matching**, 2) **inverse propensity score weighting** (IPW), and 3) **stratification based on propensity score**, respectively
- But the solutions deal with the more realistic case when the **lack of overlap is due to multiple variables**

Diagnosing the problem: the Propensity Score

- We defined overlap as the condition $0 < P(D = 1|\boldsymbol{X}_i) < 1$ for all $\boldsymbol{X}_i \in \varphi$, where φ is the support (domain) of the covariates
- As I mentioned in a previous class, $P(D = 1|\boldsymbol{X}_i)$ is the definition of the propensity score:
$$p(\boldsymbol{X}_i) \equiv P(D = 1|\boldsymbol{X}_i)$$
- The propensity score, $p(\boldsymbol{X}_i)$, for unit i is the conditional probability of receiving treatment given observed covariates $\boldsymbol{X}$ (the propensity to receive treatment)
- Obviously, the probability of not receiving treatment is $1 - p(\boldsymbol{X}_i)$
- The importance of the propensity scores is presented in Rosenbaum and Rubin (1983), so we'll go to the source

Rosenbaum and Rubin (1983)



Biometrika (1983), 70, 1, *pp.* 41–55
Printed in Great Britain

The central role of the propensity score in observational studies for causal effects

By PAUL R. ROSENBAUM
Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A.

and DONALD B. RUBIN
University of Chicago, Chicago, Illinois, U.S.A.

Summary

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching, (*ii*) multivariate adjustment by subclassification on the propensity score where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (iii) visual representation of multivariate covariance adjustment by a two-dimensional plot.

Some key words: Covariance adjustment; Direct adjustment; Discriminant matching; Matched sampling; Nonrandomized study; Standardization; Stratification; Subclassification.

Why is the propensity score important?

- Rosenbaum and Rubin presented the propensity score as a **balancing** score, meaning this (I changed the notation to match ours):

  Theorem 1. Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is: $\boldsymbol{X} \perp D | p(\boldsymbol{X})$
- "Theorem 1 implies that if a subclass of units or a matched treatment-control pair is homogeneous in $p(\boldsymbol{X})$, then the treated and control units in that subclass or matched pair will have the same distribution of $\boldsymbol{X}$."
- Said another way, comparing the propensity score of treatment and control units is the same as comparing the distribution of covariates used to estimate the propensity score. That's something. So we can check overlap on **all** covariates by **checking the distribution of the propensity score**
- Note too that Theorem 1 implies mean independence given the propensity score in the sense that the propensity score will achieve balance

Big picture

- The way Theorem 1 is stated created a lot of confusion. Some interpreted it as saying that we only need to control for the propensity score rather than the covariates (the abstract doesn't help: "...adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates"), but that has multiple drawbacks
- However, they only proposed using the propensity score for matching and stratification, not as a covariate in a regression model. Using it as an inverse weight came later
- **Alert (!)**: Notice something subtle but **very** important: if overlap is satisfied, as in randomization, then using the propensity score (matching, stratification, IPW) should give very similar estimates as regression adjustment. The vector of covariates $\boldsymbol{X}$ are also balancing. The propensity score won't achieve any more balance if $\boldsymbol{X} \perp D$ already holds. That's Theorem 3
- More recent research suggests some advantages of extensions of IPW, like doubly robust methods (robust to misspecification of functional form). You get two chances to get it right (more on this on the second part of the class)

Preview: Using the propensity score

- We are going to go over the propensity score in more detail, including better ways of specifying the propensity score, but here is a preview

```
* Estimate the propensity score
qui logit intervention ndrinks age female  drugrisk, nolog
predict double pscore if e(sample)

* Calculate statistics to check overlap
tabstat pscore, by(intervention) stats(N mean median min max)

Summary for variables: pscore
     by categories of: intervention (1 if received intervention)

intervention |         N      mean       p50       min       max
-------------+--------------------------------------------------
           0 |       209  .4343591  .4511057   .000152  .8146361
           1 |       243  .6264154  .6891437  .0836827  .8161083
-------------+--------------------------------------------------
       Total |       452  .5376106  .6060343   .000152  .8161083
----------------------------------------------------------------

* Create IP weight
gen     ipw = 1/pscore if intervention == 1
replace ipw = 1/(1-pscore) if intervention ==0
```

Check the lack of overlap

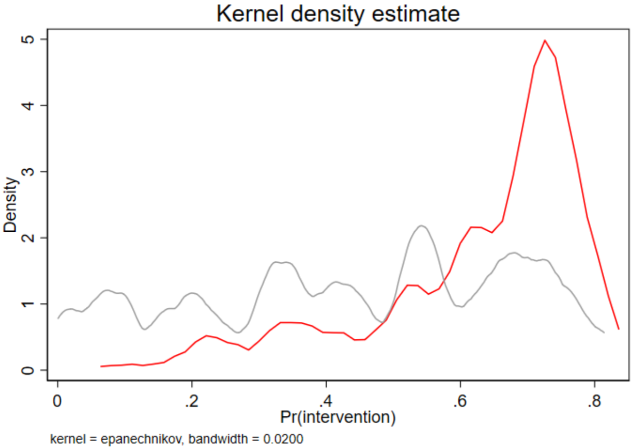
- Note the min and max above. The region where they overlap is the **common support** area
- What are the characteristics of those with PS less than 0.083 in the control group?

```
sum age female ndrinks drugrisk if intervention ==0 & pscore < 0.083

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |         22    41.59091    6.973822         31         56
      female |         22    .2727273    .4558423          0          1
     ndrinks |         22          74    26.35653         51        142
    drugrisk |         22    1.909091    5.126115          0         18
```

- Magic! They are the ones with ndrinks ≥ 51. Cool, isn't it? We knew that, but the overlap could be due to multiple variables at the same time
- The propensity score is also a **summary score** because in one number (scalar) that provides information on the distribution of all covariates $\boldsymbol{X}$

Check the distribution of the propensity score

```
kdensity pscore if intervention ==1, color(red) bw(0.02) ///
    addplot(kdensity pscore if intervention ==0, bw(0.02)) legend(off)
graph export ps_kernel.png, replace
```



Use the propensity score as a weight

- We are going to use the inverse of the propensity score as a weight. Analogous to survey design in which units are weighted based on the (inverse) probability of being surveyed
- The weight gives more importance to some observations. We can check sample characteristics using the weights

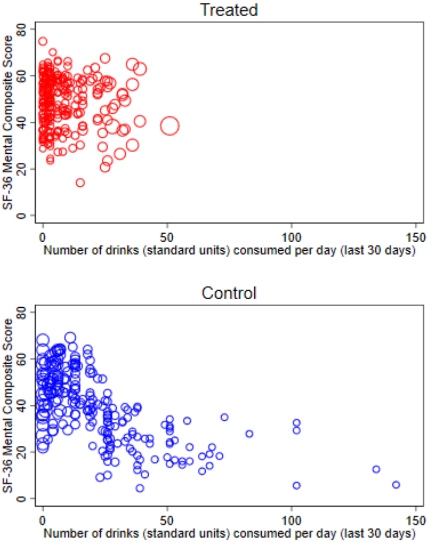
```
bysort intervention: sum age female ndrinks drugrisk [aweight=ipw]
---------------------------------------------------------------------------------------------
-> intervention = 0
    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         age |     209  450.540093    35.67488   8.281236         19         60
      female |     209  450.540093    .2286887   .4209967          0          1
     ndrinks |     209  450.540093    15.10925   19.02263          0        142
    drugrisk |     209  450.540093    1.810754   4.315663          0         21
---------------------------------------------------------------------------------------------
-> intervention = 1
    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         age |     243  441.817607    35.58435   7.058303         21         58
      female |     243  441.817607    .2210832   .4158329          0          1
     ndrinks |     243  441.817607    12.82159   13.24991          0         51
    drugrisk |     243  441.817607    1.925442   4.516877          0         21
```

- **Magic!!!!** Look how much better the balance is now. Before, average number of drinks was 8.09 and 23.03 for intervention and control. Now 15.10 and 12.8. All the other variables are closer too

Intuition: Use IP weights to change size of symbols

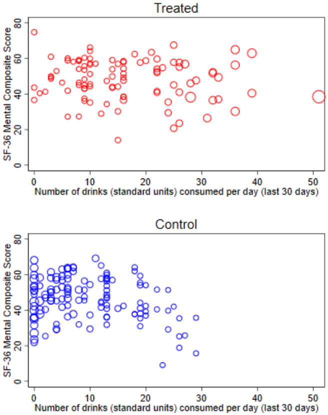
```
* Bubble plot
scatter pcs ndrinks [pweight=ipw] if intervention ==1, msymbol(circle_hollow) msize(small) ///
        color(red) saving(bubl_treated.gph, replace) title("Treated")
scatter pcs ndrinks [pweight=ipw] if intervention ==0, msymbol(circle_hollow) msize(small) ///
        color(blue) saving(bubl_cont.gph, replace) title("Control")
graph combine bubl_treated.gph bubl_cont.gph, col(1) xcommon ysize(10) xsize(8)
graph export buble.png, replace

* Keep IP weights larger than the median weight (ipw >1.55)
scatter pcs ndrinks [pweight=ipw] if intervention ==1 & ipw > 1.55, msymbol(circle_hollow) msize(small) ///
        color(red) saving(bubl_treated1.gph, replace) title("Treated")
scatter pcs ndrinks [pweight=ipw] if intervention ==0 & ipw > 1.55, msymbol(circle_hollow) msize(small) ///
        color(blue) saving(bubl_cont1.gph, replace) title("Control")
graph combine bubl_treated1.gph bubl_cont1.gph, col(1) xcommon ysize(10) xsize(8)
graph export buble1.png, replace
```

Intuition: Use IP weights to change size of symbols



Keep weights larger than median weights

- The 50% largest weights do not include any observation with ndrinks $>$ 51. Cool things: why is the weight for the treated observation (ndrinks around 50) so large? Go back to the graph with all the observations



Intuition about weights (see do file)

```
* Digression: some intuition about weights
preserve
  * make a smaller dataset so changes are easier to see
  keep if _n <=20
  gen w = 1

  * The regression below
  reg pcs age female ndrinks
  * is the same as regression in which everybody is given the same weights
  reg pcs age female ndrinks [pweight=w]

  * Now suppose we want the 20th observation to count for 10
  replace w = 10 if _n==20
  * the model below
  reg pcs age female ndrinks [pweight=w]
  est sto weighted
  * is the same as a model that creates 10 replicas of the 20th observation
  * Stata has a command for that: expand
  expand 10 if _n==20
  reg pcs age female ndrinks
  est sto expanded_noweight

  * The expanded version SEs need to be corrected
  est table weighted expanded_noweight, se stats(N)
restore
```

All magic tricks are illusions

- As the previous slides shows, the propensity score is a **balancing score**
- The analogy that it is like magic is actually accurate. It's also an illusion that has led, and continues to lead, to bad empirical research
- We have balance on **observed** variables, but not on unobservables. **We still need to assume ignorability**
- Showing that groups are balanced after using propensity scores helps make the case that you are reducing the overalp problem by giving more importance to some observations to achieve better balance
- But you still may not be controlling for all confounders
- We'll check balance using **standardized mean differences** and **variance ratios**

Outcome model

- We can now estimate the outcome model to obtain treatment effects (remember, we are **pretending** that we have ignorability)
- We use the inverse weight IPW, but we can also control for covariates in the outcome (we will dig deeper on this)

```
reg pcs intervention age female ndrinks drugrisk [pweight= ipw], robust
(sum of wgt is 892.3577007055283)
<... output omitted ...>
------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
intervention |   5.198854   1.201281     4.33   0.000     2.837981    7.559728
         age |  -.2360235   .0767722    -3.07   0.002    -.3869037   -.0851433
      female |  -5.687987   1.373758    -4.14   0.000     -8.38783   -2.988144
     ndrinks |  -.3369474   .0346501    -9.72   0.000    -.4050452   -.2688495
    drugrisk |  -.4088917   .1090059    -3.75   0.000    -.6231207   -.1946627
       _cons |   57.96827   2.828527    20.49   0.000     52.40937    63.52717
------------------------------------------------------------------------------
```

- Is 5.19 ATE? Well, yes, but also a sort of LATE. We are giving more importance to some observations
- Not that different from regression adjustment (teffects ra): 6.58

Preview

- Just to preview results, we can do the same with command teffects ipwra
- There are some key differences. teffects ipwra estimates the propensity score and the outcome model simultaneously using GMM (SEs are correct) and the outcome model follows the logic of teffects ra
- With teffects you can check balance and do other fun things

```
. teffects ipwra (pcs age female ndrinks drugrisk) ///
>              (intervention ndrinks  age female drugrisk)
Iteration 0:   EE criterion =  2.140e-21
Iteration 1:   EE criterion =  9.134e-30
Treatment-effects estimation                    Number of obs     =        452
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATE          |
intervention |
   (1 vs 0)  |   5.670275   1.212875     4.68   0.000     3.293084    8.047465
-------------+----------------------------------------------------------------
POmean       |
intervention |
          0  |   42.36709   .9414163    45.00   0.000     40.52195    44.21223
------------------------------------------------------------------------------
```

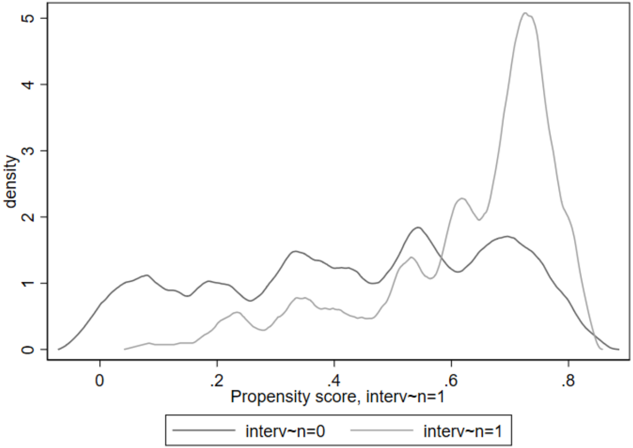
Preview

- Using postestimation commands for teffects
- Rule of thumb is that standardized difference should be less than 0.25 (absolute value). Ideally, ratio of variances should be close to 1
- Below, raw is the observed differences. We went from ndrinks being 0.83 (high) to 0.13 (acceptable). Variance ratio still problematic, but not as important. Maybe we should just focus the comparison restricting to ndrinks ≤ 51 (i.e. some form of matching)

```
tebalance summarize
Covariate balance summary
                                                  Raw     Weighted
                     -----------------------------------------
                     Number of obs =          452        452.0
                     Treated obs   =          243        223.8
                     Control obs   =          209        228.2
                     -----------------------------------------
  -----------------------------------------------------------------
                 |Standardized differences          Variance ratio
                 |        Raw     Weighted           Raw   Weighted
  ----------------+------------------------------------------------
          ndrinks |  -.8314587   -.1395673      .1725134   .4855282
              age |  -.1650646   -.0117661      .7451948   .7270095
           female |   .1998801   -.0181769      1.289522   .9763603
         drugrisk |  -.0786428    .0259628      .7077652   1.096254
  -----------------------------------------------------------------
```

Check the distribution of the propensity score - teffects

```
qui teffects ipwra (pcs age female ndrinks drugrisk) ///
             (intervention ndrinks  age female drugrisk)
teffects overlap, ptl(1)
graph export ps_kernel_te.png, replace
```



Loose ends: matching teffects ipwra

- We can match teffect ipwra manually. Remember that GMM estimates both steps at the same time, so SEs are better. ATE is the difference of POMs

```
/// --- Matching IPWRA
reg pcs age female ndrinks drugrisk [pweight= ipw] if intervention ==1
predict double pom_t
reg pcs age female ndrinks drugrisk [pweight= ipw] if intervention ==0
predict double pom_c

mean pom_c pom_t
Mean estimation                   Number of obs   =        452

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       pom_c |   42.36709   .4157372      41.55007    43.18411
       pom_t |   48.03737   .2113928      47.62193     48.4528
--------------------------------------------------------------

teffects ipwra (pcs age female ndrinks drugrisk) ///
             (intervention ndrinks  age female drugrisk), pom
------------------------------------------------------------------------------
             |               Robust
         pcs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
POmeans      |
intervention |
          0  |   42.36709   .9414163    45.00   0.000     40.52195    44.21223
          1  |   48.03737   .8567154    56.07   0.000     46.35823     49.7165
------------------------------------------------------------------------------
```

Important considerations

- We could improve the specification of the propensity score. At minimum, an interactions between ndrinks and other variables. We don't have large sample sizes in this example. We could even try a nonparameetric or semiparametric propensity score
- Of course, there is the issue of picking and choosing. Choose the specification that gives the larger treatment effect. In this, Stata failed: tebalance summarize is only available *after* you estimate treatment effects. At least we should use the **quietly** command before teffects
- We want to choose the PS specification that achieves balance, not the one that makes treatment effects go in the direction we want
- There is a chi-square test developed to check for balance (see do file). In this example, we don't achieve balance
- We could try matching or a stratified analysis that would essentially amount to ignoring those with ndrinks > 51 – a type of LATE