CHPR290/STAT266 2018

April

Su	Mo	Tu	We	Th	Fr	Sa	
1	2	3	4	5	6	7	W1
8	9	10	11	12	13	14	W2
15	16	17	18	19	20	21	W3
22	23	24	25	26	27	28	W4
29	30						W5

May

Su	Mo	Tu	We	Th	Fr	Sa	
		1	2	3	4	5	
6	7	8	9	10	11	12	W6
13	14	15	16	17	18	19	W7
20	21	22	23	24	25	26	W8
27	~~28~~	29	30	31			Holiday

June

Su	Mo	Tu	We	Th	Fr	Sa	
					1	2	
3	4	5	6	7	8	9	W9,DW
10	11	12	13	14	15	16	Exam Week

class meetings
take home exercises posted
take home exercises due

CHPR290/Stat266/HRP292-- Course Files, Readings, Examples

Week 1--Course Introduction; Matching Methods part 1 (intro and theory)

In the news

Going to concerts helps you live longer (UK) Going to Concerts Can Help You Live Longer, (Newsweek) Going to a concert is better for you than yoga (NYPost) Science says gig-going can help you live longer and increases wellbeing (o2 UK) Going to Concerts Is Good for Your Health (Variety)

Lecture Topics Lecture 1 slide deck

1. Course outline and logistics
2. A matched observational study (DOS, Chap 7)
3. Study design versus inference
4. Basic tools of multivariate matching (DOS, Secs 8.1-8.4)

Text Readings

Rosenbaum DOS: Chapters 7 and 8 (8.1-8.4)

Additional Resources

Observational Studies according to Donald B. Rubin

For objective causal inference, design trumps analysis Annals of Applied Statistics, Volume 2, Number 3 (2008), 808-840. Rubin talk . Another Rubin overview of matching: Matching Methods for Causal Inference Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. Chapter 11 (pp. 155-176) in Best Practices in Quantitative Social Science. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications.

Computing Corner: Extended Data Analysis Examples

Lalonde NSW data (DOS sec 2.1). Subclassification/Stratification and Full matching.

Week 1 handout Rogosa R-session (using R 3.3.3) 4/1/18 redo in R 3.4.4 (sparse) pdf slides shown in class MatchIt vignette

Week 1 Review Questions

From Computing Corner

1. In Week 1 Computing Corner with the Lalonde data (effect of job training on earnings), we started out (see R-session) by showing the ubiquitous [epidemiology to economics] analysis for observational data of an analysis of covariance, aka tossing the treatment variable and all the confounders into a regression equation predicting outcome and hoping for the best (c.f 2016 Week 1 *in the news* analyses: mom fish consumption on child cognition). The statement made in class (technical details week 1 stat209) is that regression does not "control" for confounders; instead the coefficient of treament (putative causal effect) is obtained from a straight-line regression of outcome on the residuals from a prediction of treatment by all the other predictors in the regression. Demonstrate that equivalence using the ancova in CC1.

Solution for Review Question 1

2. RQ1 uses the Week 1 Computing Corner Lalonde data (effect of job training on earnings) analysis of covariance: tossing the treatment variable and all the confounders into a regression equation predicting outcome and hoping for the best. Compare that ancova with an ancova the uses just the significant predictors of re78. Also compare with an ancova which uses the single available covariate/confounder having the highest correlation with outcome. Are these analyses consistent?

Solution for Review Question 2

From Lecture

3. We will be working a lot with matching based techniques. One of the best thinkers/writers on the topic of matching is Elizabeth Stuart from Johns Hopkins. For this problem, take a look at her paper: "Matching Methods for Causal Inference: A Review and a Look Forward." In lecture 01 you were introduced to "balance tables" (a.k.a. "Table 1") which summarizes the covariate distribution of the observations. A handful of questions: (a) as concisely as possible, state why we focus on balance assessments as part of our argumentation when attempting to perform causal inference, (b) in addition to a balance table, name other tools used to report balance, (c) why do we use standardized mean differences instead of p-values to assess balance when assessing the quality of a match design?, and (d) why is it kinda weird to use a p-value of the covariates in a randomized trial to assess balance?

Solution for Review Question 3

4. In lecture 1 we quickly outlined some of the big challenges to causal inference when using observational data (see slide 41, "There should be strong effort to show the two groups are similar..."). These challenges include: inclusion/exclusion of observations, observational units that may be completely missing (censored, survival bias), missing data, imbalances in observed data, and imbalances in unobserved data. We'll address each of these at different points in the course. But let's focus on the decision to include/exclude observations. What we're doing when matching -- i.e., removing observations that do not have adequate counterparts

Package 'MatchIt'

February 22, 2017

Version 2.4-22

Date 2017-02-22

Title Nonparametric Preprocessing for Parametric Casual Inference

Author Daniel Ho <daniel.e.ho@gmail.com>,
Kosuke Imai <kimai@Princeton.Edu>,
Gary King <king@harvard.edu>,
Elizabeth Stuart <stuart@stat.harvard.edu>

Maintainer Kosuke Imai <kimai@Princeton.Edu>

Depends R (>= 2.6), MASS

Suggests cem, optmatch, Matching, nnet, rpart, mgcv, WhatIf, R.rsp

VignetteBuilder R.rsp

Description Selects matched samples of the original treated and control groups with similar covariate distributions -- can be used to match exactly on covariates, to match on propensity scores, or perform a variety of other matching procedures. The package also implements a series of recommendations offered in Ho, Imai, King, and Stuart (2007) <DOI:10.1093/pan/mpl013>.

LazyLoad yes

LazyData yes

License GPL (>= 2)

URL http://gking.harvard.edu/matchit

NeedsCompilation no

Repository CRAN

Date/Publication 2017-02-22 14:13:05

R topics documented:

> vignette(package = "MatchIt")
Vignettes in package 'MatchIt': linked
matchit MatchIt: Nonparametric Preprocessing for Parametric Causal Inference
(source, pdf)

help.matchit *HTML Help for Matchit Commands and Models*

Description

The help.matchit command launches html help for Matchit commands and supported methods. The full manual is available online at http://gking.harvard.edu/matchit.

Usage

```
help.matchit(object)
```

Arguments

object	a character string representing a Matchit command or model. help.matchit("command") will take you to an index of Matchit commands and help.matchit("method") will take you to a list of matching methods. The following inputs are currently available: exact, nearest, subclass, full, optimal.

Author(s)

Daniel Ho <<daniel.ho@yale.edu>>; Kosuke Imai <<kimai@princeton.edu>>; Gary King <<king@harvard.edu>>; Elizabeth Stuart<<estuart@jhsph.edu>>

See Also

The complete document is available online at http://gking.harvard.edu/matchit.

lalonde *Data from National Supported Work Demonstration and PSID, as analyzed by Dehejia and Wahba (1999).*

Description

This is a subsample of the data from the treated group in the National Supported Work Demonstration (NSW) and the comparison sample from the Current Population Survey (CPS). This data was previously analyzed extensively by Lalonde (1986) and Dehejia and Wahba (1999). The full dataset is available at http://www.columbia.edu/~rd247/nswdata.html.

Usage

```
data(lalonde)
```

Format

A data frame with 614 observations (185 treated, 429 control). There are 10 variables measured for each individual. "treat" is the treatment assignment (1=treated, 0=control). "age" is age in years. "educ" is education in number of years of schooling. "black" is an indicator for African-American (1=African-American, 0=not). "hispan" is an indicator for being of Hispanic origin (1=Hispanic, 0=not). "married" is an indicator for married (1=married, 0=not married). "nodegree" is an indicator for whether the individual has a high school degree (1=no degree, 0=degree). "re74" is income in 1974, in U.S. dollars. "re75" is income in 1975, in U.S. dollars. "re78" is income in 1978, in U.S. dollars.

Source

http://www.columbia.edu/~rd247/nswdata.html

References

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76: 604-620.

Dehejia, R.H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94: 1053-1062.

`match.data` *Output Matched Data Sets*

Description

`match.data` outputs matched data sets from `matchit()`.

Usage

```
  match.data(object, group="all", distance = "distance",
weights = "weights", subclass = "subclass")
```

Arguments

`object`	The output object from `matchit`. This is a required input.
`group`	This argument specifies for which matched group the user wants to extract the data. Available options are `"all"` (all matched units), `"treat"` (matched units in the treatment group), and `"control"` (matched units in the control group). The default is `"all"`.
`distance`	This argument specifies the variable name used to store the distance measure. The default is `"distance"`.
`weights`	This argument specifies the variable name used to store the resulting weights from matching. The default is `"weights"`.
`subclass`	This argument specifies the variable name used to store the subclass indicator. The default is `"subclass"`.

```
R version 3.2.2 (2015-08-14) -- "Fire Safety"  #### Week 1 session. Lalonde data
# If you start from a relatively clean install, get MatchIt and optmatch
# some years order matters because of complication with license for optmatch algorithms this year appea

> install.packages("optmatch")
> library(optmatch)
> install.packages("MatchIt")
> library(MatchIt)

################################
> data(lalonde) # in MatchIt package
# get lalonde data from MatchIt, there are different versions in existence under this name
help(lalonde)  #produces
-----------------------------------
lalonde               package:MatchIt               R Documentation

Data from National Supported Work Demonstration and PSID, as analyzed by Dehejia and Wahba (1999).

Description:

     This is a subsample of the data from the treated group in the
     National Supported Work Demonstration (NSW) and the comparison
     sample from the Current Population Survey (CPS).   This data was
     previously analyzed extensively by Lalonde (1986) and Dehejia and
     Wahba (1999).   The full dataset is available at <URL:
     http://www.columbia.edu/~rd247/nswdata.html>. [note: broken link still in current documentation]

Usage:

     data(lalonde)

Format:

     A data frame with 313 [sic, 614] observations (185 treated, 429 control).
     There are 10 variables measured for each individual.  "treat" is
     the treatment assignment (1=treated, 0=control).   "age" is age in
     years.  "educ" is education in number of years of schooling.
     "black" is an indicator for African-American (1=African-American,
     0=not).  "hispan" is an indicator for being of Hispanic origin
     (1=Hispanic, 0=not).  "married" is an indicator for married
     (1=married, 0=not married).  "nodegree" is an indicator for
     whether the individual has a high school degree (1=no degree,
     0=degree). "re74" is income in 1974, in U.S. dollars.  "re75" is
     income in 1975, in U.S. dollars.  "re78" is income in 1978, in
     U.S. dollars.


References:

     Lalonde, R. (1986). Evaluating the econometric evaluations of
     training programs  with experimental data. American Economic
     Review 76: 604-620.

     Dehejia, R.H. and Wahba, S. (1999).  Causal Effects in
     Nonexperimental Studies:  Re-Evaluating the Evaluation  of
     Training Programs.  Journal of the American  Statistical
     Association 94: 1053-1062.
-----------------------------------------------
> dim(lalonde)
[1] 614  10
> attach(lalonde)
> table(treat)  # so these summaries synch with data description
treat
  0   1
429 185
```

```
> head(lalonde)
     treat age educ black hispan married nodegree re74 re75       re78
NSW1     1  37   11     1      0       1        1    0    0  9930.0460
NSW2     1  22    9     0      1       0        1    0    0  3595.8940
NSW3     1  30   12     1      0       0        0    0    0 24909.4500
NSW4     1  27   11     1      0       0        1    0    0  7506.1460
NSW5     1  33    8     1      0       0        1    0    0   289.7899
NSW6     1  22    9     1      0       0        1    0    0  4056.4940

################## prelim compare groups on outcome measure
> tapply(re78, treat, median)
       0        1
4975.505 4232.309
> tapply(re78, treat, fivenum)
$`0`
[1]     0.0000   220.1813  4975.5050 11688.8200 25564.6700
$`1`
[1]     0.0000   485.2298  4232.3090  9642.9990 60307.9300

> t.test(re78 ~ treat)
        Welch Two Sample t-test
data:  re78 by treat
t = 0.93773, df = 326.41, p-value = 0.3491
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -697.192 1967.244
sample estimates:
mean in group 0 mean in group 1
       6984.170        6349.144

> wilcox.test(re78 ~ treat, conf.int = TRUE)
        Wilcoxon rank sum test with continuity correction
data:  re78 by treat
W = 41840, p-value = 0.2818
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -4.664401e-05  1.053159e+03
sample estimates:
difference in location
          5.053114e-05

> #####But wait, some say "we are never done until the ancova is run" see Fish
> # as we see the social science, life science practice is to put in the treatment variable  and
> # a whole bunch of other variables to "control" for self-selection, nonequivalence etc.
> #  equivalent to analysis of covariance by whatever name
>  ancova.lalonde = lm( re78 ~ treat + age + educ + black + hispan + married + nodegree + re74 + re75)
>  summary(ancova.lalonde)
Call:
lm(formula = re78 ~ treat + age + educ + black + hispan + married +
    nodegree + re74 + re75)

Residuals:
   Min     1Q Median     3Q    Max
-13595  -4894  -1662   3929  54570

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.651e+01  2.437e+03   0.027   0.9782
treat        1.548e+03  7.813e+02   1.982   0.0480 *
age          1.298e+01  3.249e+01   0.399   0.6897
educ         4.039e+02  1.589e+02   2.542   0.0113 *
black       -1.241e+03  7.688e+02  -1.614   0.1071
hispan       4.989e+02  9.419e+02   0.530   0.5966
married      4.066e+02  6.955e+02   0.585   0.5590
nodegree     2.598e+02  8.474e+02   0.307   0.7593
re74         2.964e-01  5.827e-02   5.086 4.89e-07 ***
```

Standard Analysis (ancova)

OUTCOME ~ TREATMENT + (binary, contin)

CONFOUNDERS (controls)

see FISH (in the news) example

American Journal of Epidemiology

Vol. 183, No. 3
DOI: 10.1093/aje/kwv195
Advance Access publication:
January 5, 2016

Original Contribution

Maternal Consumption of Seafood in Pregnancy and Child Neuropsychological Development: A Longitudinal Study Based on a Population With High Consumption Levels

Jordi Julvez*, Michelle Méndez, Silvia Fernandez-Barres, Dora Romaguera, Jesus Vioque, Sabrina Llop, Jesus Ibarluzea, Monica Guxens, Claudia Avella-Garcia, Adonina Tardón, Isolina Riaño, Ainara Andiarena, Oliver Robinson, Victoria Arija, Mikel Esnaola, Ferran Ballester, and Jordi Sunyer

* Correspondence to Dr. Jordi Julvez, ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona Biomedical Research Park, C. Doctor Aiguader 8, 08003 Barcelona, Spain (e-mail: jjulvez@creal.cat).

Initially submitted March 13, 2015; accepted for publication July 14, 2015.

Seafood consumption during pregnancy is thought to be beneficial for child neuropsychological development, but to our knowledge no large cohort studies with high fatty fish consumption have analyzed the association by seafood subtype. We evaluated 1,892 and 1,589 mother-child pairs at the ages of 14 months and 5 years, respectively, in a population-based Spanish birth cohort established during 2004–2008. Bayley and McCarthy scales and the Childhood Asperger Syndrome Test were used to assess neuropsychological development. Results from multivariate linear regression models were adjusted for sociodemographic characteristics and further adjusted for umbilical cord blood mercury or long-chain polyunsaturated fatty acid concentrations. Overall, consumption of seafood above the recommended limit of 340 g/week was associated with 10-g/week increments in neuropsychological scores. By subtype, in addition to lean fish, consumption of large fatty fish showed a positive association; offspring of persons within the highest quantile (>238 g/week) had an adjusted increase of 2.29 points in McCarthy general cognitive score (95% confidence interval: 0.42, 4.16). Similar findings were observed for the Childhood Asperger Syndrome Test. Beta coefficients diminished 15%–30% after adjustment for mercury or long-chain polyunsaturated fatty acid concentrations. Consumption of large fatty fish during pregnancy presents moderate child neuropsychological benefits, including improvements in cognitive functioning and some protection from autism-spectrum traits.

Table 3. Associations Between Maternal Seafood Consumption in the First Trimester of Pregnancy and Child's Score on the McCarthy General Cognitive Scale at Age 5 Years, Spanish Childhood and Environment (INMA) Project, 2004–2008

Seafood Intake[a]	No. of Subjects	Difference in Child's Neurobehavioral Score[b]			
		Minimally Adjusted[c]		Fully Adjusted[d]	
		β	95% CI	β	95% CI
All Seafood					
Continuous variable, 10 g/week[e]	1,589	0.03[f]	0.00, 0.05	0.02[g]	0.00, 0.05
Quintiles					
1[h]	320	0.00	Referent	0.00	Referent
2	340	1.91[g]	−0.23, 4.04	1.61	−0.43, 3.65
3	299	3.46[f]	1.24, 5.67	2.13[f]	0.00, 4.26
4	323	3.60[f]	1.41, 5.79	2.84[f]	0.74, 4.94
5	308	2.93[f]	0.72, 5.14	2.08[g]	−0.04, 4.21
P for trend		0.007		0.049	
Large Fatty Fish					
Continuous variable, 10 g/week	1,589	0.10[f]	0.02, 0.17	0.06[g]	−0.00, 0.13
Quartiles					
1	704	0.00	Referent	0.00	Referent
2	285	2.99[f]	1.05, 4.93	2.26[f]	0.40, 4.11
3	296	2.36[f]	0.43, 4.30	1.93[f]	0.09, 3.79
4	304	3.46[f]	1.51, 5.40	2.29[f]	0.42, 4.16
P for trend		0.001		0.02	
Small Fatty Fish					
Continuous variable, 10 g/week	1,589	−0.03	−0.11, 0.05	−0.03	−0.10, 0.05
Quartiles					
1	736	0.00	Referent	0.00	Referent
2	280	1.41	−0.61, 3.44	0.60	−1.33, 2.53
3	288	0.94	−0.98, 2.87	1.25	−0.59, 3.10
4	285	1.27	−0.67, 3.21	0.91	−0.93, 2.76
P for trend		0.18		0.25	

Table continues

Median intakes in specific quantiles (Q), in g/week:
total seafood—Q1, 195; Q2, 338; Q3, 461; Q4, 600; Q5, 854;
large fatty fish—Q1, none; Q2, 48; Q3, 92; Q4, 238;

Abbreviations: CI, confidence interval; INMA, Infancia y Medio Ambiente.

[a] Median seafood intake in each quantile (g/week) is shown in the Table 2 footnotes.

[b] McCarthy Scales of Children's Abilities (16).

[c] Regression models adjusted for sex of the child, age during testing, cohort, quality of the test, and maternal energy intake (kcal/day) during pregnancy.

[d] Regression models additionally adjusted for child's birth weight, gestational age, duration of breastfeeding, maternal age, educational level, social class, prepregnancy body mass index, parity, and country of origin/birth.

[e] Per 10-g/week increase.

[f] $P < 0.05$.

[g] $P < 0.10$.

[h] Results were similar when the reference group included all mothers with seafood consumption less than or equal to 340 g/week (Web Table 5). Further inclusion of all seafood subtypes in the final model showed similar association patterns (data not shown).

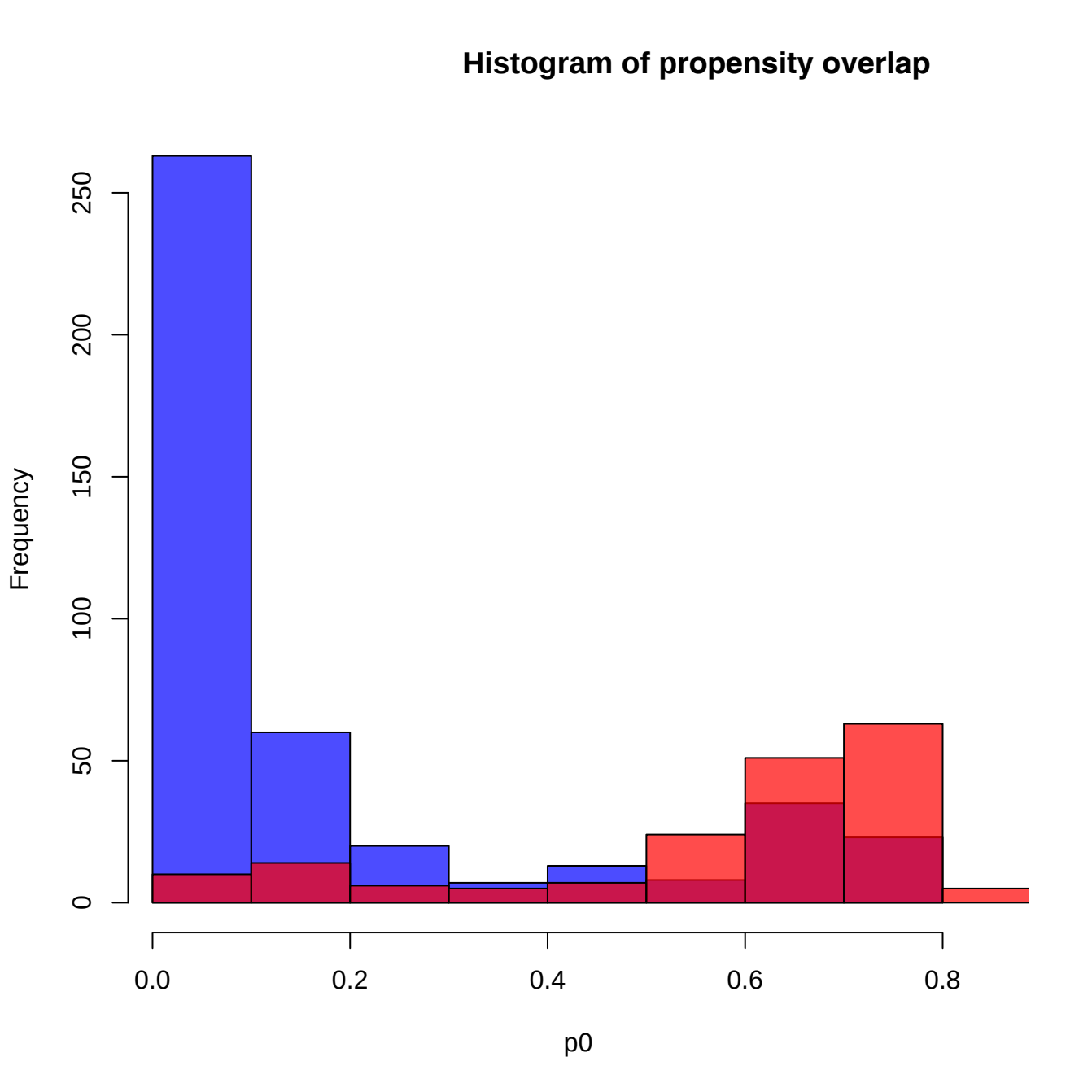
```
re75         2.315e-01  1.046e-01   2.213   0.0273 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6948 on 604 degrees of freedom
Multiple R-squared:  0.1478,    Adjusted R-squared:  0.1351
F-statistic: 11.64 on 9 and 604 DF,  p-value: < 2.2e-16

> # so treatment is significantly helpful ??
```

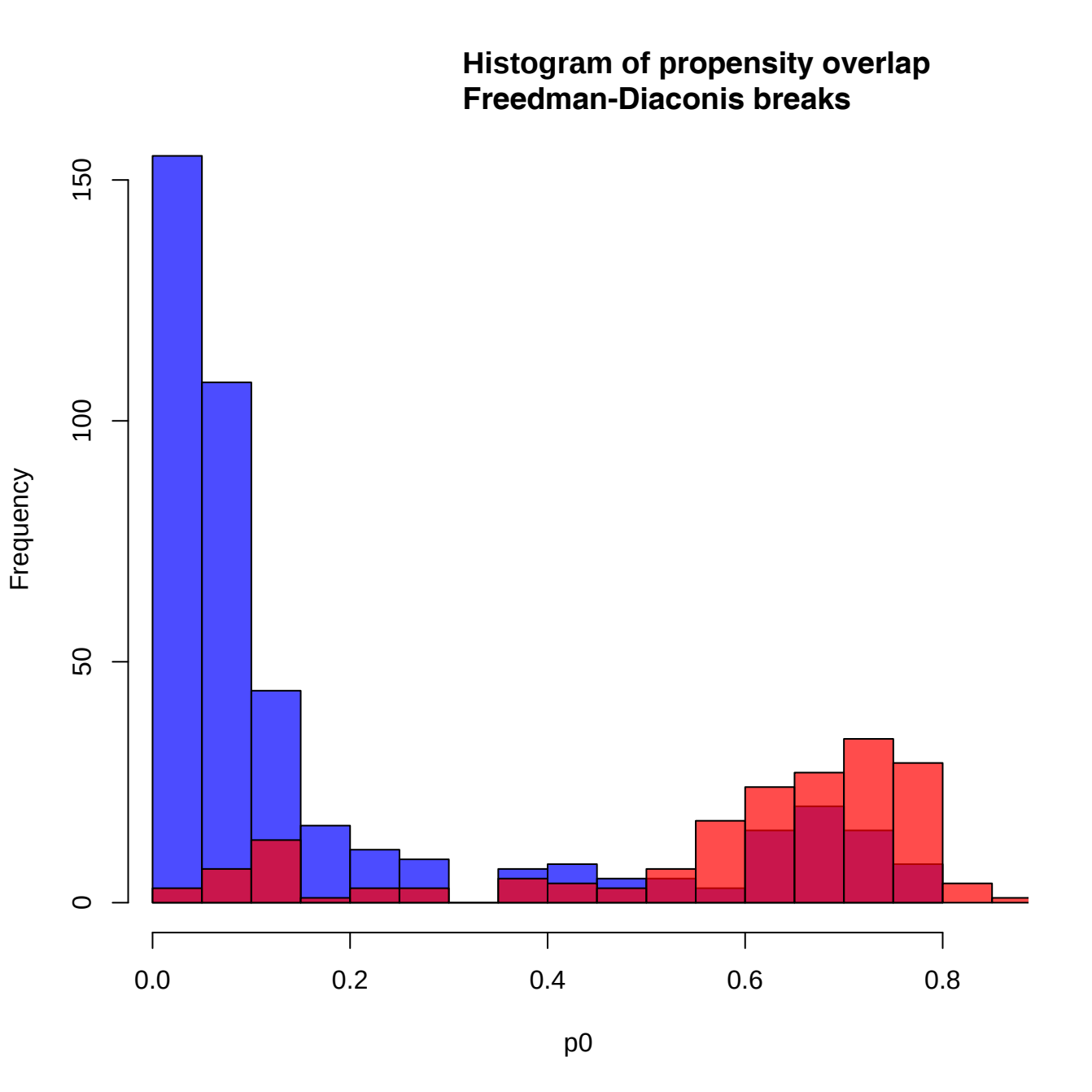
First by hand, then by algorithm

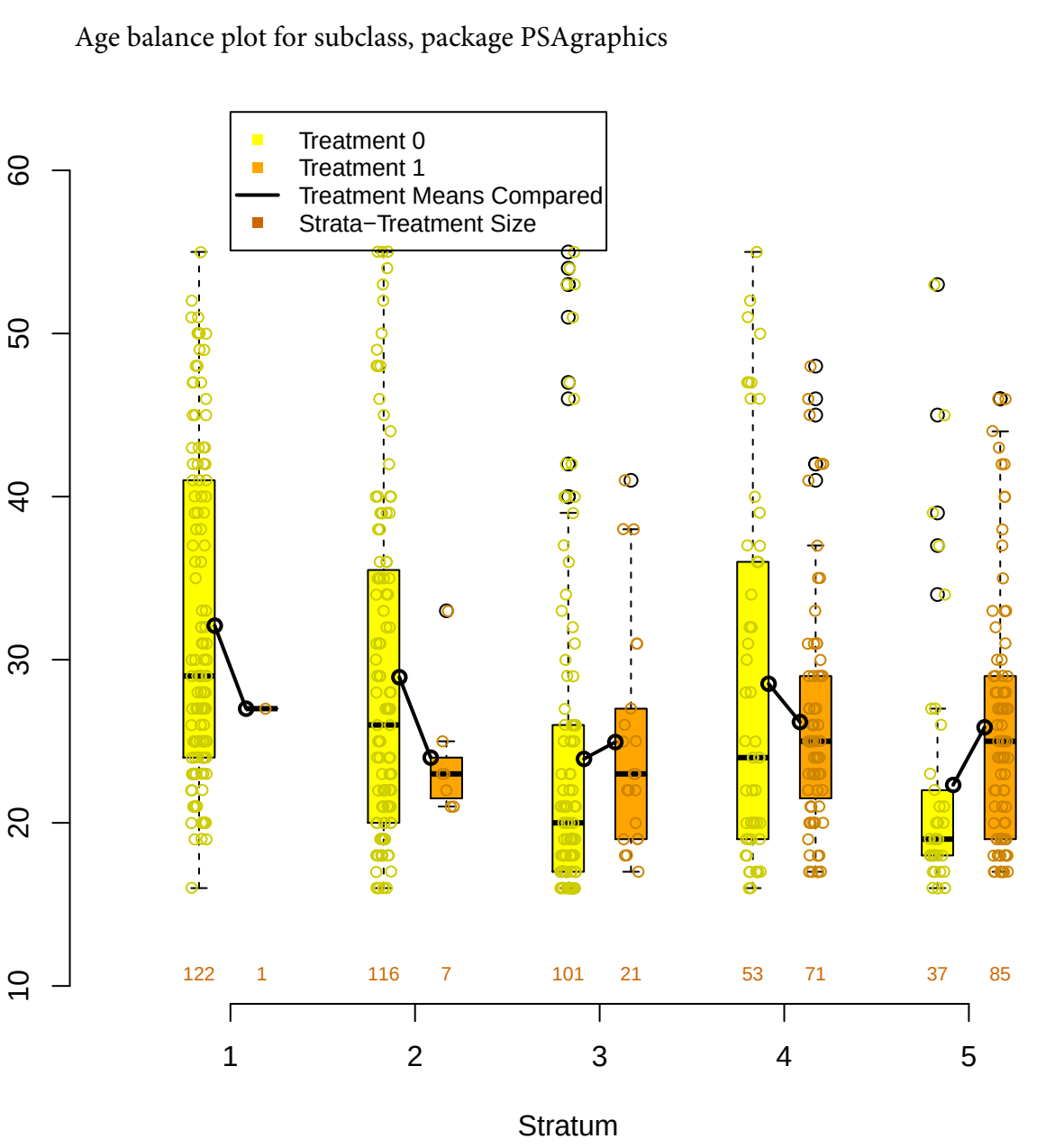
```
########### Begin matching analysis; Quintile Subclassification with Propensity Scores
## original Rosenbaum-Rubin, cardiac; Rubin breast cancer

> # now do the logistic regression that computes propensity scores (matching packages will do this for
>  glm.p = glm( treat ~ age + educ + black + hispan + married + nodegree + re74 + re75, data = lalonde,
>  summary(glm.p)
Call:
glm(formula = treat ~ age + educ + black + hispan + married +
    nodegree + re74 + re75, family = binomial, data = lalonde)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7645  -0.4736  -0.2862   0.7508   2.7169

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.729e+00  1.017e+00  -4.649 3.33e-06 ***
age          1.578e-02  1.358e-02   1.162  0.24521
educ         1.613e-01  6.513e-02   2.477  0.01325 *
black        3.065e+00  2.865e-01  10.699  < 2e-16 ***
hispan       9.836e-01  4.257e-01   2.311  0.02084 *
married     -8.321e-01  2.903e-01  -2.866  0.00415 **
nodegree     7.073e-01  3.377e-01   2.095  0.03620 *
re74        -7.178e-05  2.875e-05  -2.497  0.01253 *
re75         5.345e-05  4.635e-05   1.153  0.24884
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 751.49  on 613  degrees of freedom
Residual deviance: 487.84  on 605  degrees of freedom
AIC: 505.84
Number of Fisher Scoring iterations: 5

> propen = fitted(glm.p)  # now we have the propensity scores
> quantile(propen) # overall distrib
         0%          25%          50%          75%         100%
0.009080193 0.048536484 0.120676493 0.638715991 0.853152844
> tapply(propen, treat, quantile) # look at overlap via 5-number summary (or side-by-side boxplots) not
$`0`
         0%          25%          50%          75%         100%
0.009080193 0.038880745 0.075849106 0.195135746 0.789172834
$`1`
        0%         25%         50%         75%        100%
0.02495179 0.52646352 0.65368426 0.72659995 0.85315284

> # as we are fitting prob(treat = 1) fits for those in treatment group will be larger, we need good ov
> boxplot(propen ~ treat) #gives side-by-side boxplots, you can add labels, not wonderful overlap
> detach(lalonde)
> lalonde$propen = propen
> attach(lalonde)

#### looking at overlap, histograms
>  p1 = propen[treat == 1]
>  length(p1)
[1] 185
>  p0 = propen[treat == 0]
```

0.0
0.2
0.4
0.6
0.8
0
1

```
>  length(p0)
[1] 429
>  fivenum(p1)
    NSW124     NSW156      NSW50     NSW119     NSW178
0.02495179 0.52646352 0.65368426 0.72659995 0.85315284
>  fivenum(p0)
    PSID296     PSID347     PSID221     PSID334     PSID118
0.009000193 0.030000745 0.075049186 0.195135746 0.789172834
>  hist(p0,col=rgb(0,0,1,0.7),xlim=range(c(p0,p1)))
>  hist(p1,col=rgb(1,0,0,0.7),add=T)
>  # superimposed propensity histograms, like Ben Hansen SAT, contol is blue, treatment is red, overlap
> hist(p0, breaks = "FD", col=rgb(0,0,1,0.7),xlim=range(c(p0,p1)))
> hist(p1, breaks = "FD", col=rgb(1,0,0,0.7),add=T)

### make quintiles of propensity distribution
>  pbin = cut(propen, quantile(propen, seq(0, 1, 1/5)), include.lowest = TRUE, labels = FALSE)
> detach(lalonde)
> lalonde$bins = pbin
> attach(lalonde)
>  table(pbin, treat)
    treat
pbin   0   1
   1 122   1
   2 116   7
   3 101  21
   4  53  71
   5  37  85

##### examples of checking balance (more to come)
> tapply(age, list(bins, treat), median)
   0  1
1 29 27
2 26 23
3 20 23
4 24 25
5 19 25

> ### install.packages("PSAgraphics")
>  library(PSAgraphics)
>  box.psa(age, treat, bins)

##################################### examine outcome by strata
> tapply(re78, list(bins, treat),mean)  # here are the mean diffs in re78 (the outcome) stratified by p
          0        1
1 10467.064    0.000
2  5796.548 7919.316
3  6043.316 9210.726
4  4977.401 5819.143
5  4666.221 6030.258
> # direction of mean diffs favors treatment, job training

> # contrast that with the comparison ignoring any concerns about self-selection (selection bias), effe
>  tapply(re78, treat, mean)
       0        1
6984.170 6349.144

> ##### can do t-tests by subclassification (strata)
> # e.g. for the 3 upper quintiles is the mean difference significant? since we are doing 3 of these be
> ##  we won't find any evidence for the effectiveness of job training looking at each of the subclasse
>
> ##### lmer, a better way to do the t-tests  ########################
> library(lme4)
Loading required package: Matrix
> propen.lmer = lmer(re78 ~ treat + (1 + treat|bins), data = lalonde)
> summary(propen.lmer)
Linear mixed model fit by REML ['lmerMod']
```


Histogram of propensity overlap
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
p0


Histogram of propensity overlap
Freedman-Diaconis breaks
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
p0


Age balance plot for subclass, package PSAgraphics
Treatment 0
Treatment 1
Treatment Means Compared
Strata−Treatment Size
60
50
40
30
20
10
122
1
116
7
101
21
53
71
37
85
1
2
3
4
5
Stratum

Optmatch: Flexible, Optimal Matching for Observational Studies

Ben B. Hansen

Observational studies compare subjects who received a specified treatment to others who did not, without controlling assignment to treatment and comparison groups. When the groups differ at baseline in ways that are relevant to the outcome, the study has to adjust for the differences. An old and particularly direct method of making these adjustments is to match treated subjects to controls who are similar in terms of their pretreatment characteristics, then conduct an outcome analysis conditioning upon the matched sets. Adjustments of this type enjoy properties of robustness (Rubin, 1979) and transparency not shared with purely model-based adjustments, such as covariance adjustment without matching or stratification; and with the introduction of propensity scores to matching (Rosenbaum and Rubin, 1985), the approach was shown to be more broadly applicable than was previously thought. Arguably, the reach of techniques based on matching now exceeds that of purely model-based adjustment (Hansen, 2004).

To achieve these benefits, matched adjustment requires the analyst to articulate a distinction between desirable and undesirable potential matches, and then to match treated and control subjects in such a way as to favor the more desirable pairings. Propensity scoring fits under the first of these tasks, as do the construction of Mahalanobis matching metrics (Rosenbaum and Rubin, 1985), prognostic scoring (Hansen, 2006b), and the distance metric optimization of Diamond and Sekhon (2006). The second task, matching itself, is less statistical in nature, but doing it well can substantially improve the power and robustness of matched inference (Hansen and Klopfer, 2006; Hansen, 2004). The main purpose of **optmatch** is to relieve the analyst of responsibility for this important, if potentially tedious, undertaking, freeing attention for other aspects of the analysis. Given discrepancies between each treatment and control subject that might potentially be matched, **optmatch** places them into non-overlapping matched sets, in the process solving the discrete optimization problems needed to make sums of matched discrepancies as small as possible; after this, the analysis can proceed using permutation inference (Rosenbaum, 2002; Hothorn et al., 2006; Bowers and Hansen, 2006), conditional inference (Breslow and Day, 1980; Cox and Snell, 1989; Hansen, 2004; Lumley and Therneau, 2006), approximately conditional inference (Pierce and Peters, 1992; Brazzale, 2005; Brazzale et al., 2006), or multilevel models (Smith, 1997; Raudenbush and Bryk, 2002; Gelman and Hill, 2006).

Optimal matching of two groups

To illustrate the meaning of optimal matching, consider Cox and Snell's (1981, p.81) study of costs of nuclear power. Of 26 light water reactor plants constructed in the U.S. between 1967 and 1972, seven had been built on the site of existing plants. The problem is to estimate the cost benefit (or penalty) of building on an existing site as opposed to a new one. A matched analysis seeks to adjust for background characteristics determinative of cost, such as the date of construction and the capacity of the plant, by linking similar refurbished and new plants: plants of about the same capacity and constructed at about the same time, for example. To highlight the analogy with intervention studies, I refer to existing-site plants as "treatments" and new-site plants as "controls."

Consider the problem of arranging the plants in disjoint triples, each containing one treatment and two controls, placing each treatment and 14 of the 19 controls into some matched triple or another. A straightforward way to create such a match is to move down the list of treatments, pairing each to the two most similar controls that have not yet been matched; this is *nearest-available matching*. Figure 1 shows the 26 plants, their capacities and dates of construction, and a 1 : 2 matching constructed in this way. First A was matched to I and J, then B to L and N, and so forth. This example is discussed by Rosenbaum (2002, ch.10).

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

"date" is date of construction, in years after 1965; "capacity" is net capacity of the power plant, in MWe above 400.

Figure 1: 1:2 matching by a nearest-available algorithm.

How might this process be improved? To complete step i, the nearest-available algorithm requires

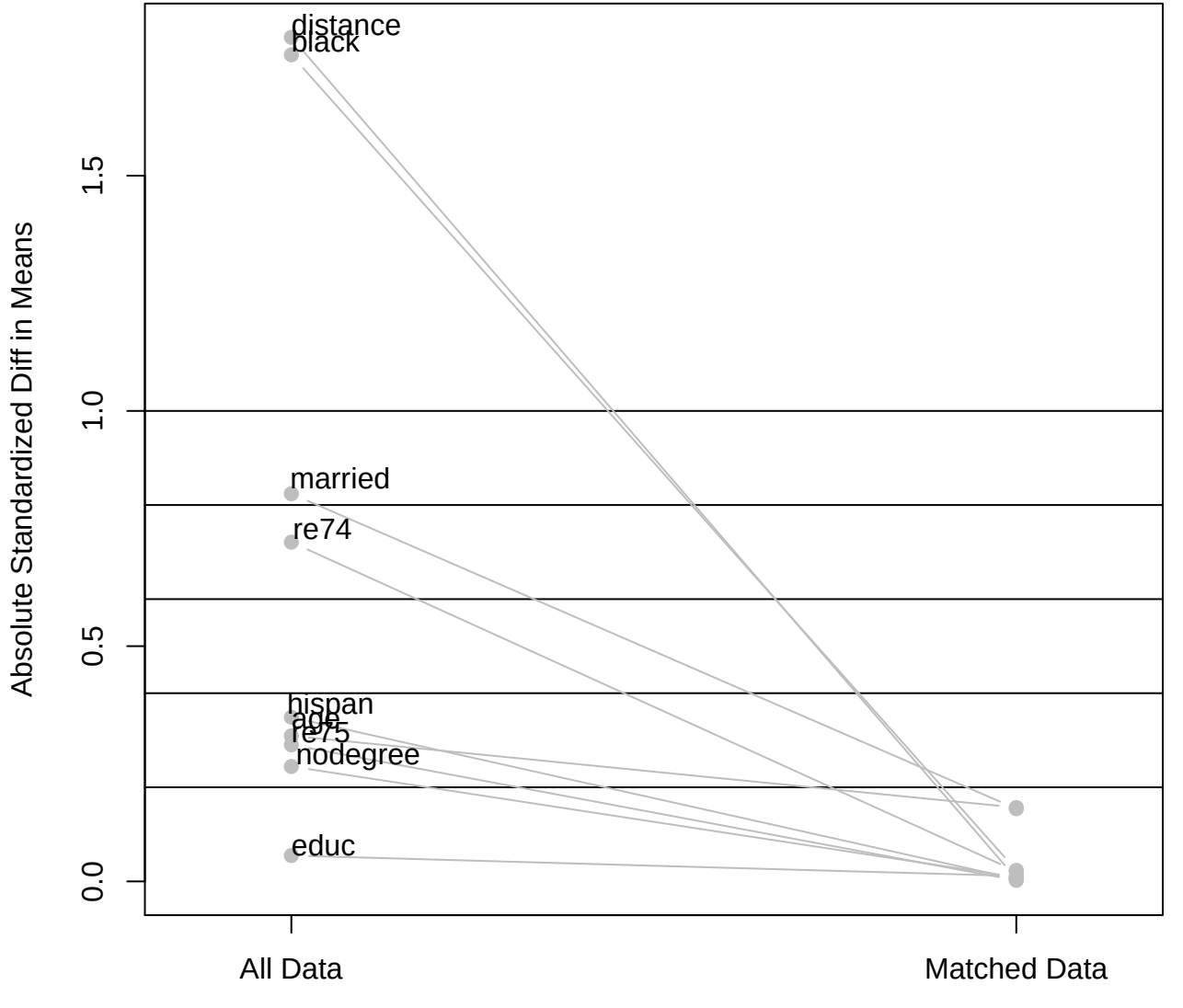
ISSN 1609-3631

```
Formula: re78 ~ treat + (1 + treat | bins)
   Data: lalonde

REML criterion at convergence: 12637.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.3976 -0.7541 -0.2878  0.5408  7.4535

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 bins     (Intercept)  5208943 2282
          treat        2069963 1439     -1.00
 Residual             52597981 7252
Number of obs: 614, groups:  bins, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6434.2     1090.2   5.902
treat          385.7      950.8   0.406

Correlation of Fixed Effects:
      (Intr)
treat -0.795

# so here we have an overall estimate of the effect of the treat on re78 of positive $386, but
# far from significant. Much smaller point estimate than in some of the individual strata

> confint(propen.lmer) # bombs
> confint(propen.lmer, method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...

                  2.5 %   97.5 %
.sig01        414.81230 4084.578
.sig02         -1.00000    1.000
.sig03         54.74858 3644.981
.sigma       6846.49101 7654.434
(Intercept)  4432.91940 8695.198
treat       -1681.75647 2565.802

  some bootstrap runs failed (7/1000)
```

```
########################## Full Matching (Hansen, via Rosenbaum, using MatchIt)

>  m2full.out = matchit(treat ~ re74 + re75 + educ + black + hispan + age + married + nodegree, data =
Warning message:
In fullmatch(d, ...) :
  Without 'data' argument the order of the match is not guaranteed
    to be the same as your original data.

>  summary(m2full.out)
Call:
matchit(formula = treat ~ re74 + re75 + educ + black + hispan +
    age + married + nodegree, data = lalonde, method = "full")

Summary of balance for all data:
         Means Treated Means Control  Mean Diff   eQQ Med  eQQ Mean   eQQ Max
distance        0.5774        0.1822     0.3952    0.5176    0.3955    0.5966
re74         2095.5737     5619.2365 -3523.6628 2425.5720 3620.9240 9216.5000
re75         1532.0553     2466.4844  -934.4291  981.0968 1060.6582 6795.0100
educ           10.3459       10.2354     0.1105    1.0000    0.7027    4.0000
black           0.8432        0.2028     0.6404    1.0000    0.6432    1.0000
hispan          0.0595        0.1422    -0.0827    0.0000    0.0811    1.0000
age            25.8162       28.0303    -2.2141    1.0000    3.2649   10.0000
```

```
married          0.1892        0.5128    -0.3236    0.0000    0.3243    1.0000
nodegree         0.7081        0.5967     0.1114    0.0000    0.1135    1.0000


Summary of balance for matched data:
         Means Treated Means Control Mean Diff  eQQ Med eQQ Mean   eQQ Max
distance        0.5774        0.5761    0.0013   0.0026   0.0066     0.096
re74         2095.5737     2199.7126 -104.1390  72.6510 512.7210 13121.750
re75         1532.0553     1524.8362    7.2191 209.6655 460.5643 12746.050
educ           10.3459       10.3227    0.0233   0.0000   0.4596     4.000
black           0.8432        0.8347    0.0086   0.0000   0.0020     1.000
hispan          0.0595        0.0583    0.0012   0.0000   0.0012     1.000
age            25.8162       24.6928    1.1235   3.0000   3.3100     9.000
married         0.1892        0.1285    0.0607   0.0000   0.0544     1.000
nodegree        0.7081        0.7040    0.0041   0.0000   0.0028     1.000

Percent Balance Improvement:
         Mean Diff.   eQQ Med eQQ Mean  eQQ Max
distance    99.6662   99.5001  98.3388  83.9052
re74        97.0446   97.0048  85.8401 -42.3724
re75        99.2274   78.6295  56.5775 -87.5796
educ        78.9494  100.0000  34.5954   0.0000
black       98.6582  100.0000  99.6891   0.0000
hispan      98.5858    0.0000  98.5200   0.0000
age         49.2583 -200.0000  -1.3825  10.0000
married     81.2495    0.0000  83.2267   0.0000
nodegree    96.3435    0.0000  97.5333   0.0000

Sample sizes:
          Control Treated
All           429     185
Matched       429     185
Unmatched       0       0
Discarded       0       0



> summary(m2full.out, standardize = T)
Call:
matchit(formula = treat ~ re74 + re75 + educ + black + hispan +
    age + married + nodegree, data = lalonde, method = "full")

Summary of balance for all data:
         Means Treated Means Control Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance        0.5774        0.1822          1.7941   0.3964    0.3774   0.6444
re74         2095.5737     5619.2365         -0.7211   0.2335    0.2248   0.4470
re75         1532.0553     2466.4844         -0.2903   0.1355    0.1342   0.2876
educ           10.3459       10.2354          0.0550   0.0228    0.0347   0.1114
black           0.8432        0.2028          1.7568   0.3202    0.3202   0.6404
hispan          0.0595        0.1422         -0.3489   0.0414    0.0414   0.0827
age            25.8162       28.0303         -0.3094   0.0827    0.0813   0.1577
married         0.1892        0.5128         -0.8241   0.1618    0.1618   0.3236
nodegree        0.7081        0.5967          0.2443   0.0557    0.0557   0.1114


Summary of balance for matched data:
         Means Treated Means Control Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance        0.5774        0.5761          0.0060   0.0060    0.0085   0.0596
re74         2095.5737     2199.7126         -0.0213   0.0160    0.0476   0.2268
re75         1532.0553     1524.8362          0.0022   0.0348    0.0693   0.2324
educ           10.3459       10.3227          0.0116   0.0286    0.0275   0.0568
black           0.8432        0.8347          0.0236   0.0104    0.0104   0.0208
hispan          0.0595        0.0583          0.0049   0.0036    0.0036   0.0072
age            25.8162       24.6928          0.1570   0.0416    0.0857   0.3436
married         0.1892        0.1285          0.1545   0.0366    0.0366   0.0732
nodegree        0.7081        0.7040          0.0089   0.0008    0.0008   0.0016
```

```
Percent Balance Improvement:
          Std. Mean Diff. eCDF Med eCDF Mean  eCDF Max
distance          99.6662  98.4863   97.7452   90.7506
re74              97.0446  93.1488   78.8321   49.2658
re75              99.2274  74.3198   48.3597   19.2062
educ              78.9494 -25.6137   20.8722   48.9995
black             98.6582  96.7523   96.7523   96.7523
hispan            98.5858  91.2972   91.2972   91.2972
age               49.2583  49.7246   -5.3122 -117.8448
married           81.2495  77.3817   77.3817   77.3817
nodegree          96.3435  98.5634   98.5634   98.5634

Sample sizes:
           Control Treated
All            429     185
Matched        429     185
Unmatched        0       0
Discarded        0       0

>  plot(summary(m2full.out, standardize = T))
[1] "To identify the variables, use first mouse button; to stop, use second."
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
warning: nearest point already identified
integer(0)
>    setwd("D:\\drr16\\somgen290\\week1\\")
>  plot(m2full.out)
Waiting to confirm page change...
Waiting to confirm page change...
> # gives you QQ plots for each var

> detach(lalonde)
>  m2full.dat = match.data(m2full.out)  # obtain results from the full matching
>  head(m2full.dat)
     treat age educ black hispan married nodegree re74 re75       re78    propen bins  distance weights
NSW1     1  37   11     1      0       1        1    0    0  9930.0460 0.6387699    4 0.6387699       1
NSW2     1  22    9     0      1       0        1    0    0  3595.8940 0.2246342    3 0.2246342       1
NSW3     1  30   12     1      0       0        0    0    0 24909.4500 0.6782439    5 0.6782439       1
NSW4     1  27   11     1      0       0        1    0    0  7506.1460 0.7763241    5 0.7763241       1
NSW5     1  33    8     1      0       0        1    0    0   289.7899 0.7016387    5 0.7016387       1
NSW6     1  22    9     1      0       0        1    0    0  4056.4940 0.6990699    5 0.6990699       1

> dim(m2full.dat)
[1] 614  15
> head(m2full.dat)
     treat age educ black hispan married nodegree re74 re75       re78    propen bins  distance weights
NSW1     1  37   11     1      0       1        1    0    0  9930.0460 0.6387699    4 0.6387699       1
NSW2     1  22    9     0      1       0        1    0    0  3595.8940 0.2246342    3 0.2246342       1
NSW3     1  30   12     1      0       0        0    0    0 24909.4500 0.6782439    5 0.6782439       1
NSW4     1  27   11     1      0       0        1    0    0  7506.1460 0.7763241    5 0.7763241       1
NSW5     1  33    8     1      0       0        1    0    0   289.7899 0.7016387    5 0.7016387       1
NSW6     1  22    9     1      0       0        1    0    0  4056.4940 0.6990699    5 0.6990699       1
> attach(m2full.dat)

>  # so you can see match.data appends 3 colums "distance" "weights"  "subclass" to the original data s
>  table(m2full.dat$subclass) #the 104 subclasses have various sizes

  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26
  2  13   2   7   3   5   3   2   4   2   8   3   2   2   9   4   2   9   6  14   3   2   2   6   3   4
 36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61
 14   5   3   3   2   6   2   5   3   2  10   2   4   8   3   2  14   7   2  14   2   2   4  40   2   2
 71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96
```



Other Examples: (propen success) SAT coaching ASPIRIN

```
  2   3  70   2   5   6   2   2  13   2   2   2   2   2   7   3   2   2   3   2   2   2   2   3   6   4

###### outcome comparison over the (matched) subclasses
> mfull.lmer = lmer(re78 ~ treat + (1 + treat|subclass), data = m2full.dat) # like for the quintiles in
> summary(mfull.lmer)
Linear mixed model fit by REML ['lmerMod']
Formula: re78 ~ treat + (1 + treat | subclass)
   Data: m2full.dat

REML criterion at convergence: 12633.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5267 -0.7497 -0.2851  0.5165  7.4616

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subclass (Intercept)  4027897 2007
          treat        3810081 1952     -0.89
 Residual             51103906 7149
Number of obs: 614, groups:  subclass, 104

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5862.9      507.8  11.546
treat          504.5      736.2   0.685 ## about the same as seen in base section 384 (952)

Correlation of Fixed Effects:
      (Intr)
treat -0.679
> confint(mfull.lmer)
Computing profile confidence intervals ...
                2.5 %   97.5 %
.sig01      1216.8647 3011.968
.sig02        -1.0000    1.000
.sig03         0.0000      Inf
.sigma      6740.8624 7581.414
(Intercept) 4807.1941 6873.722
treat       -985.7685 1977.973
There were 50 or more warnings (use warnings() to see the first 50)
>
```

Software (R software with no guarantees)

Two R Packages for Sensitivity Analysis in Observational Studies

sensitivitymv (R package at cran)

sensitivitymw (Rpackage at cran)

"A new u-statistic..." Biometrics 2011 R-Session (Supplement 2): txt.document

Match Functions from Design of Observational Studies R workspace

Selected Data Sets from Design of Observational Studies R workspace

Appendix 3.9 from Design of Observational Studies R workspace

Software supplement to "Imposing minimax constraints..." pdf aamatch package local files zip tar.gz

Suggested R Packages for Matching

Ben Hanson's optmatch (at cran)

Sam Pimentel's rcbalance (at cran)

Bo Lu, Robert Greevy, Xinyi Xu and Cole Beck's nbpMatching (at cran)

Dan Yang's finebalance package (archived but working at cran)

Jose Zubizarreta's mipmatch (requires special installation)

Adaptive sensitivity analysis

Dylan Small's SensitivityCaseControl (at cran) including adaptive.noether.brown

Stat 209

Week 8 Propensity Scores

Let $z = 1, 0$ T/C $\underset{\sim}{X}$ vector of covariates

propensity score $e(\underset{\sim}{X}) = Pr(z=1 | \underset{\sim}{X})$ scalar $\hat{e}(\underset{\sim}{X})$

cond'l prob unit w vector $\underset{\sim}{X}$ observed cov. assigned to $T(z=1)$

Thm Balancing score $b(\underset{\sim}{X})$ s.t. conditional distrib of $\underset{\sim}{X}$ given $b(\underset{\sim}{X})$ same of treated and control units

$\underset{\sim}{X} \perp\!\!\!\perp z | b(\underset{\sim}{X})$. Coarsest (low dimen) balancing score is propensity score. $Pr(X, z|e) = Pr\{\underset{\sim}{X}|e\} Pr(z|e)$

Thm (result) Approx 90% reduction in bias for subclassifying at quintiles of population propensity score. $B_I = E(f(\underset{\sim}{x})|z=1) - E(f(\underset{\sim}{x})|z=0)$, B_S after strat, treatm percent reduction in bias $100(1 - B_S/B_I) \approx 90\%$

Ros Rubin
1983 Biometrika
1984 JASA

(i) The propensity score is a balancing score.
(ii) Any score that is 'finer' than the propensity score is a balancing score; moreover, x is the finest balancing score and the propensity score is the coarsest.
(iii) If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score
(iv) At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects.
(v) Using sample estimates of balancing scores can produce sample balance on x.

Applications: Rubin Breast Cancer, Love (RR '84) CAD, Love Aspirin, Hansen SAT coaching, Substance Abuse (UNC), Rosenbaum, Danish downers

Rubin AnnIntMedicine

Table 3: Estimated 5-year Survival Rates for Node-Negative Patients in SEER from Tables 5 and 7 in U.S. GAO Report (1994). AIM pub

Propensity Score

Subclass	Treatment	n	Estimate	n*	Estimate*
1	Breast Conservation	56	85.6%	54	88.8%
	Mastectomy	1,008	86.7%	966	90.5%
2	Breast Conservation	106	82.8%	102	86.0%
	Mastectomy	964	82.8%	917	87.7%
3	Breast Conservation	193	85.2%	184	89.4%
	Mastectomy	866	88.8%	841	91.4%
4	Breast Conservation	289	88.7%	279	92.0%
	Mastectomy	978	87.3%	742	91.5%
5	Breast Conservation	462	89.0%	453	90.7%
	Mastectomy	604	88.5%	589	90.7%

* omitting patients whose deaths were unrelated to cancer.

Lalonde data

Lab 4 stratification

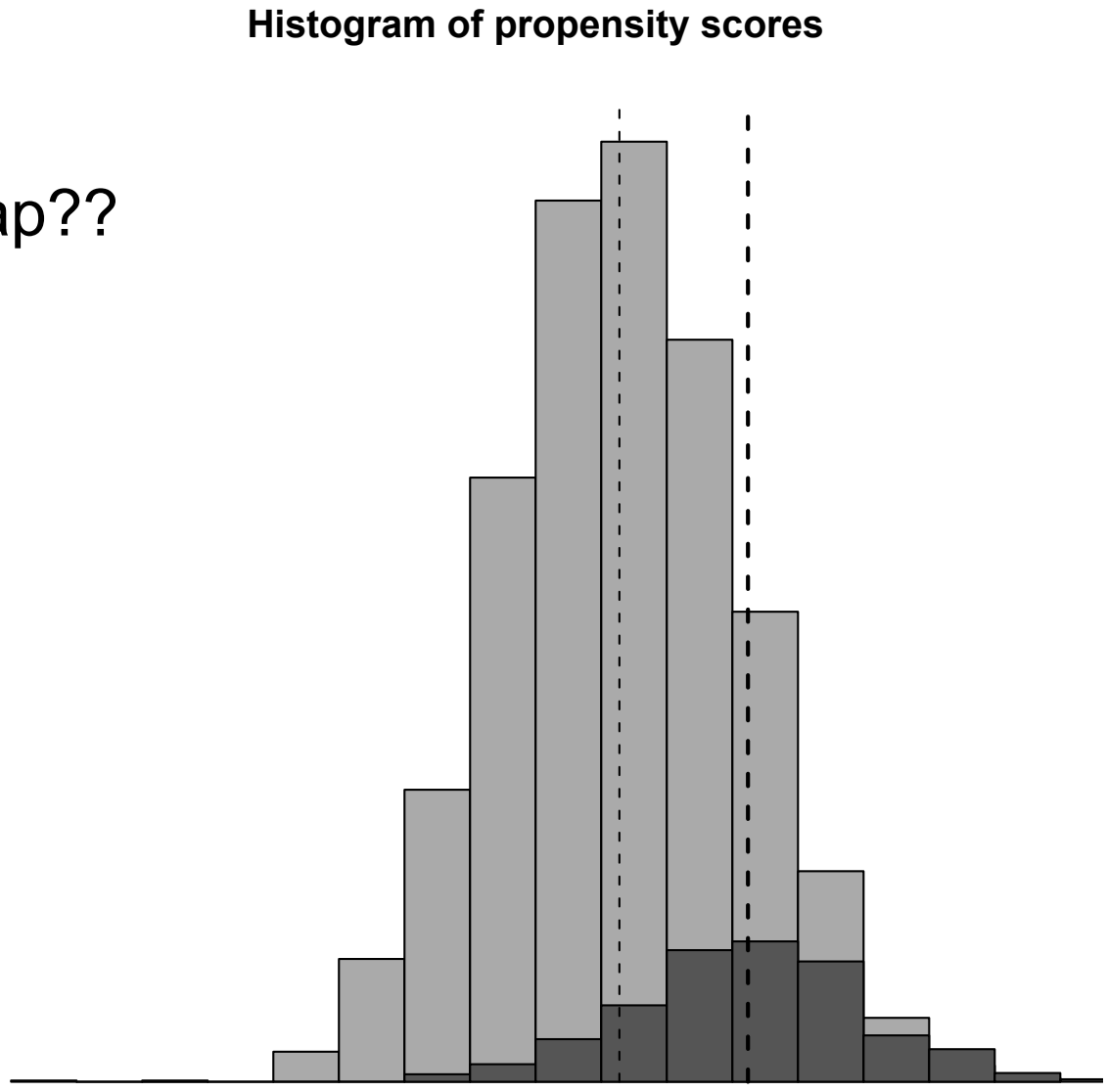
```
> table(propbin, treat)
                 treat
propbin            0   1
  (0,0.0401]     122   1
  (0.0401,0.0872] 116   7
  (0.0872,0.27]  101  21
  (0.27,0.671]    53  71
  (0.671,1]       37  85
> tapply(re78, list(propbin, treat),mean)
                    0    1
(0,0.0401]      10467    0
(0.0401,0.0872]  5797 7919
(0.0872,0.27]    6043 9211
(0.27,0.671]     4977 5819
(0.671,1]        4666 6030
```

counts

means re78

The challenge of matching on the propensity to be coached

Overlap??



fullmatch

Table 1. Selected Pretreatment Variables

Variable	Range of values	Standardized bias	Percentage of sample
Math section of PSAT	20–43	−.1	18
	45–51	.1	17
	52–57	−.1	16
	58–80	.1	15
	Not taken	.1	34
Mean SAT at respondent's first-choice college	787–987	−.3	16
	988–1,060	−.2	16
	1,061–1,123	.1	16
	1,124–1,336	.3	16
	No response	.0	36
Father's education	High school	−.4	40
	A.A. or B.A.	−.1	26
	Graduate	.4	25
	No response	.2	9
Average math grade	"Excellent"	.1	35
	"Good"–"fail"	−.1	59
	No response	.1	6
Foreign language years taken	0–2	−.3	64
	3–4	.3	27
	No response	.1	9

well as scores on previous SAT–I or PSAT tests and their answers to the Student Descriptive Questionnaire (SDQ), which all SAT–I registrants are asked to complete. By their responses to questions about extracurricular SAT preparation, respondents split into a treated and a control group, and the data describe the results of a classical quasiexperiment (Campbell and Stanley 1966).

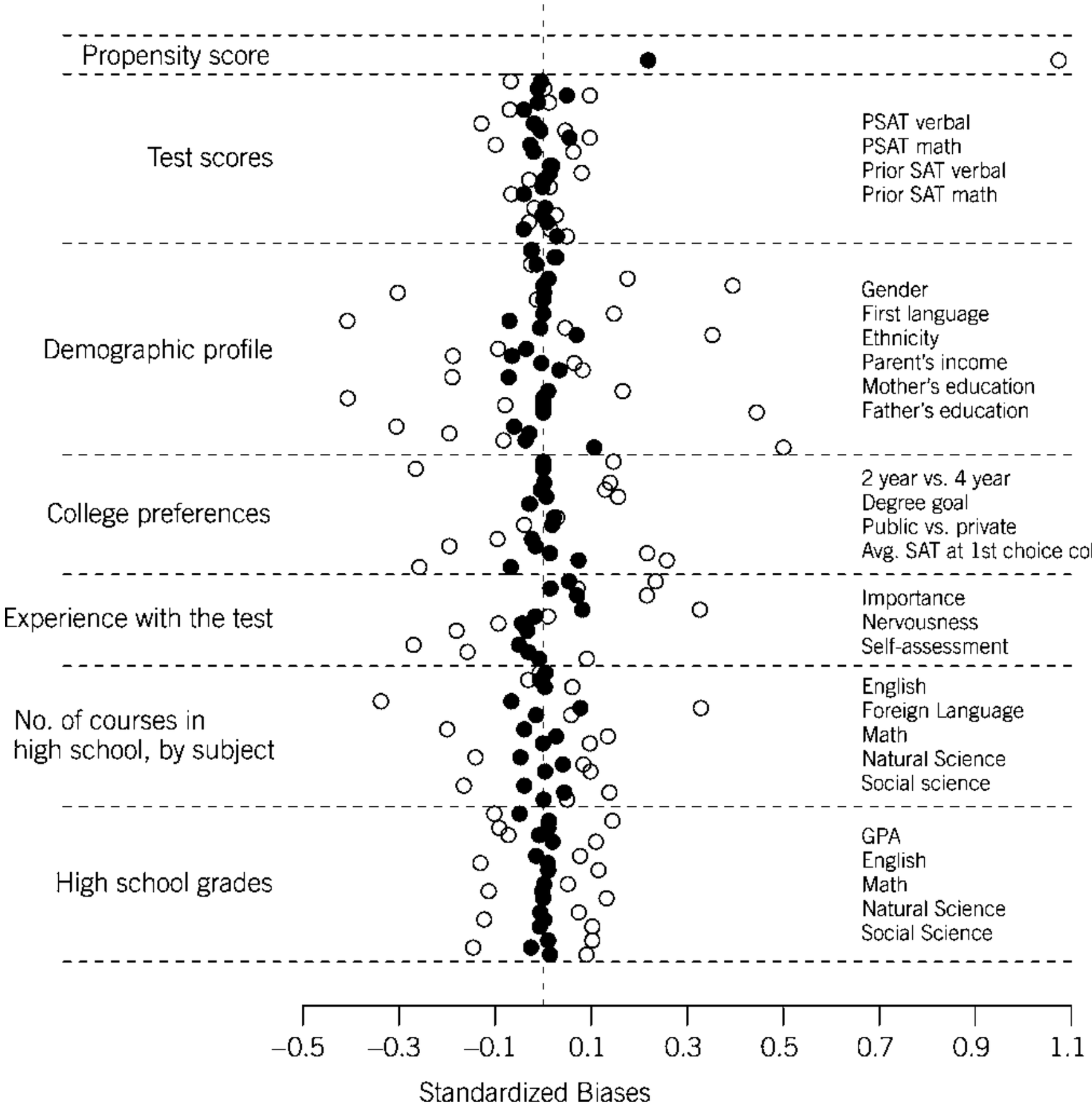
Nineteen in twenty of the survey respondents actually took the spring 1996 or fall 1995 exam for which they had registered. The analysis given below restricts itself to these 3,994 students, using the corresponding SAT scores as outcome measures. Thus the record gives coaching status and SAT outcomes for all students in the sample to be analyzed; among the additional measures, each available for some fraction of the students, are pretest scores, racial and socioeconomic indicators, various data about their academic preparation, and responses to a survey item that, by eliciting students' first choices in colleges, recovered an unusually discriminating measure of students' educational aspirations. In all, there are 27 pretreatment variables.

The coached and uncoached groups differ appreciably in these recorded measures—as do high and low scorers on the SAT. Table 1 offers some illustration of this, giving overall incidences of various covariate attributes and comparing their relative incidences in the coached and uncoached groups. (The statistic here used to effect these comparisons is the *standardized bias*, given for a variable v by $(\bar{v}_t - \bar{v}_c)/s_p$, where $\bar{v}_t$ and $\bar{v}_c$ are the average values of v in the treatment and control groups, respectively, and s_p^2 is the pooled within-group variance in v.) Yet the table shows only five covariates; the analysis must address biases on all 27 of them.

1.2 Missing and Misleading Data in Regression and in Subclassification

In regression-based adjustment, the simplest way to handle missing data on a covariate is to reject cases without complete information. In adjustment based on matching or stratification,

see RItools, PSAgraphics



Multivariate Matching with the Propensity Score

- Match subjects so that they balance on multiple covariates using one scalar score.
- Goal: Emulate a RCT in matching, then use standard analyses to compare matched sets.
- Design: Treated subjects matched to people who didn't receive treatment but who had similar propensity to receive treatment (match the treated to untreated "clones").

Pairmatch 1:1

Aspirin Use and Mortality

- 6174 consecutive adults at CCF undergoing stress echocardiography for evaluation of known or suspected coronary disease.
- 2310 (37%) were taking aspirin (treatment).
- Main Outcome: all-cause mortality
- Median follow-up: 3.1 years
- Univariate Analysis: 4.5% of aspirin patients died, and 4.5% of non-aspirin patients died…
- Unadjusted Hazard Ratio: 1.08 (0.85, 1.39)

Gum et al. (2001) http://www.ncbi.nlm.nih.gov/pubmed/11559263

JAMA. 2001 Sep 12;286(10):1187-94. http://jama.jamanetwork.com/article.aspx?articleid=194177
Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. Gum PA1, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS.

Propensity Score Model for Aspirin Use

- Logistic Regression predicting aspirin use
- 31 covariates included in the model:
  - Demographics, Clinical history, Medication use
  - Cardiovascular assessment and Exercise capacity
- Estimated propensity scores for aspirin use range from .03 to .98
  - ROC Area shows good discrimination (C = .83)
- But does the propensity score model work?
- Are the covariates balanced?

Baseline Characteristics By Aspirin Use (in %) (before matching)

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	P value
Men	77.0	56.1	< .001
Clinical history: diabetes	16.8	11.2	< .001
hypertension	53.0	40.6	< .001
prior coronary artery disease	69.7	20.1	< .001
congestive heart failure	5.5	4.6	.12
Medication use: Beta-blocker	35.1	14.2	< .001
ACE inhibitor	13.0	11.4	< .001

- Baseline characteristics appear very dissimilar: 25 of 31 covariates have $p < .001$, 28 of 31 have $p < .05$.
- Aspirin user covariates indicate higher mortality risk.

How Were The Aspirin Subjects Matched?

- Tried to match each aspirin user to a unique non-user with a PS identical to 5 digits.
- If not possible, proceeded to a 4-digit match, then 3-digit, 2-digit, and finally a 1-digit match (i.e., propensity scores within .099).
- Result: matches for 1351 (58%) of the 2310 aspirin patients to 1351 unique non-users.

SAS macro: http://www2.sas.com/proceedings/sugi26/p214-26.pdf

Propensity Matcher Results

ID	Treated?	Propensity	Linear Propensity	Match?	Partner ID
1	1	0.2	-1.386	No	-999
2	1	0.3	-0.847	Yes	8
3	1	0.4	-0.405	Yes	10
4	1	0.6	0.405	No	-999
5	1	0.7	0.847	No	-999

logit

SE (Linear Propensity):	0.1829
x % Selected:	0.6
x % of SE:	0.1097

Baseline Characteristics By Aspirin Use [%] (after matching)

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P value
Men	70.4	72.1	.33
Clinical history: diabetes	15.0	15.3	.83
hypertension	50.3	51.7	.46
prior coronary artery disease	48.3	48.8	.79
congestive heart failure	5.8	6.6	.43
Medication use: Beta-blocker	26.1	26.5	.79
ACE inhibitor	15.5	15.8	.79

- Baseline characteristics similar in matched users and non-users.
- 30 of 31 covariates show NS difference between matched users and non-users. [Peak exercise capacity for men is p = .01]

Using Standardized Differences to Measure Covariate Balance

- Standardized Differences are appropriate summaries of Covariate Balance for both Continuous and Categorical Variables

$$d = \frac{100\left(\bar{x}_{Treatment} - \bar{x}_{Control}\right)}{\sqrt{\frac{s^2_{Treatment} + s^2_{Control}}{2}}}$$ for continuous variables

$$d = \frac{100\left(p_{Treatment} - p_{Control}\right)}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}}$$ for binary variables

|Standardized Differences| > 10% Indicate Serious Imbalance

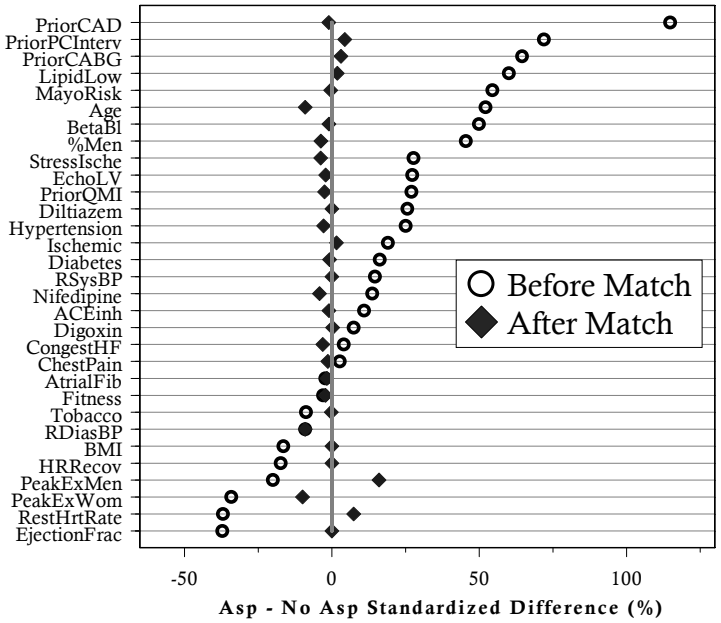
Before Match:

- 811/2310 (35.1%) Aspirin users used β-blockers
- 550/3864 (14.2%) non-Aspirin users used β-blockers
- Standardized Difference is 49.9%
- P value for difference is < .001

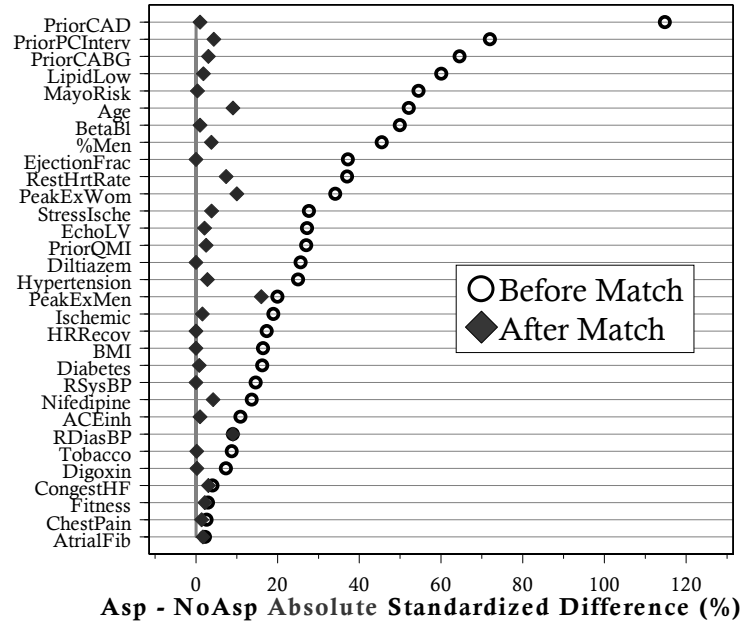
After Match:

- 352/1351 (26.1%) Aspirin users used β-blockers
- 358/1351 (26.5%) non-Aspirin users used β-blockers
- Standardized Difference is –1.0%
- P value for difference is .79

Covariate Balance for Aspirin Study



Absolute Standardized Differences



Week 1 Computing Corner

Stat 266
CHPR 290

```
> data(lalonde) # in MatchIt package, help(lalonde)
> dim(lalonde)  > attach(lalonde)
[1] 614  10
> table(treat)
treat
  0   1
429 185
> head(lalonde)
     treat age educ black hispan married nodegree re74 re75       re78
NSW1     1  37   11     1      0       1        1    0    0  9930.0460
NSW2     1  22    9     0      1       0        1    0    0  3595.8940
NSW3     1  30   12     1      0       0        0    0    0 24909.4500
NSW4     1  27   11     1      0       0        1    0    0  7506.1460
NSW5     1  33    8     1      0       0        1    0    0   289.7899
NSW6     1  22    9     1      0       0        1    0    0  4056.4940

################## prelim compare groups on outcome measure
> tapply(re78, treat, median)
       0        1
4975.505 4232.309
> t.test(re78 ~ treat)
        Welch Two Sample t-test
data:  re78 by treat
t = 0.93773, df = 326.41, p-value = 0.3491
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:   -697.192 1967.244
sample estimates: mean in group 0 mean in group 1
                        6984.170        6349.144

> #####But wait, some say "we are never done until the ancova is run" see Fish
> # as we see the social science, life science practice is to put in the treatment variable  and
> # a whole bunch of other variables to "control" for self-selection, nonequivalence etc.
> #  equivalent to analysis of covariance by whatever name
>  ancova.lalonde = lm( re78 ~ treat + age + educ + black + hispan + married + nodegree + re74 + re75)
>  summary(ancova.lalonde)
Call: lm(formula = re78 ~ treat + age + educ + black + hispan + married +  nodegree + re74 + re75)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.651e+01  2.437e+03   0.027   0.9782
treat        1.548e+03  7.813e+02   1.982   0.0480 *
age          1.298e+01  3.249e+01   0.399   0.6897
educ         4.039e+02  1.589e+02   2.542   0.0113 *
black       -1.241e+03  7.688e+02  -1.614   0.1071
hispan       4.989e+02  9.419e+02   0.530   0.5966
married      4.066e+02  6.955e+02   0.585   0.5590
nodegree     2.598e+02  8.474e+02   0.307   0.7593
re74         2.964e-01  5.827e-02   5.086 4.89e-07 ***
re75         2.315e-01  1.046e-01   2.213   0.0273 *
---
> # so treatment is significantly helpful ??

########### Begin matching analysis; Quintile Subclassification with Propensity Scores
## original Rosenbaum-Rubin, cardiac; Rubin breast cancer

> # now do the logistic regression that computes propensity scores
  # matching packages will do this for you with propen as distance measure
>  glm.p = glm( treat ~ age + educ + black + hispan + married + nodegree + re74 + re75,
                                                  data = lalonde, family = binomial)

>  summary(glm.p)
Call: glm(formula = treat ~ age + educ + black + hispan + married +
    nodegree + re74 + re75, family = binomial, data = lalonde)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.729e+00  1.017e+00  -4.649 3.33e-06 ***
age          1.578e-02  1.358e-02   1.162  0.24521
educ         1.613e-01  6.513e-02   2.477  0.01325 *
black        3.065e+00  2.865e-01  10.699  < 2e-16 ***
```

training (treatment)

outcome

control has higher wages (re78)

First approach, vintage

fit from logistic regression

Week 1 p2
Stat 266

```
hispan       9.836e-01  4.257e-01   2.311  0.02084 *
married     -8.321e-01  2.903e-01  -2.866  0.00415 **
nodegree     7.073e-01  3.377e-01   2.095  0.03620 *
re74        -7.178e-05  2.875e-05  -2.497  0.01253 *
re75         5.345e-05  4.635e-05   1.153  0.24884
---

> propen = fitted(glm.p)  # now we have the propensity scores

> quantile(propen) # overall distrib
         0%          25%          50%          75%         100%
0.009080193 0.048536484 0.120676493 0.638715991 0.853152844
 # look at overlap via 5-number summary (or side-by-side boxplots) not good overlap,
> tapply(propen, treat, quantile)
$`0`
         0%          25%          50%          75%         100%
0.009080193 0.038880745 0.075849106 0.195135746 0.789172834
$`1`
        0%         25%         50%         75%        100%
0.02495179 0.52646352 0.65368426 0.72659995 0.85315284

> # as we are fitting prob(treat = 1) fits for those in treatment group will be larger,
   # we need good overlap for matching purposes
> detach(lalonde) > lalonde$propen = propen  > attach(lalonde)
> boxplot(propen ~ treat) #gives side-by-side boxplots, you can add labels, not wonderful overlap
```

see pictures

```
#### looking at overlap, histograms
>  p1 = propen[treat == 1]  >  p0 = propen[treat == 0]
>  hist(p0,col=rgb(0,0,1,0.7),xlim=range(c(p0,p1)))
>  hist(p1,col=rgb(1,0,0,0.7),add=T)  # superimposed propensity histograms, like Ben Hansen SAT,
   control is blue, treatment is red, overlap close to perfect Stanford Cardinal red
> hist(p0, breaks = "FD", col=rgb(0,0,1,0.7),xlim=range(c(p0,p1)))
> hist(p1, breaks = "FD", col=rgb(1,0,0,0.7),add=T) # Freedman-Diaconis breakpoints

### make quintiles of propensity distribution to to subclassification/strata matching
>  pbin = cut(propen, quantile(propen, seq(0, 1, 1/5)), include.lowest = TRUE, labels = FALSE)
> detach(lalonde) > lalonde$bins = pbin > attach(lalonde)
>  table(pbin, treat)  #each bin of size 122,123
    treat
pbin   0   1
   1 122   1
   2 116   7
   3 101  21
   4  53  71
   5  37  85
```

a pbin classification for each subject

```
##### examples of checking balance (more to come)
> tapply(age, list(bins, treat), median)
   0  1
1 29 27
2 26 23
3 20 23
4 24 25
5 19 25
```

not great see picture

```
> ### install.packages("PSAgraphics") >  library(PSAgraphics)
>  box.psa(age, treat, bins) # see picture
################################### examine outcome re78 by strata
> tapply(re78, list(bins, treat),mean)  # mean diffs in re78 stratified by propensity quintile
          0        1
1 10467.064    0.000
2  5796.548 7919.316
3  6043.316 9210.726
4  4977.401 5819.143
5  4666.221 6030.258
> # direction of mean diffs favors treatment, job training

> # contrast that with the comparison ignoring any concerns about self-selection (selection bias),
```

```
     effect in the other direction, but not significant
>  tapply(re78, treat, mean)
       0        1
6984.170 6349.144

> ##### can do t-tests by subclassification (strata) e.g. for the 3 upper quintiles
> ##### lmer, a better way to do the t-tests  ########################
> library(lme4)
> propen.lmer = lmer(re78 ~ treat + (1 + treat|bins), data = lalonde)
> summary(propen.lmer)
Linear mixed model fit by REML ['lmerMod']
Formula: re78 ~ treat + (1 + treat | bins)   Data: lalonde
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 bins     (Intercept)  5208943 2282
          treat        2069963 1439     -1.00
 Residual             52597981 7252
Number of obs: 614, groups:  bins, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6434.2     1090.2   5.902
treat          385.7      950.8   0.406

# so here we have an overall estimate of the effect of the treat on re78 of positive $386, but
# far from significant. Much smaller point estimate than in some of the individual strata

> confint(propen.lmer) # bombs
> confint(propen.lmer, method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...
                 2.5 %   97.5 %
.sig01        414.81230 4084.578
.sig02         -1.00000    1.000
.sig03         54.74858 3644.981
.sigma       6846.49101 7654.434
(Intercept)  4432.91940 8695.198
treat       -1681.75647 2565.802  some bootstrap runs failed (7/1000)
```

second, another approach

```
############################ Full Matching (Hansen, via Rosenbaum, using MatchIt)

>  m2full.out = matchit(treat ~ re74 + re75 + educ + black + hispan + age + married + nodegree,
                                                                  data = lalonde, method = "full")
>  summary(m2full.out)
Call: matchit(formula = treat ~ re74 + re75 + educ + black + hispan +
    age + married + nodegree, data = lalonde, method = "full")
Summary of balance for all data:
         Means Treated Means Control  Mean Diff   eQQ Med  eQQ Mean   eQQ Max
distance        0.5774        0.1822     0.3952    0.5176     0.3955    0.5966
re74         2095.5737     5619.2365 -3523.6628 2425.5720 3620.9240 9216.5000
re75         1532.0553     2466.4844  -934.4291  981.0968 1060.6582 6795.0100
educ           10.3459       10.2354     0.1105    1.0000     0.7027    4.0000
black           0.8432        0.2028     0.6404    1.0000     0.6432    1.0000
hispan          0.0595        0.1422    -0.0827    0.0000     0.0811    1.0000
age            25.8162       28.0303    -2.2141    1.0000     3.2649   10.0000
married         0.1892        0.5128    -0.3236    0.0000     0.3243    1.0000
nodegree        0.7081        0.5967     0.1114    0.0000     0.1135    1.0000
Summary of balance for matched data:
         Means Treated Means Control Mean Diff  eQQ Med eQQ Mean   eQQ Max
distance        0.5774        0.5761    0.0013   0.0026   0.0066     0.096
re74         2095.5737     2199.7126 -104.1390  72.6510 512.7210 13121.750
re75         1532.0553     1524.8362    7.2191 209.6655 460.5643 12746.050
educ           10.3459       10.3227    0.0233   0.0000   0.4596     4.000
black           0.8432        0.8347    0.0086   0.0000   0.0020     1.000
hispan          0.0595        0.0583    0.0012   0.0000   0.0012     1.000
age            25.8162       24.6928    1.1235   3.0000   3.3100     9.000
married         0.1892        0.1285    0.0607   0.0000   0.0544     1.000
```

```
nodegree        0.7081        0.7040    0.0041   0.0000   0.0028     1.000
Percent Balance Improvement:
          Mean Diff.   eQQ Med eQQ Mean  eQQ Max
distance     99.6662   99.5001  98.3388  83.9052
re74         97.0446   97.0048  85.8401 -42.3724
re75         99.2274   78.6295  56.5775 -87.5796
educ         78.9494  100.0000  34.5954   0.0000
black        98.6582  100.0000  99.6891   0.0000
hispan       98.5858    0.0000  98.5200   0.0000
age          49.2583 -200.0000  -1.3825  10.0000
married      81.2495    0.0000  83.2267   0.0000
nodegree     96.3435    0.0000  97.5333   0.0000

Sample sizes:
          Control Treated   # uses all cases, as do 'inferior' IPTW methods
All           429     185
Matched       429     185
Unmatched       0       0
Discarded       0       0
```

(twang)

alternative optimal 2:1 or 1:1
see RQ 4.1

```
>  summary(m2full.out, standardize = T)
>  plot(summary(m2full.out, standardize = T))  # see picture. 10% criteria
>  plot(m2full.out) > # gives you QQ plots for each var

> detach(lalonde)
>  m2full.dat = match.data(m2full.out)  # obtain results from the full matching
> dim(m2full.dat)
[1] 614  15
> head(m2full.dat) > attach(m2full.dat)
```

get matching data

```
>  # so you can see match.data appends 3 colums "distance" "weights"  "subclass" to the original data s
>  table(m2full.dat$subclass) #the 104 subclasses have various sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26
  2  13   2   7   3   5   3   2   4   2   8   3   2   2   9   4   2   9   6  14   3   2   2   6   3   4
 36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61
 14   5   3   3   2   6   2   5   3   2  10   2   4   8   3   2  14   7   2  14   2   2   4  40   2   2
 71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96
  2   3  70   2   5   6   2   2  13   2   2   2   2   2   7   3   2   2   3   2   2   2   2   3   6   4

###### outcome comparison over the (matched) subclasses # like for the quintiles
> mfull.lmer = lmer(re78 ~ treat + (1 + treat|subclass), data = m2full.dat)
> summary(mfull.lmer)
Linear mixed model fit by REML ['lmerMod']
Formula: re78 ~ treat + (1 + treat | subclass)
   Data: m2full.dat
Number of obs: 614, groups:  subclass, 104

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5862.9      507.8  11.546
treat          504.5      736.2   0.685 ## about the same as seen in base section 384 (952)
```

analog to paired t-test

```
> confint(mfull.lmer)
Computing profile confidence intervals ...
                2.5 %   97.5 %
.sig01      1216.8647 3011.968
.sig02        -1.0000    1.000
.sig03         0.0000      Inf
.sigma      6740.8624 7581.414
(Intercept) 4807.1941 6873.722
treat       -985.7685 1977.973
There were 50 or more warnings (use warnings() to see the first 50)
>
```

a little tighter CI

breastfeeding

By Nadia Kounang

Updated 4:15 AM ET, Mon March 27, 2017

Women breastfeed their babies at the Hirshhorn Museum in Washington in 2011.

Story highlights

Study finds some short-term cognitive benefit to breastfeeding

Differences between breastfed and non-breastfed children lost by age five

(CNN) — While the medical benefits of breastfeeding for helping newborns fight infections and helping pre-term infants get stronger are fairly well established, the long-term impact is much less so.

While new mothers may debate what they believe to be long-term benefits, a new study published in the journal Pediatrics finds that breastfeeding has little impact on long-term cognitive development and behavior.

The study followed 7,478 Irish children born full term, from the time they were 9 months old. They were then evaluated at three years and again at five years of age.

At three, the children's parents were asked to fill out questionnaires evaluating vocabulary and problem-solving skills to assess cognition and behavior. At age five, both parents and teachers were asked the same questions.

While the researchers found that those children who were s of s at three,

neighbor matching, the sample is randomly ordered with matching occurring sequentially between the treatment (breastfed) and control (not breastfed) group based on participants' propensity scores. Typically, the pair is then removed from the list and the next match is created. To ensure optimal matches, we imposed a caliper so that pairs could only be matched if the propensity score was within a tenth of a SD of the other. We also allowed matching with replacement given the low rates of longer durations and full breastfeeding in this cohort. Although matching with replacement has been argued to increase variance in the data, it also arguably reduces bias in the sample by ensuring better quality of matches.[36] Balance checks in all models revealed substantial reductions of bias between matched groups on all individual confounders (ie, 0%–13.9% remaining bias in partial breastfeeding models, 0%–18.1% remaining bias in full models; data available on request). The remaining overall mean bias across models ranged from 3.2% to 8.5%. The ≤20% remaining bias has been suggested as the acceptable cutoff after matching.[37] Thus, we concluded that the analytic matching technique resulted in good matches between conditions. Matching resulted in all participants falling within the area of common support. The average treatment effect on those who were treated (ie, children who were breastfed) is reported. Adjustments were made for multiple hypothesis testing by using the Holmes-Bonferroni method. All statistical analyses for PSM were conducted by using Stata version 13 software (Stata Corp, College Station, TX).

To note, although PSM is advantageous in mimicking random assignment, a drawback is the challenge in evaluating a linear dose-response association, which has previously been found. Structural equation modeling (SEM) offers an alternative approach to examining this dose-response association. Additionally, SEM uses the full sample and has greater power. Thus, the data were also modeled by using SEM, where confounders were treated as correlated exogenous variables, the duration of breastfeeding was treated as a continuous mediating variable, and child outcomes were treated as correlated, which could be influenced by both breastfeeding and confounders. These results can be found in the Supplemental Material.

RESULTS

Postmatching results for children fully breastfed up to 31 days revealed no statistically significant differences between groups on any outcome at age 3 or 5 years (Table 3). Similarly, for children who were fully breastfed between 32 and 180 days, no statistically significant differences were found for any outcomes at either age postmatching (Table 4). Finally, for children who were fully breastfed for ≥6, statistically significant differences were found postmatching for only 2 outcomes, problem solving and hyperactivity at age 3 years. Children who were fully breastfed scored 2.95 (SE = 1.39, $P = .048$) points higher on the problem-solving scale compared with children who were never breastfed and −0.84 (SE = 0.25, $P \leq .001$) points lower on the hyperactivity scale. After adjustment for multiple testing, cognition was no longer statistically significant. However, children who were fully breastfed had slightly lower parent-rated hyperactivity compared with controls, and this remained statistically significant after adjustment (Table 5). Of note, results of the partial breastfeeding models were similar to the full models, however, after adjustment for multiple testing, neither cognitive ability nor hyperactivity at age 3 years remained statistically significant. These results can be found in the Supplemental Material.

DISCUSSION

Without randomized controlled trials, the issue of causality will necessarily remain open, however the present results contribute important insights to the long-standing debate of potential "causal effects" versus artifacts of confounding that are not properly accounted for. This study also provides new perspectives on breastfeeding and children's externalizing behavior. To the best of our knowledge, this is among the first studies to examine expressive vocabulary as an individual outcome and to consider externalizing behavior. It should be noted that our results apply only to infants born full term.

After adjustment for multiple testing, the initial support found for breastfeeding and better problem solving at age 3 years if the child was breastfed for a minimum of 6 months was no longer statistically significant. In addition, no statistically significant effects were found for cognitive ability at age 5 years. These results are in contrast to some studies that have used PSM techniques to examine the effects of breastfeeding and general cognitive abilities.[38–40] However, differences in both analytical choices of the PSM approach used (eg, replacement, calipers) and differing selection of covariates may help to explain these differences across studies. Nonetheless, our findings were surprising in the context of the nutrients in breast milk being responsible for increased cognitive development. Regarding expressive vocabulary, no statistically significant advantages were observed for children who were breastfed at either age 3 or age 5.

The limited research on breastfeeding and behavior problems is inconsistent, despite the relatively consistent reliance on the SDQ. Of interest, studies that have dichotomized the SDQ scales into abnormal scores (ie, at the 85th or 90th percentile) have not found