after installing (Lab4) Week 8

Ben Hansen matching Exs using MatchIt wrapper
Stat 209

```
> library(MatchIt)
Loading required package: MASS
Loading required package: optmatch
Loading required package: Matching
##  MatchIt (Version 2.3-1, built: 2007-10-11)
##  Please refer to http://gking.harvard.edu/matchit for full documentation
> data(nuclear)  # get from "boot"; library(boot)
> nuclear
    cost  date t1 t2  cap pr ne ct bw cum.n pt
1  460.1 68.58 14 46  687  0  1  0  0    14  0
2  453.0 67.33 10 73 1065  0  0  1  0     1  0
3  443.2 67.33 10 85 1065  1  0  1  0     1  0
4  652.3 68.00 11 67 1065  0  1  1  0    12  0
5  642.2 68.00 11 78 1065  1  1  1  0    12  0
.............................................
24 608.8 70.08 19 58  821  1  0  0  0     3  0
29 284.9 67.83 12 63  886  0  0  0  1    11  1
32 270.7 67.83  7 80  886  1  0  0  1    11  1
```

```
> library(boot)
> data(nuclear)
```

insert

orig data set 'nuclearplants'

```
> bennuke = subset(nuclear, subset = pt == 0, select = c(date,cap,pr))
> bennuke$cap = bennuke$cap - 400 > bennuke$date = bennuke$date - 65
> bennuke
   date cap pr                    date cap pr
1  3.58 287  0                 14 3.92 650  0
2  2.33 665  0                 15 3.92 450  0
3  2.33 665  1                 16 3.42 378  0
4  3.00 665  0                 17 4.50 445  0
5  3.00 665  1                 18 3.42 130  1
6  2.92 114  0                 19 4.17 690  0
7  3.17 422  0                 20 3.92 650  1
8  3.42  57  0                 21 3.75 513  0
9  3.42 422  1                 22 5.92 428  1
10 3.33 392  0                 23 4.67 386  0
11 3.58 160  0                 24 5.08 421  1
12 3.75 390  0                 25 5.42 138  0
13 3.42 130  0                 26 6.08 730  0
```

subset and transform to get B/H cx

7 prior plants for matching

2 to 1 match

```
> mopt2= matchit(pr ~ date + cap, data = bennuke, method = "optimal", ratio = 2)
> summary(mopt2)
Percent Balance Improvement:
Sample sizes:          Control Treated
  All                     19       7
  Matched                 14       7
  Unmatched                5       0
  Discarded                0       0
```

diagnostics, plots Lab4

see online for balance

see Gender p.2

lots of options full, manual see Lab4, MatchIt docs p.2 handout

```
> #subclass gives you the matching
> mopt2.data = match.data(mopt2)   > mopt2.data
   date cap pr distance weights subclass
2  2.33 665  0   0.3409       1        1
3  2.33 665  1   0.3409       1        1
4  3.00 665  0   0.3513       1        2
5  3.00 665  1   0.3513       1        2
6  2.92 114  0   0.1514       1        4
7  3.17 422  0   0.2518       1        6
9  3.42 422  1   0.2551       1        3
10 3.33 392  0   0.2427       1        3
12 3.75 390  0   0.2473       1        3
13 3.42 130  0   0.1601       1        4
14 3.92 650  0   0.3590       1        2
15 3.92 450  0   0.2727       1        7
17 4.50 445  0   0.2787       1        6
18 3.42 130  1   0.1601       1        4
19 4.17 690  0   0.3817       1        5
20 3.92 650  1   0.3590       1        5
21 3.75 513  0   0.2960       1        1
22 5.92 428  1   0.2917       1        6
23 4.67 386  0   0.2578       1        7
24 5.08 421  1   0.2770       1        7
26 6.08 730  0   0.4328       1        5
> #this corresponds to hansen slide 45
```

create matching data, then list

subclass marks the matches

default distance measure is propensity score demonstrated in "more nuke"

Check Balance 2:1 match handout

```
>  mopt2= matchit(pr ~ date + cap, data = bennuke, method = "optimal", ratio = 2)
Installing package(s) into 'C:/Users/rag/Documents/R/win-library/2.14'
(as 'lib' is unspecified)
trying URL 'http://cran.cnr.Berkeley.edu/bin/windows/contrib/2.14/optmatch_0.7-1.zip'
Content type 'application/zip' length 330389 bytes (322 Kb)
opened URL
downloaded 322 Kb

package 'optmatch' successfully unpacked and MD5 sums checked

The downloaded packages are in
        C:\Users\rag\AppData\Local\Temp\RtmpWi648r\downloaded_packages
Loading required package: optmatch
> summary(mopt2)

Call:
matchit(formula = pr ~ date + cap, data = bennuke, method = "optimal",
    ratio = 2)

Summary of balance for all data:
propensity      Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean  eQQ Max
distance           0.2907        0.2613     0.0832    0.0294  0.0398   0.0454   0.0852
score date         3.8700        3.8079     0.8807    0.0621  0.1600   0.1529   0.5800
cap             483.0000      403.2632   214.1816   79.7368 65.0000 100.1429 283.0000


Summary of balance for matched data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean  eQQ Max
distance        0.2907        0.2874     0.0799    0.0033   0.019   0.0263   0.0738
date            3.8700        3.7807     0.9177    0.0893   0.090   0.2129   0.9100
cap           483.0000      474.4286   194.3885    8.5714  30.000  42.8571 137.0000

Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean  eQQ Max
distance    88.8018 52.3558  42.1473  13.3177
date       -43.7651 43.7500 -39.2523 -56.8966
cap         89.2504 53.8462  57.2040  51.5901

Sample sizes:
          Control Treated
All            19       7
Matched        14       7
Unmatched       5       0
Discarded       0       0
```

Week 8

p.2

p2 stat209 Ben Hansen, Gender Equity Example

```
> geneq  = read.table(file="D:\\drr08\\stat209\\week8\\genderdata", header = T)
> geneq
  log10Grant gender
1        5.7      W
2        4.0      W
3        3.4      W
4        3.1      W
5        5.5      M
6        5.3      M
7        4.9      M
8        4.9      M
9        3.9      M
```

second Hansen talk ex.
do in HW 8

```
# had some problems matchit; outcome has to be 1=W, 0=M or do a as.numeric
> #outcome has to be numeric categorical, not W/M

> mfullgen= matchit(gender ~ Grant, data = geneq, method = "full")
> summary(mfullgen)
Call:
matchit(formula = gender ~ Grant, data = geneq, method = "full")

Summary of balance for all data:
         Means Treated Means Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance         0.576         0.339     0.237   0.275    0.252   0.421
Grant            4.050         4.900    -0.850   0.850    0.850   1.500

Summary of balance for matched data:
         Means Treated Means Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance         0.576         0.532     0.045   0.129    0.121   0.463
Grant            4.050         4.213    -0.163   0.500    0.507   1.800

Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean eQQ Max
distance      81.18   53.10    51.96  -9.927
Grant         80.88   41.18    40.37 -20.000

Sample sizes:
          Control Treated
All             5       4
Matched         5       4
Unmatched       0       0
Discarded       0       0

> mfullgen.dat = match.data(mfullgen)
> mfullgen.dat
  Grant gender distance weights subclass
1   5.7      1   0.1561  1.0000        1
2   4.0      1   0.5903  1.0000        2
3   3.4      1   0.7484  1.0000        2
4   3.1      1   0.8103  1.0000        2
5   5.5      0   0.1907  0.3125        1
6   5.3      0   0.2307  0.3125        1
7   4.9      0   0.3271  0.3125        1
8   4.9      0   0.3271  0.3125        1
9   3.9      0   0.6192  3.7500        2
> #corresponds to slide, subclass
```

see Hansen talk

An Optimal Full Matching for the Gender Equity Example

Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9

As compared to the optimal 1:(1 or 2) match, full matching:

- decreases the largest discrepancy from 1.5 to 0.8; and
- decreases the sum of discrepancies from 3.8 to 3.6.

In global terms, it gives a tighter match. In local terms, it gives a *much* tighter match.

more nuke matching Week 8 Stat 209

```
R version 2.8.1 (2008-12-22)
> install.packages("MatchIt")
> data(nuclear)
> bennuke = subset(nuclear, subset = pt == 0, select = c(date,cap,pr))
> bennuke$cap = bennuke$cap - 400
> bennuke$date = bennuke$date - 65
> attach(bennuke)
> mopt1= matchit(pr ~ date + cap, data = bennuke, method = "optimal")
> summary(mopt1)
Call:
matchit(formula = pr ~ date + cap, data = bennuke, method = "optimal")
Summary of balance for all data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.291         0.261      0.083     0.029    0.04    0.045   0.085
date              3.870         3.808      0.881     0.062    0.16    0.153   0.580
cap             483.000       403.263    214.182    79.737   65.00  100.143 283.000
Summary of balance for matched data:
          Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance          0.291         0.292      0.070    -0.001       0    0.001   0.004
date              3.870         3.656      0.823     0.214       0    0.309   1.250
cap             483.000       493.429    195.677   -10.429       0   20.429  85.000
Percent Balance Improvement:
          Mean Diff. eQQ Med eQQ Mean eQQ Max
distance       95.81     100    97.29   94.91
date         -245.04     100  -101.87 -115.52
cap            86.92     100    79.60   69.97
Sample sizes:
           Control Treated
All             19       7
Matched          7       7
Unmatched       12       0
Discarded        0       0
```

cf Lab 4

see [redacted] handout

optimal match 1:1

Before and After matching

← only asked for 1:1 ratio=1

maybe get better balance, maybe not compared to ratio=2

```
> mopt.data = match.data(mopt1)
> mopt.data
   date cap pr distance weights subclass
2  2.33 665  0   0.3409       1        1
3  2.33 665  1   0.3409       1        1
4  3.00 665  0   0.3513       1        2
5  3.00 665  1   0.3513       1        2
9  3.42 422  1   0.2551       1        3
13 3.42 130  0   0.1601       1        4
14 3.92 650  0   0.3590       1        5
17 4.50 445  0   0.2787       1        7
18 3.42 130  1   0.1601       1        4
20 3.92 650  1   0.3590       1        5
21 3.75 513  0   0.2960       1        6
22 5.92 428  1   0.2917       1        6
23 4.67 386  0   0.2578       1        3
24 5.08 421  1   0.2770       1        7
```

create a matched data set "subclass" tells us the matches

ID

default distance is propensity score

```
> glmnuke = glm(pr ~ date + cap, family = binomial, data = bennuke)
> propen = fitted(glmnuke)
> propen
     1      2      3      4      5      6      7      8      9     10     11     12     13
0.2089 0.3409 0.3409 0.3513 0.3513 0.1514 0.2518 0.1414 0.2551 0.2427 0.1699 0.2473 0.1601
    14     15     16     17     18     19     20     21     22     23     24     25     26
0.3590 0.2727 0.2387 0.2787 0.1601 0.3817 0.3590 0.2960 0.2917 0.2578 0.2770 0.1819 0.4328
```

create the propensity score by hand

⊗ in mopt.data

```
> attach(mopt.data)
> names(mopt.data)
[1] "date"     "cap"      "pr"       "distance" "weights"  "subclass"
> tapply(distance, list(pr,subclass), mean)
       1      2      3      4      5      6      7
0 0.3409 0.3513 0.2578 0.1601 0.3590 0.2960 0.2787
1 0.3409 0.3513 0.2551 0.1601 0.3590 0.2917 0.2770
> tapply(cap, list(pr,subclass), mean)
   1   2   3   4   5   6   7
0 665 665 386 130 650 513 445
1 665 665 422 130 650 428 421
> tapply(date, list(pr,subclass), mean)
     1 2    3    4    5    6    7
0 2.33 3 4.67 3.42 3.92 3.75 4.50
1 2.33 3 3.42 3.42 3.92 5.92 5.08
```

you can do your own comparisons of balance after matching

cf have Aspirin ex 10% effect size rule

```
> mnear1= matchit(pr ~ date + cap, data = bennuke, method = "nearest")
> summary(mnear1)
```

compare with nearest neighbor 1:1 matching

```
Call:
matchit(formula = pr ~ date + cap, data = bennuke, method = "nearest")

Summary of balance for all data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance         0.291         0.261      0.083     0.029    0.04    0.045   0.085
date             3.870         3.808      0.881     0.062    0.16    0.153   0.580
cap            483.000       403.263    214.182    79.737   65.00  100.143 283.000


Summary of balance for matched data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance         0.291         0.292      0.070    -0.001       0    0.001   0.004
date             3.870         3.656      0.823     0.214       0    0.309   1.250
cap            483.000       493.429    195.677   -10.429       0   20.429  85.000

Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean eQQ Max
distance      95.81     100    97.29   94.91
date        -245.04     100  -101.87 -115.52
cap           86.92     100    79.60   69.97

Sample sizes:
           Control Treated
All             19       7
Matched          7       7
Unmatched       12       0
Discarded        0       0

> mnear1.data = match.data(mnear1)
> mnear1.data
   date cap pr distance weights
2  2.33 665  0   0.3409       1
3  2.33 665  1   0.3409       1
4  3.00 665  0   0.3513       1
5  3.00 665  1   0.3513       1
9  3.42 422  1   0.2551       1
13 3.42 130  0   0.1601       1
14 3.92 650  0   0.3590       1
17 4.50 445  0   0.2787       1
18 3.42 130  1   0.1601       1
20 3.92 650  1   0.3590       1
21 3.75 513  0   0.2960       1
22 5.92 428  1   0.2917       1
23 4.67 386  0   0.2578       1
24 5.08 421  1   0.2770       1

>
```

matched data show same result as optimal for this small example.

Rubin An Int Med Cochrans ex, matching on Age

Table 1. Comparison of Mortality Rates for Three Smoking Groups in Three Databases*

Variable	Canadian Study			United Kingdom Study			United States Study		
	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers
Mortality rates per 1000 person-years, %	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
Average age, *y*	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
Adjusted mortality rates using subclasses, %									
2 subclasses	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3 subclasses	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9–11 subclasses	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

* Adapted from Tables 1–3 in Cochran (2).

subclassification p757 Rubin for bias reduction

```
>  mfull= matchit(pr ~ date + cap, data = bennuke, method = "full")
> summary(mfull)
Call: matchit(formula = pr ~ date + cap, data = bennuke, method = "full")
Summary of balance for all data:
         Means Treated Means Control Mean Diff eQQ Med eQQ Mean  eQQ Max
distance        0.2907        0.2613    0.0294  0.0398   0.0454   0.0852
date            3.8700        3.8079    0.0621  0.1600   0.1529   0.5800
cap           483.0000      403.2632   79.7368 65.0000 100.1429 283.0000
Summary of balance for matched data:
         Means Treated Means Control Mean Diff eQQ Med eQQ Mean  eQQ Max
distance        0.2907        0.2939   -0.0032  0.0043   0.0113   0.0738
date            3.8700        3.6312    0.2388  0.1600   0.3107   1.4200
cap           483.0000      496.5905  -13.5905 23.0000  35.0232 261.0000
Percent Balance Improvement:
         Mean Diff. eQQ Med  eQQ Mean   eQQ Max
distance    89.2703 89.1190   75.2162   13.3177
date      -284.4471  0.0000 -103.2460 -144.8276
cap         82.9558 64.6154   65.0268    7.7739
Sample sizes:
          Control Treated
All            19       7
Matched        19       7
Unmatched       0       0
Discarded       0       0

> mfull.data = match.data(mfull)        > mfull.data
   date cap pr  distance   weights subclass
1  3.58 287  0 0.2089340 0.4523810        3
2  2.33 665  0 0.3408701 2.7142857        1
3  2.33 665  1 0.3408701 1.0000000        1
4  3.00 665  0 0.3513216 2.7142857        2
5  3.00 665  1 0.3513216 1.0000000        2
6  2.92 114  0 0.1513795 0.5428571        4
7  3.17 422  0 0.2518308 0.4523810        3
8  3.42  57  0 0.1414043 0.5428571        4
9  3.42 422  1 0.2550915 1.0000000        3
10 3.33 392  0 0.2426859 0.4523810        3
11 3.58 160  0 0.1699374 0.5428571        4
12 3.75 390  0 0.2472990 0.4523810        3
13 3.42 130  0 0.1601291 0.5428571        4
14 3.92 650  0 0.3589552 0.9047619        5
15 3.92 450  0 0.2726900 1.3571429        7
16 3.42 378  0 0.2386883 0.4523810        3
17 4.50 445  0 0.2786707 1.3571429        7
18 3.42 130  1 0.1601291 1.0000000        4
19 4.17 690  0 0.3816748 0.9047619        5
20 3.92 650  1 0.3589552 1.0000000        5
21 3.75 513  0 0.2960073 2.7142857        6
22 5.92 428  1 0.2916721 1.0000000        6
23 4.67 386  0 0.2577539 0.4523810        3
24 5.08 421  1 0.2770347 1.0000000        7
25 5.42 138  0 0.1819249 0.5428571        4
26 6.08 730  0 0.4327688 0.9047619        5
> attach(mfull.data)
> table(subclass)
subclass
1 2 3 4 5 6 7
2 2 7 6 4 2 3
> tapply(distance, list(pr,subclass), mean)
          1         2         3         4         5         6         7
0 0.3408701 0.3513216 0.2411986 0.1609550 0.3911329 0.2960073 0.2756804
1 0.3408701 0.3513216 0.2550915 0.1601291 0.3589552 0.2916721 0.2770347
> tapply(cap, list(pr,subclass), mean)         > tapply(date, list(pr,subclass), mean)
    1   2         3     4   5   6     7              1 2         3     4         5    6    7
0 665 665 375.8333 119.8 690 513 447.5         0 2.33 3 3.653333 3.752 4.723333 3.75 4.21
1 665 665 422.0000 130.0 650 428 421.0         1 2.33 3 3.420000 3.420 3.920000 5.92 5.08
```