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 ~~for~~ 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 01

```
>  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)

> 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)
>
```

analog to paired t-test

a little tighter CI