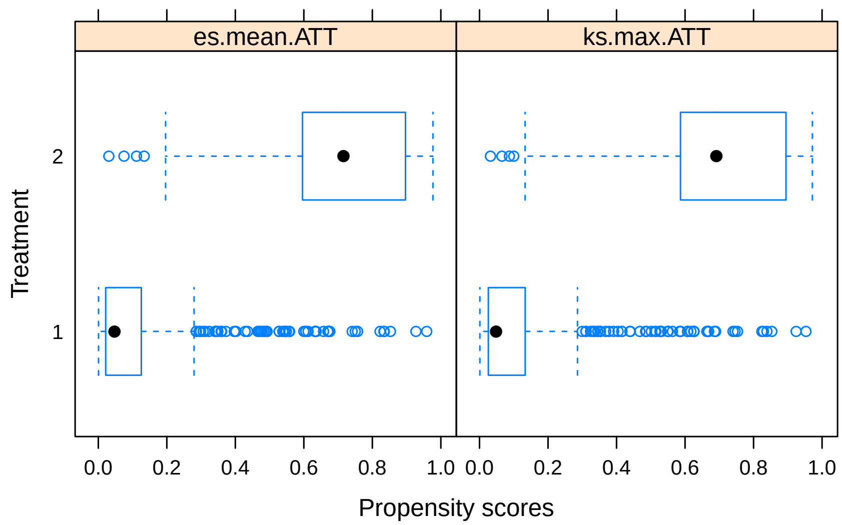
```
    ##### Week5 Computing Corner: twang package, IPTW with boosteg regression for propensity
 R version 3.2.2 (2015-08-14) -- "Fire Safety"

 > install.packages("twang")
 > library(twang)

 > set.seed(1) # for tutorial match
 > data(lalonde) # like week1 ComCo

 > # Details IPTW for later
 > # Weights for ATT are 1 for the treatment cases and p/(1-p) for the control cases.
 > # Weights for ATE are 1/p for the treatment cases and 1/(1-p) for the control cases

 ### propensity score from boosted regression (calls gbm from week 4)
  ## tuning params etc from tutorial, ATT is goal here
 > ps.lalonde <- ps(treat ~ age + educ + black + hispan + nodegree + + married + re74 + re75, data = lal
 +      n.trees=5000, interaction.depth=2, shrinkage=0.01, perm.test.iters=0,
 +   stop.method=c("es.mean","ks.max"), estimand = "ATT", verbose=FALSE)

 # plots page 10 tutorial
 > plot(ps.lalonde, plots = 2) # propen boxplots (es and ks from stop.method) poor overlap
 > plot(ps.lalonde, plots = 3) # standardized imbalance plot like from MatchIt
 > lalonde.balance = bal.table(ps.lalonde) # like MatchIt tables
 > lalonde.balance
 $unw
             tx.mn    tx.sd    ct.mn    ct.sd std.eff.sz   stat     p    ks ks.pval
 age        25.816    7.155   28.030   10.787     -0.309 -2.994 0.003 0.158   0.003
 educ       10.346    2.011   10.235    2.855      0.055  0.547 0.584 0.111   0.074
 black       0.843    0.365    0.203    0.403      1.757 19.371 0.000 0.640   0.000
 hispan      0.059    0.237    0.142    0.350     -0.349 -3.413 0.001 0.083   0.317
 nodegree    0.708    0.456    0.597    0.491      0.244  2.716 0.007 0.111   0.074
 married     0.189    0.393    0.513    0.500     -0.824 -8.607 0.000 0.324   0.000
 re74     2095.574 4886.620 5619.237 6788.751     -0.721 -7.254 0.000 0.447   0.000
 re75     1532.055 3219.251 2466.484 3291.996     -0.290 -3.282 0.001 0.288   0.000

 $es.mean.ATT
             tx.mn    tx.sd    ct.mn    ct.sd std.eff.sz   stat     p    ks ks.pval
 age        25.816    7.155   25.802    7.279      0.002  0.015 0.988 0.122   0.892
 educ       10.346    2.011   10.573    2.089     -0.113 -0.706 0.480 0.099   0.977
 black       0.843    0.365    0.842    0.365      0.003  0.027 0.978 0.001   1.000
 hispan      0.059    0.237    0.042    0.202      0.072  0.804 0.421 0.017   1.000
 nodegree    0.708    0.456    0.609    0.489      0.218  0.967 0.334 0.099   0.977
 married     0.189    0.393    0.189    0.392      0.002  0.012 0.990 0.001   1.000
 re74     2095.574 4886.620 1556.930 3801.566      0.110  1.027 0.305 0.066   1.000
 re75     1532.055 3219.251 1211.575 2647.615      0.100  0.833 0.405 0.103   0.969

 $ks.max.ATT
             tx.mn    tx.sd    ct.mn    ct.sd std.eff.sz   stat     p    ks ks.pval
 age        25.816    7.155   25.764    7.408      0.007  0.055 0.956 0.107   0.919
 educ       10.346    2.011   10.572    2.140     -0.113 -0.712 0.477 0.107   0.919
 black       0.843    0.365    0.835    0.371      0.022  0.187 0.852 0.008   1.000
 hispan      0.059    0.237    0.043    0.203      0.069  0.779 0.436 0.016   1.000
 nodegree    0.708    0.456    0.601    0.490      0.235  1.100 0.272 0.107   0.919
 married     0.189    0.393    0.199    0.400     -0.024 -0.169 0.866 0.010   1.000
 re74     2095.574 4886.620 1673.666 3944.600      0.086  0.800 0.424 0.054   1.000
 re75     1532.055 3219.251 1257.242 2674.922      0.085  0.722 0.471 0.094   0.971

 > summary(ps.lalonde)  # note ess (effective sample size) !!
             n.treat n.ctrl ess.treat  ess.ctrl    max.es    mean.es    max.ks max.ks.p    mean.ks iter
 unw             185    429       185 429.00000 1.7567745 0.56872589 0.6404460       NA 0.27024507   NA
 es.mean.ATT     185    429       185  22.96430 0.2177817 0.07746175 0.1223384       NA 0.06361021 2127
 ks.max.ATT      185    429       185  27.05472 0.2348846 0.08025994 0.1070761       NA 0.06282432 1756

 # boosted regression look # bar chart for free
 > summary(ps.lalonde$gbm.obj)
```

of the estimated propensity scores in the treatment and comparison groups. Whereas propensity score stratification requires considerable overlap in these spreads, excellent covariate balance can often be achieved with weights, even when the propensity scores estimated for the treatment and control groups show little overlap.

```
> plot(ps.lalonde, plots=2)
```

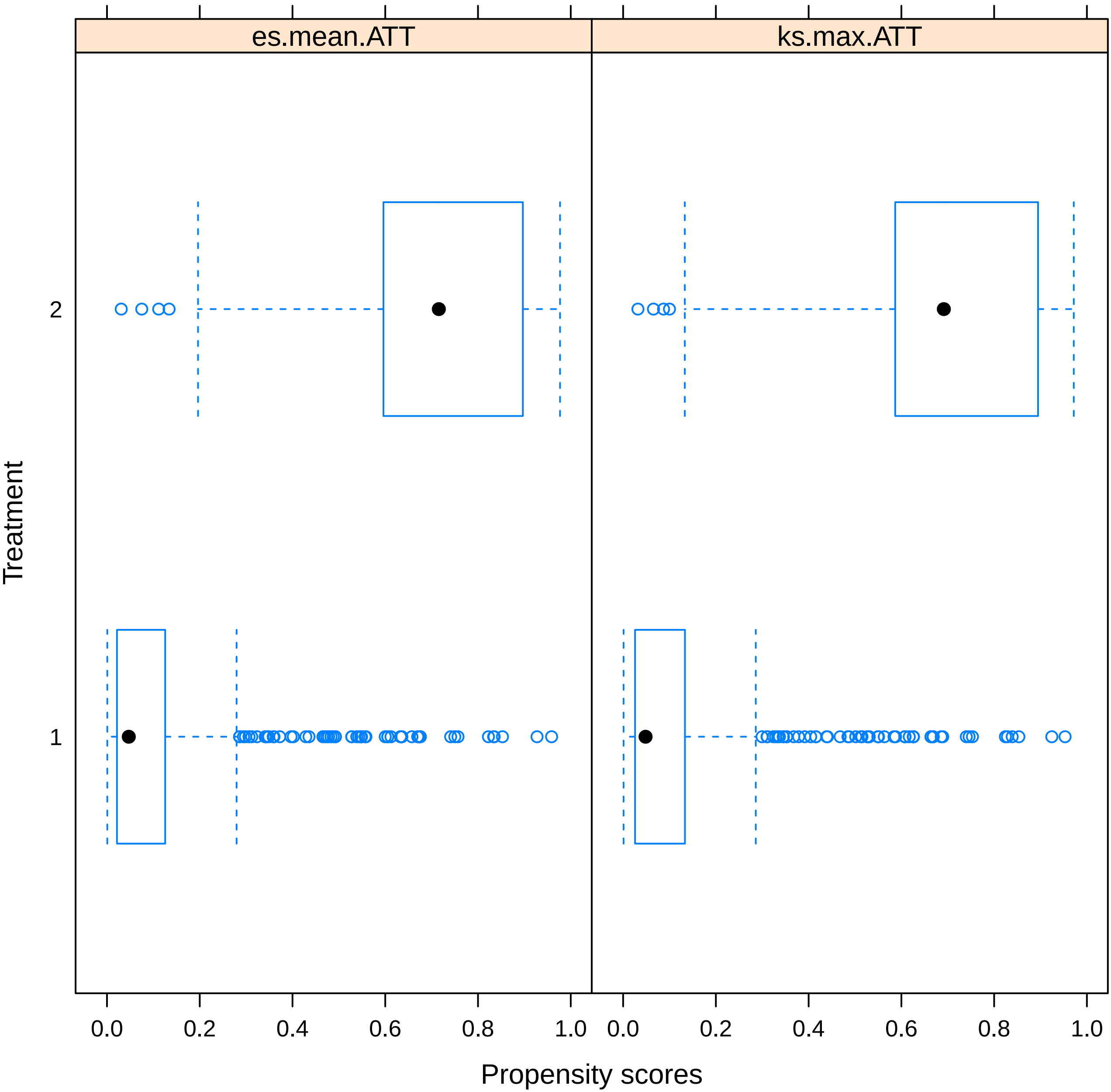


Descriptive argument	Numeric argument	Description
`"optimize"`	1	Balance measure as a function of GBM iterations
`"boxplot"`	2	Boxplot of treatment/control propensity scores
`"es"`	3	Standardized effect size of pretreatment variables
`"t"`	4	t-test p-values for weighted pretreatment variables
`"ks"`	5	Kolmogorov-Smirnov p-values for weighted pretreatment variables
`"histogram"`	6	Histogram of weights for treatment/control

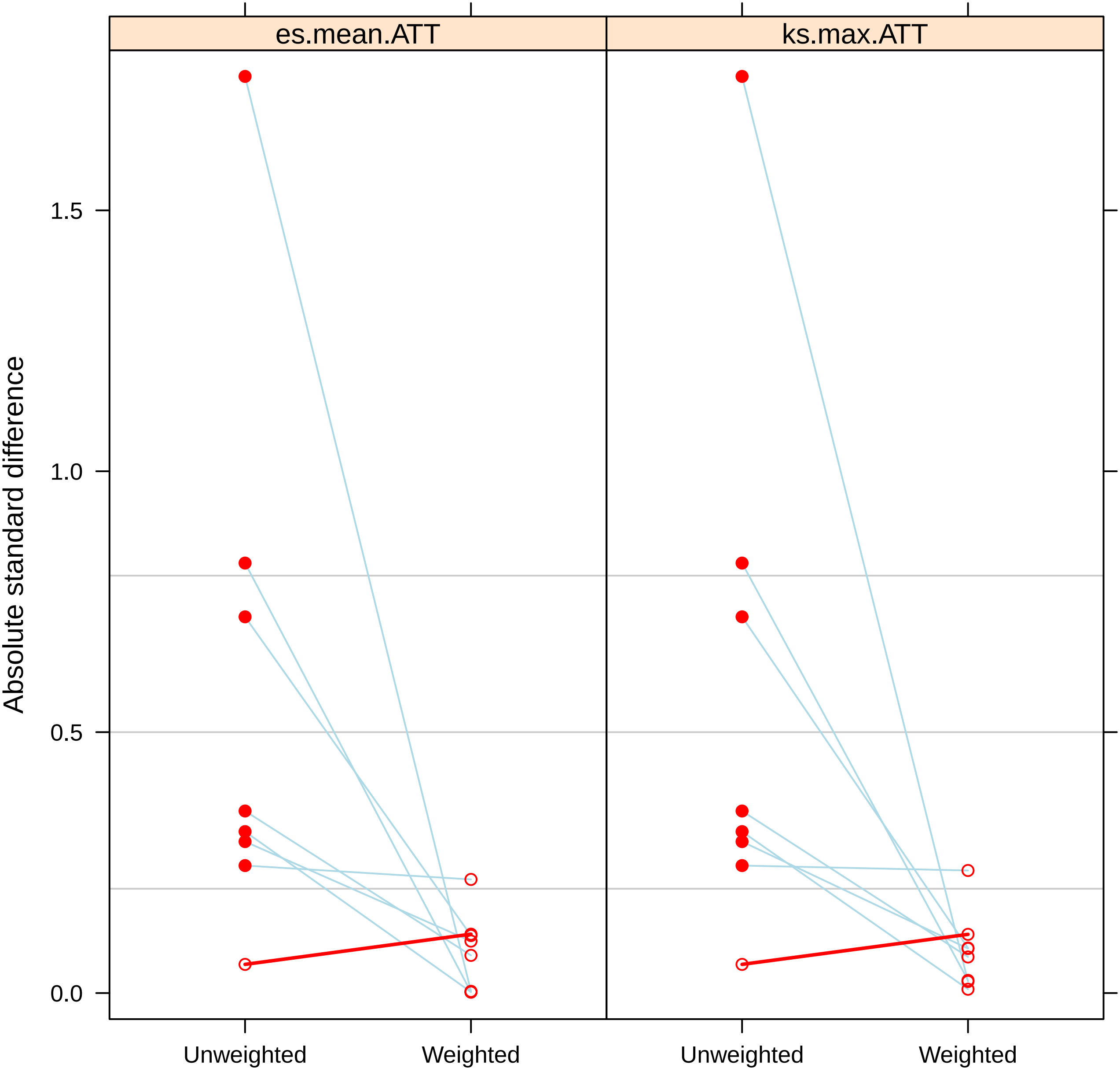
Table 2: Available options for `plots` argument to `plot()` function.

The effect size plot illustrates the effect of weights on the magnitude of differences between groups on each pretreatment covariate. These magnitudes are standardized using the standardized effect size described earlier. In these plots, substantial reductions in effect sizes are observed for most variables (blue lines), with only one variable showing an increase in effect size (red lines), but only a seemingly trivial increase. Closed red circles indicate a statistically significant difference, many of which occur before weighting, none after. In some analyses variables can have very little variance in the treatment group sample or the entire sample and group differences can be very large relative to the standard deviations. In these situations, the user is warned that some effect sizes are too large to plot.

```
> plot(ps.lalonde, plots=3)
```



es.mean.ATT
ks.max.ATT
Treatment
2
1
0.0
0.2
0.4
0.6
0.8
1.0
Propensity scores



es.mean.ATT
ks.max.ATT
Absolute standard difference
0.0
0.5
1.0
1.5
Unweighted
Weighted
Unweighted
Weighted

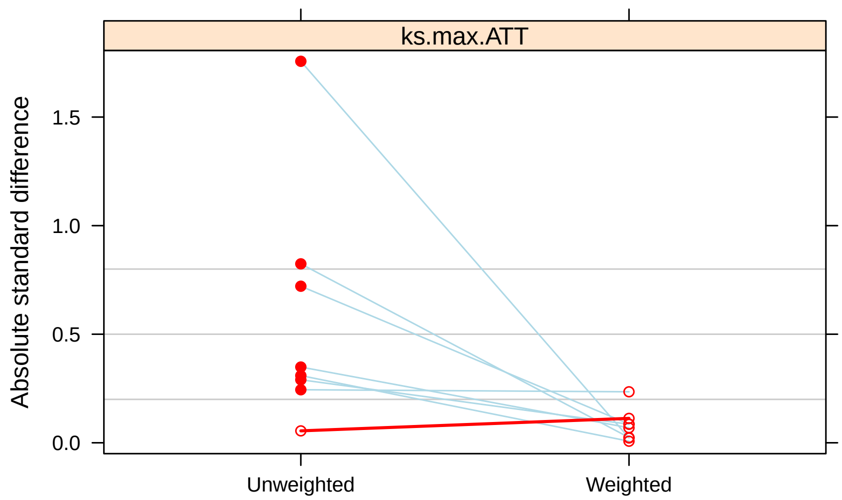
```
              var    rel.inf
black       black 52.5951837
re74         re74 17.4828827
age           age 16.8346839
re75         re75  6.3199135
educ         educ  3.4137190
married   married  2.8185068
nodegree nodegree  0.4291914
hispan     hispan  0.1059190
```

```
> attach(lalonde)
> cor(treat,lalonde) # black has highest cor with treatment
     treat        age       educ     black     hispan    married  nodegree      re74       re75
[1,]     1 -0.1028929 0.01930817 0.6009066 -0.1179833 -0.3013337 0.1058572 -0.249779 -0.1301972 -0.0390
> detach(lalonde)

> propen1 = ps.lalonde$ps # extract propensity scores--note this is a data frame: both es and ks criter
> str(propen1)
'data.frame':   614 obs. of  2 variables:
 $ es.mean.ATT: num  0.595 0.738 0.927 0.959 0.953 ...
 $ ks.max.ATT : num  0.615 0.692 0.924 0.953 0.948 ...
> fivenum(propen1$es.mean.ATT)
[1] 0.0006532284 0.0329630726 0.1080491150 0.6055660930 0.9768987928
> boxplot(propen1$es.mean.ATT) # replicates twang plot
# add treatment and outcome to my little data frame
> propen1$treat = lalonde$treat
> propen1$re78 = lalonde$re78
> head(propen1)
  es.mean.ATT ks.max.ATT treat       re78
1   0.5945568  0.6151896     1  9930.0460
2   0.7382721  0.6917407     1  3595.8940
3   0.9272562  0.9235889     1 24909.4500
4   0.9587267  0.9529411     1  7506.1460
5   0.9534908  0.9484507     1   289.7899
6   0.9591846  0.9529411     1  4056.4940
> boxplot(propen1$re78 ~ propen1$treat)
> boxplot(propen1$es.mean.ATT ~ propen1$treat)

> # do by hand the ATT estimation (see cc_3 session)
> propen1$weight.ATT =
    ifelse(propen1$treat ==1, 1, propen1$es.mean.ATT/(1 - propen1$es.mean.ATT))

> lm.ATT = lm(propen1$re78 ~ propen1$treat, data = propen1, weights = (propen1$weight.ATT))
> summary(lm.ATT)
Call: lm(formula = propen1$re78 ~ propen1$treat, data = propen1, weights = (propen1$weight.ATT))

Weighted Residuals:
   Min     1Q Median     3Q    Max
-20052  -1947   -284   1478  53959

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     5616.6      430.4  13.051   <2e-16 ***
propen1$treat    732.5      574.4   1.275    0.203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5175 on 612 degrees of freedom
Multiple R-squared:  0.00265,   Adjusted R-squared:  0.001021
F-statistic: 1.626 on 1 and 612 DF,  p-value: 0.2027

> # point estimate matches tutorial which uses weighted regression from survey package
> # is the standard IPTW method too optimistic? survey gives se of 1057!
# smoking paper, itn_4 used bootstrap of ATE regression to get s.e.

> confint(lm.ATT)
```

```
                  2.5 %   97.5 %
(Intercept)   4771.4763 6461.778
propen1$treat -395.5411 1860.574
> plot(lm.ATT) # standard diagnostics
```

```
> # tutorial sec2.5 accomodates logistic propen; dx.wts, bal.table give nice covariate balance statisti
> # tutorial sec 2.4 repeats week1 ComCo-- not done till ancova is run
> # tutorial section 3 does Lindner, with ATE on lifepres
```

```
> plot(ps.lalonde, plots = 3, subset = 2)
```



2.3 Analysis of outcomes

A separate R package, the `survey` package, is useful for performing the outcomes analyses using weights. Its statistical methods account for the weights when computing standard error estimates. It is not a part of the standard R installation but installing `twang` should automatically install `survey` as well.

```
> library(survey)
```

The `get.weights()` function extracts the propensity score weights from a `ps` object. Those weights may then be used as case weights in a `svydesign` object. By default, it returns weights corresponding to the estimand (ATE or ATT) that was specified in the original call to `ps()`. If needed, the user can override the default via the optional `estimand` argument.

```
> lalonde$w <- get.weights(ps.lalonde, stop.method="es.mean")
> design.ps <- svydesign(ids=~1, weights=~w, data=lalonde)
```

The `stop.method` argument specifies which GBM model, and consequently which weights, to utilize.

The `svydesign` function from the `survey` package creates an object that stores the dataset along with design information needed for analyses. See `help(svydesign)` for more details on setting up `svydesign` objects.

The aim of the National Supported Work Demonstration analysis is to determine whether the program was effective at increasing earnings in 1978. The propensity score adjusted test can be computed with `svyglm`.

```
> glm1 <- svyglm(re78 ~ treat, design=design.ps)
> summary(glm1)
```

```
Call:
svyglm(formula = re78 ~ treat, design = design.ps)

Survey design:
svydesign(ids = ~1, weights = ~w, data = lalonde)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5616.6      884.9   6.347 4.28e-10 ***
treat          732.5     1056.6   0.693    0.488
---
Signif. codes:
0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for gaussian family taken to be 49804197)

Number of Fisher Scoring iterations: 2
```

The analysis estimates an increase in earnings of $733 for those that participated in the NSW compared with similarly situated people observed in the CPS. The effect, however, does not appear to be statistically significant.

Some authors have recommended utilizing both propensity score adjustment and additional covariate adjustment to minimize mean square error or to obtain "doubly robust" estimates of the treatment effect (Huppler-Hullsiek & Louis 2002, Bang & Robins 2005). These estimators are consistent if either the propensity scores are estimated correctly *or* the regression model is specified correctly. For example, note that the balance table for `ks.max.ATT` made the two groups more similar on `nodegree`, but still some differences remained, 70.8% of the treatment group had no degree while 60.1% of the comparison group had no degree. While linear regression is sensitive to model misspecification when the treatment and comparison groups are dissimilar, the propensity score weighting has made them more similar, perhaps enough so that additional modeling with covariates can adjust for any remaining differences. In addition to potential bias reduction, the inclusion of additional covariates can reduce the standard error of the treatment effect if some of the covariates are strongly related to the outcome.

```
> glm2 <- svyglm(re78 ~ treat + nodegree, design=design.ps)
> summary(glm2)

Call:
svyglm(formula = re78 ~ treat + nodegree, design = design.ps)

Survey design:
svydesign(ids = ~1, weights = ~w, data = lalonde)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6768.4     1471.0   4.601 5.11e-06 ***
treat          920.3     1082.8   0.850    0.396
nodegree     -1891.8     1261.9  -1.499    0.134
---
Signif. codes:
0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

Week 1 p3
Stat 266

```
    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
```

```
(Dispersion parameter for gaussian family taken to be 49013778)

Number of Fisher Scoring iterations: 2
```

Adjusting for the remaining group difference in the `nodegree` variable slightly increased the estimate of the program's effect to $920, but the difference is still not statistically significant. We can further adjust for the other covariates, but that too in this case has little effect on the estimated program effect.

```
> glm3 <- svyglm(re78 ~ treat + age + educ + black + hispan + nodegree +
+                        married + re74 + re75,
+                 design=design.ps)
> summary(glm3)

Call:
svyglm(formula = re78 ~ treat + age + educ + black + hispan +
    nodegree + married + re74 + re75, design = design.ps)

Survey design:
svydesign(ids = ~1, weights = ~w, data = lalonde)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.459e+03  4.289e+03  -0.573  0.56671
treat        7.585e+02  1.019e+03   0.745  0.45674
age          3.005e+00  5.558e+01   0.054  0.95691
educ         7.488e+02  2.596e+02   2.884  0.00406 **
black       -7.627e+02  1.012e+03  -0.753  0.45153
hispan       6.106e+02  1.711e+03   0.357  0.72123
nodegree     5.350e+02  1.626e+03   0.329  0.74227
married      4.918e+02  1.072e+03   0.459  0.64660
re74         5.699e-02  1.801e-01   0.316  0.75176
re75         1.568e-01  1.946e-01   0.806  0.42076
---
Signif. codes:
0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for gaussian family taken to be 47150852)

Number of Fisher Scoring iterations: 2
```

2.4 Estimating the program effect using linear regression

The more traditional regression approach to estimating the program effect would fit a linear model with a treatment indicator and linear terms for each of the covariates.

```
> glm4 <- lm(re78 ~ treat + age + educ + black + hispan + nodegree +
+                    married + re74 + re75,
+             data=lalonde)
> summary(glm4)
```

```
Call:
lm(formula = re78 ~ treat + age + educ + black + hispan + nodegree +
    married + re74 + re75, data = lalonde)

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
nodegree     2.598e+02  8.474e+02   0.307   0.7593
married      4.066e+02  6.955e+02   0.585   0.5590
re74         2.964e-01  5.827e-02   5.086 4.89e-07 ***
re75         2.315e-01  1.046e-01   2.213   0.0273 *
---
Signif. codes:
0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 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
```

This model estimates a rather strong treatment effect, estimating a program effect of \$1548 with a p-value=0.048. Several variations of this regression approach also estimate strong program effects. For example using square root transforms on the earnings variables yields a p-value=0.016. These estimates, however, are very sensitive to the model structure since the treatment and control subjects differ greatly as seen in the unweighted balance comparison (\$unw) from `bal.table(ps.lalonde)`.

2.5 Propensity scores estimated from logistic regression

Propensity score analysis is intended to avoid problems associated with the misspecification of covariate adjusted models of outcomes, but the quality of the balance and the treatment effect estimates can be sensitive to the method used to estimate the propensity scores. Consider estimating the propensity scores using logistic regression instead of `ps()`.

```
> ps.logit <- glm(treat ~ age + educ + black + hispan + nodegree +
+                          married + re74 + re75,
+                  data = lalonde,
+                  family = binomial)
> lalonde$w.logit <- rep(1,nrow(lalonde))
> lalonde$w.logit[lalonde$treat==0] <- exp(predict(ps.logit,subset(lalonde,treat==0)))
```

`predict()` for logistic regression model produces estimates on the log-odds scale by default. Exponentiating those predictions for the comparison subjects gives the ATT weights $p/(1-p)$.

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
```

training (treatment)

outcome

```
################## 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
```

control has
higher wages (re78)

```
> #####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 ??
```

First approach, vintage

```
########### 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 ***
```

fit from
logistic
regression

```
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 w1

```
>  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
```

analog to paired t-test

a little tighter CI

```
There were 50 or more warnings (use warnings() to see the first 50)
>
```

The analysis estimates an increase in earnings of $1214 for those that participated in the NSW compared with similarly situated people observed in the CPS. Table 5 compares all of the treatment effect estimates.

Treatment effect	PS estimate	Linear adjustment
$733	GBM, minimize KS	none
$920	GBM, minimize KS	nodegree
$758	GBM, minimize KS	all
$1548	None	all
$1214	Logistic regression	none
$1237	Logistic regression	all

Table 5: Treatment effect estimates by various methods

3 An ATE example

In the analysis of Section 2, we focused on estimating ATT for the `lalonde` dataset. In this situation, the ATE is not of great substantive interest because not all people who are offered entrance into the program could be expected to take advantage of the opportunity. Further, there is some evidence that the treated subjects were drawn from a subset of the covariate space. In particular, in an ATE analysis, we see that we are unable to achieve balance, especially for the "black" indicator.

We now turn to an ATE analysis that is feasible and meaningful. We focus on the `lindner` dataset, which was included in the `USPS` package (Obenchain 2011), and is now included in `twang` for convenience. A tutorial by Helmreich and Pruzek (2009; HP) for the `PSAgraphics` package also uses propensity scores to analyze a portion of these data. HP describe the data as follows on p. 3 with our minor recodings in square braces:

> The `lindner` data contain data on 996 patients treated at the Lindner Center, Christ Hospital, Cincinnati in 1997. Patients received a Percutaneous Coronary Intervention (PCI). The data consists of 10 variables. Two are outcomes: [`sixMonthSurvive`] ranges over two values... depending on whether patients surved to six months post treatment [denoted by `TRUE`] or did not survive to six months [`FALSE`]... Secondly, `cardbill` contains the costs in 1998 dollars for the first six months (or less if the patient did not survive) after treatment... The treatment variable is `abcix`, where 0 indicates PCI treatment and 1 indicates standard PCI treatment and additional treatment in some form with abciximab. Covariates include `acutemi`, 1 indicating a recent acute myocardial infarction and 0 not; `ejecfrac` for the left ventricle ejection fraction, a percentage from 0 to 90; `ves1proc` giving the number of vessels (0 to 5) involved in the initial PCI; `stent` with 1 indicating coronary stent inserted, 0 not; `diabetic` where 1 indicates that the patient has been diagnosed with diabetes, 0 not; `height` in centimeters and `female` coding the sex of the patent, 1 for female, 0 for male.

HP focus on `cardbill` — the cost for the first months after treatment — as their outcome of interest. However, since not all patients survived to six months, it is not clear whether a lower value of `cardbill` is good or not. For this reason, we choose six-month survival (`sixMonthSurvive`) as our outcome of interest.

Ignoring pre-treatment variables, we see that `abcix` is associated with lower rates of 6-month mortality:

3.4 Sensitivity Analysis: People Who Look Comparable May Differ

What is sensitivity analysis?

If the naïve model (3.5)–(3.8) were true, the distribution of treatment assignments $\mathbf{Z}$ in a randomized paired experiment could be reconstructed by matching for the observed covariate, $\mathbf{x}$. It is common for a critic to argue that, in a particular study, the naïve model may be false. Indeed, it may be false. Typically, the critic accepts that the investigators matched for the observed covariates, $\mathbf{x}$, so treated and control subjects are seen to be comparable in terms of $\mathbf{x}$, but the critic points out that the investigators did not measure a specific covariate u, did not match for u, and so are in no position to assert that treated and control groups are comparable in terms of u. This criticism could be dismissed in a randomized experiment — randomization does tend to balance unobserved covariates — but the criticism cannot be dismissed in an observational study. This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: "This would never happen in my laboratory."

It is important to recognize at the outset that our critic may be, but need not be, on the side of the angels. The tobacco industry and its (sometimes distinguished) consultants criticized, in precisely this way, observational studies linking smoking with lung cancer [103]. In this instance, the criticism was wrong. Investigators and their critics stand on level ground [8].

It is difficult if not impossible to give form to arguments of this sort until one has a way of speaking about the degree to which the naïve model is false. In an observational study, one could never assert with warranted conviction that the naïve model is precisely true. Trivially small deviations from the naïve model will have a trivially small impact on the study's conclusions. Sufficiently large deviations from the naïve model will overturn the results of any study. Because these two facts are always true, they quickly exhaust their usefulness. Therefore, the magnitude of the deviation is all-important. The sensitivity of an observational study to bias from an unmeasured covariate u is the magnitude of the departure from the naïve model that would need to be present to materially alter the study's conclusions.[11]

The first sensitivity analysis in an observational study concerned smoking and lung cancer. In 1959, Jerry Cornfield and his colleagues [15] asked about the magnitude of the bias from an unobserved covariate u needed to alter the conclusion

[11] In general, a sensitivity analysis asks how the conclusion of an argument dependent upon assumptions would change if the assumptions were relaxed. The term is sometimes misused to refer to performing several parallel statistical analyses without regard to the assumptions upon which they depend. If several statistical analyses all depend upon the same assumption — for instance, the naïve model (3.5) — then performing several such analyses provides no insight into consequences of the failure of that assumption.

from observational studies that heavy smoking causes lung cancer. They concluded that the magnitude of the bias would need to be enormous.

The sensitivity analysis model: Quantitative deviation from random assignment

The naïve model (3.5)–(3.8) said that two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have the same probability of treatment given $(r_T, r_C, \mathbf{x}, u)$, i.e., $\pi_k = \pi_\ell$, where $\pi_k = \Pr(Z_k = 1 \mid r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k)$ and $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$. The sensitivity analysis model speaks about the same probabilities in (3.1), saying that the naïve model (3.5)–(3.8) may be false, but to an extent controlled by a parameter, $\Gamma \geq 1$. Specifically, it says that two people, k and ℓ, with the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, have odds[12] of treatment, $\pi_k / (1 - \pi_k)$ and $\pi_\ell / (1 - \pi_\ell)$, that differ by at most a multiplier of Γ; that is, in (3.1),

$$\boxed{\frac{1}{\Gamma} \leq \frac{\pi_k / (1 - \pi_k)}{\pi_\ell / (1 - \pi_\ell)} \leq \Gamma \quad \text{whenever} \quad \mathbf{x}_k = \mathbf{x}_\ell\,.} \tag{3.13}$$

If $\Gamma = 1$ in (3.13), then $\pi_k = \pi_\ell$, so (3.5)–(3.8) is true; that is, $\Gamma = 1$ corresponds with the naïve model. In §3.1, expression (3.1) was seen to be a representation and not a model — something that is always true for suitably defined u_ℓ — but that representation took $\pi_\ell = 0$ or $\pi_\ell = 1$, which implies $\Gamma = \infty$ in (3.13). In other words, numeric values of Γ between $\Gamma = 1$ and $\Gamma = \infty$ define a spectrum that begins with the naïve model (3.5)–(3.8) and ends with something that is hollow in the sense that it is always true, namely (3.1). The hollow statement that is always true, namely (3.1), is the statement that 'association does not imply causation,' that is, a sufficiently large departure from the naïve model can explain away as noncausal any observed association.

If $\Gamma = 2$, and if you, k, and I, ℓ, look the same, in the sense that we have the same observed covariates, $\mathbf{x}_k = \mathbf{x}_\ell$, then you might be twice as likely as I to receive the treatment because we differ in ways that have not been measured. For instance, if your $\pi_k = 2/3$ and my $\pi_\ell = 1/2$, then your odds of treatment rather than control are $\pi_k / (1 - \pi_k) = 2$ or 2-to-1, whereas my odds of treatment rather than control are $\pi_\ell / (1 - \pi_\ell) = 1$ or 1-to-1, and you are twice as likely as I to receive treatment, $\{\pi_k / (1 - \pi_k)\} / \{\pi_\ell / (1 - \pi_\ell)\} = 2$ in (3.13).[13]

---

[12] Odds are an alternative way of expressing probabilities. Probabilities and odds carry the same information in different forms. A probability of $\pi_k = 2/3$ is an odds of $\pi_k / (1 - \pi_k) = 2$ or 2-to-1. Gamblers prefer odds to probabilities because odds express the chance of an event in terms of fair betting odds, the price of a fair bet. It is easy to move from probability π_k to odds $\omega_k = \pi_k / (1 - \pi_k)$ and back again from odds ω_k to probability $\pi_k = \omega_k / (1 + \omega_k)$.

[13] Implicitly, the critic is saying that the failure to measure u is the source of the problem, or that (3.5) would be true with $(\mathbf{x}, u)$ in place of $\mathbf{x}$, but is untrue with $\mathbf{x}$ alone. That is, the critic is saying $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell, u_\ell)$. As in §3.1, because of the delicate nature of unobserved variables, this is a manner of speaking rather than a tangible distinction. If the formalities are understood to refer to $\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)$, then it is not necessary to

Table 3.3 Sensitivity analysis for the one-sided 95% confidence interval for a constant, additive treatment effect τ on DNA elution rates. As usual, the hypothesis of a constant effect $H_0: \tau = \tau_0$ is tested by testing no effect on $Y_i - \tau_0$ for the given value of Γ. The one-sided 95% confidence interval is the set of values of τ_0 not rejected in the one-sided, 0.05 level test. As Γ increases, there is greater potential deviation from random treatment assignment in (3.13), and the confidence interval grows longer. For instance, a treatment effect of $\tau_0 = 0.30$ would be implausible in a randomized experiment, $\Gamma = 1$, but not in an observational study with $\Gamma = 2$.

Γ	1	2	3
95% Interval	$[0.37, \infty)$	$[0.21, \infty)$	$[0.094, \infty)$

$$E\left(\overline{T}\,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{1}{1+\Gamma}\sum_{i=1}^{I} s_i q_i, \tag{3.26}$$

while the variance becomes

$$\operatorname{var}\left(\overline{T}\,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \operatorname{var}\left(\overline{\overline{T}}\,\middle|\, \mathscr{F}, \mathscr{Z}\right) = \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I} (s_i q_i)^2. \tag{3.27}$$

The remaining calculations are unchanged.

Sensitivity analysis for a confidence interval

Table 3.3 is the sensitivity analysis for the one-sided 95% confidence interval for an additive, constant treatment effect discussed in §2.4.2. As in a randomized experiment, the hypothesis that $H_0: r_{Tij} = r_{Cij} + \tau_0$ is tested by testing the null hypothesis of no treatment effect on the adjusted responses, $R_{ij} - \tau_0 Z_{ij}$, or equivalently on the adjusted, treated-minus-control pair differences, $Y_i - \tau_0$. The one-sided 95% confidence interval is the set of values of τ_0 not rejected by a one-sided, 0.05 level test.

From Table 3.2, the hypothesis $H_0: \tau = \tau_0$ for $\tau_0 = 0$ is barely rejected for $\Gamma = 4$ because the maximum possible one-sided P-value is 0.047. For $\Gamma = 3$, the maximum possible one-sided P-value is 0.04859 for $\tau_0 = .0935$ and is 0.05055 for $\tau_0 = .0936$, so after rounding to two significant digits, the one-sided 95% confidence interval is $[0.094, \infty)$.

Sensitivity analysis for point estimates

For each value of $\Gamma \geq 1$, a sensitivity analysis replaces a single point estimate, say $\widehat{\tau}$, by an interval of point estimates, say $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ that are the minimum and maximum point estimates for all distributions of treatment assignments satisfying (3.16)–(3.18). Unlike a test or a confidence interval, and like a point estimate, this interval $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ does not reflect sampling uncertainty; however, it does reflect uncertainty introduced by departures from random treatment assignment in (3.13) or (3.16)–(3.18).

Package ‘rbounds’

February 20, 2015

Version 2.1

Title Perform Rosenbaum bounds sensitivity tests for matched and unmatched data.

Date 2014-12-7

Author Luke J. Keele

Maintainer Luke J. Keele <ljk20@psu.edu>

Depends R (>= 2.8.1), Matching

Description Takes matched and unmatched data and calculates Rosenbaum bounds for the treatment effect. Calculates bounds for binary outcome data, Hodges-Lehmann point estimates, Wilcoxon signed-rank test for matched data and matched IV estimators, Wilcoxon sum rank test, and for data with multiple matched controls. Package is also designed to work with the Matching package and operate on Match() objects.

NeedsCompilation no

Repository CRAN

Date/Publication 2014-12-08 07:23:24

R topics documented:

Observational Studies 1 (2015) 1-17 Submitted 7/14; Published 11/14

Two R Packages for Sensitivity Analysis in Observational Studies

Paul R. Rosenbaum **rosenbaum@wharton.upenn.edu**
Department of Statistics
Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340 US

Abstract

Two `R` packages for sensitivity analysis in observational studies are described. Package `sensitivitymw` is for matched pairs with one treated subject and one control, or matched sets with one treated subject and a fixed number, $K \geq 2$, of controls. Package `sensitivitymv` is for matched sets with variable numbers of controls. The packages offer conventional statistics, such as the permutational t-test and M-statistics using Huber's weights, but they also offer less familiar test statistics that have higher power in sensitivity analyses. The packages provide several tools useful in sensitivity analyses, such as an aid, `amplify`, to the interpretation of the value of the sensitivity parameter, and a device for combining evidence from several independent sensitivity analyses, `truncatedP`, for instance, several evidence factors or several subgroups.

Keywords: M-test; observational study; permutational t-test; randomization inference; sensitivity analysis.

1. Introduction

1.1 `R` Packages `sensivitymv` and `sensitivitymw`

The two `R` packages `sensivitymv` and `sensitivitymw` perform sensitivity analyses for observational studies with matched pairs or matched sets containing multiple controls. Package `sensitivitymw` is for matched pairs or matching with a fixed number of controls, for instance matching each treated subject to two controls. In contrast, package `sensivitymv` is for matched sets with variable numbers of controls, perhaps some treatment-control pairs together with some triples containing a treated subject and two controls. Also, the packages contain several data sets and several additional functions useful in sensitivity analysis. The packages overlap considerably, but package `sensitivitymw` is faster with additional features for matched pairs and for matching with a fixed number of controls. Both packages are available at CRAN and contain documentation.

My purpose here is to present a gentle introduction to these `R` packages, with pointers to articles for technical detail and pointers to the software documentation for additional options.

1.2 Scope of the current discussion

In an observational study, a sensitivity analysis replaces qualitative claims about whether unmeasured biases are present with an objective quantitative statement about the magnitude of bias that would need to be present to change the conclusions. In this sense, a sensitivity analysis speaks to the assertion "it might be bias" in much the same way that a P-value speaks to the assertion "it might be bad luck". If someone asserted that the higher responses in the treated group in a randomized experiment "might be bad luck," an unlucky randomization with no treatment effect, then a P-value does not deny the logical possibility of bad luck, but objectively measures the quantity of bad luck that would need to be present to alter the impression that the treatment did have an effect. In parallel, a sensitivity analysis measures the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the conclusions of an observational study.

A sensitivity analysis is one tool useful in the large task of designing and interpreting an observational study. The discussion here is rather narrowly focused on carrying out such a sensitivity analysis in `R`.

1.3 What do the packages do?

In an observational study, treated and control subjects may be matched to be similar in terms of observed or measured covariates, but people who look similar in terms of measured covariates may still differ in terms of unmeasured covariates. The packages perform a sensitivity analysis asking about the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the qualitative conclusions of a naive analysis that presumes matching for observed covariates removes all bias.

In a matched randomized experiment, each subject in a matched set has the same chance of being assigned to treatment or control because randomization has ensured that this is so. Without randomization, two people who look similar may differ in their chances of receiving treatment because they differ in terms of an unmeasured covariate not controlled by matching for measured covariates. The sensitivity analysis assumes that one subject in a matched set may be $\Gamma \geq 1$ times more likely than another to receive treatment because they differ in terms of unobserved covariates. If $\Gamma = 1$, then subjects who look the same are the same: matched subjects have equal chances of treatment, as in a randomized experiment. For $\Gamma = 1$, the sensitivity analysis reports a single answer, for instance a single P-value testing the null hypothesis of no treatment effect, and that single answer is the P-value that would be appropriate in a matched randomized experiment. For $\Gamma > 1$, there is no longer a single P-value, but rather an interval of possible P-values. The sensitivity analysis asks: How large must Γ be before the interval is so long that it is inconclusive, perhaps both accepting and rejecting the null hypothesis of no effect at the 0.05 level? The interval of possible P-values would be inconclusive in this sense if it extended from below 0.05 to above 0.05. The `senmw` and `senmv` functions compute sensitivity bounds for P-values. Specifically, they compute the upper bound on the P-value, for a specific Γ, so if that upper bound is at most 0.05, then a bias of magnitude Γ is too small to lead to acceptance of the null hypothesis. The `senmwCI` function inverts bounds on P-values to obtain sensitivity bounds for confidence intervals and point estimates. For detailed discussion of this model, see Rosenbaum (2002, §4; 2007).