Introduction to Propensity Score Analysis with R

useR! 2013 Pre-Conference Tutorial

Jason M. Bryer [1,2]
jason@bryer.org

Robert M. Pruzek [2]
rpruzek@albany.edu

[1]Excelsior College
Albany, NY 12203

[2]University at Albany
Albany, NY, 12222

`http://github.com/jbryer/IntroPSAwithR`

July 9, 2013

Agenda

1. Randomized Experiments
2. Defining Propensity Scores
3. Different Methods of PSA
4. The Lalonde Example
5. Multilevel PSA
6. Matching of Non-Binary Treatments

Agenda

The Randomized Experiment

Considered to be the *gold standard* for estimating causal effects.

- Effects can be estimated using simple means between groups, or blocks in randomized block design.
- Randomization presumes unbiasedness and balance between groups.

However, randomization is often not feasible for many reasons, especially in educational contexts.

The Randomized Experiment

Considered to be the *gold standard* for estimating causal effects.

- Effects can be estimated using simple means between groups, or blocks in randomized block design.
- Randomization presumes unbiasedness and balance between groups.

However, randomization is often not feasible for many reasons, especially in educational contexts.

The strong ignorability assumption states that:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | X_i = x$$

for all X_i.

Rubin Causal Model[1]

- The causal effect of a treatment is the difference in an individual's outcome under the situation they were given the treatment and not (referred to as a counterfactual).

$$\delta_i = Y_{i1} - Y_{i0}$$

- However, it is impossible to directly observe δ_i (referred to as *The Fundamental Problem of Causal Inference*, Holland 1986).
- Rubin frames this problem as a "missing data problem."

[1] See Rubin, 1974, 1977, 1978, 1980, and Holland, 1986

Agenda

Propensity Score Analysis

Propensity score analysis (PSA) is a quasi-experimental design used to estimate causal effects in observational studies (i.e. studies where students are not randomized to treatment). PSA is conducted in two phases:

Phase I (Also referred to as the design phase) In phase one we are concerned with adjusting for selection bias. We model treatment placement using observed variables (see next slide). The propensity score is the probability of a student being in the treatment. With estimated propensity scores, clusters or matches are created for phase II.

Phase II With matches or clusters made in phase I, we compare the difference between matches or clusters on the outcome measure of interest.

Propensity Score Analysis

The propensity score is the "conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983, p. 41). The probability of being in the treatment:

$$\pi(X_i) \equiv Pr(T_i = 1|X_i)$$

The balancing property under exogeneity:

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

We can then restate the ignorability assumption with the propensity score:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \mid \pi(X_i)$$

Treatment Effects

The average treatment effect (ATE) is defined as:

$$E(r_1) - E(r_0)$$

where $E(.)$ is the expectation in the population. For a set of covariates, X, and outcomes Y where 0 denotes control and 1 treatment, we define ATE as:

$$ATE = E(Y_1 - Y_0|X) = E(Y_1|X) - E(Y_0|X)$$

The Average treatment effect on the treated (ATT), is defined as:

$$ATT = E(Y_1 - Y_0|X, C = 1) = E(Y_1|X, C = 1) - E(Y_0|X, C = 1)$$

Agenda

Propensity score methods

Matching Each treatment unit is paired with a comparison unit based upon the pre-treatment covariates.

Stratification Treatment and comparison units are divided into strata (or subclasses) so that treated and comparison units are similar within each strata. Cochran (1968) observed that creating five subclassifications (stratum) removes at least 90% of the bias in the estimated treatment effect.

Weighting Each observation is weighted by the inverse of the probability of being in that group.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\pi(X_i)} - \frac{(1-T_i)Y_i}{1-\pi(X_i)}\right)$$

Steps for Implementing Matching Methods

Stuart and Rubin (2008) outline the following steps for matching, but the same approach can be used for stratification and weighting as well.

1. Choose the covariates to be used.
2. Define a distance measure (i.e. what constitutes similar).
3. Choose the matching algorithm.
4. Diagnose the matches (or strata) obtained (iterating through steps 2 and 3 as well).
5. Estimate the treatment effect using the matches (or strata) found in step 4.

Matching Methods

There are many choices and approaches to matching, including:

- Propensity score matching.
- Limited exact matching.
- Full matching.
- Nearest neighbor matching.
- Optimal/Genetic matching.
- Mahalanobis distance matching (for quantiative covariates only).
- Matching with and without replacement.
- One-to-one or one-to-many matching.

Which matching method should you use?

Matching Methods

There are many choices and approaches to matching, including:

- Propensity score matching.
- Limited exact matching.
- Full matching.
- Nearest neighbor matching.
- Optimal/Genetic matching.
- Mahalanobis distance matching (for quantiative covariates only).
- Matching with and without replacement.
- One-to-one or one-to-many matching.

Which matching method should you use?

Whichever one gives the best balance!

See Rosenbaum (2012), *Testing one hypothesis twice in observational studies*.

Agenda

National Supported Work

The National Supported Work (NSW) Demonstration was a federally and privately funded randomized experiment done in the 1970s to estimate the effects of a job training program for disadvantaged workers.

- Participants were randomly selected to participate in the training program.
- Both groups were followed up to determine the effect of the training on wages.
- Analysis of the mean differences (unbiased given randomization), was approximately $800.

Lalonde (1986) used data from the Panel Survey of Income Dynamics (PSID) and the Current Population Survey (CPS) to investigate whether non-experimental methods would result in similar results to the randomized experiment. He found results ranging from $700 to $16,000.

National Supported Work (cont.)

Dehejia and Wahba (1999) later used propensity score matching to analyze the data. The found that,

- Comparison groups selected by Lalonde were very dissimilar to the treated group.
- By restricting the comparison group to those that were similar to the treated group, they could replicate the original NSW results.
- Using the CPS data, the range of treatment effect was between \$1,559 to \$1,681. The experimental results for the sample sample was approximately \$1,800.

The covariates available include: age, eduction level, high school degree, marital status, race, ethnicity, and earning sin 1974 and 1975.

Outcome of interest is earnings in 1978.

```
> data(lalonde, package='Matching')
```

Estimating Propensity Scores

```
> lalonde.formu <- treat~age + educ  + black + hisp + married + nodegr + re74 + re75
> glm1 <- glm(lalonde.formu, family=binomial, data=lalonde)
> summary(glm1)

Call:
glm(formula = lalonde.formu, family = binomial, data = lalonde)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.436  -0.990  -0.907   1.282   1.695

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.18e+00   1.06e+00    1.12    0.265
age          4.70e-03   1.43e-02    0.33    0.743
educ        -7.12e-02   7.17e-02   -0.99    0.321
black       -2.25e-01   3.66e-01   -0.61    0.539
hisp        -8.53e-01   5.07e-01   -1.68    0.092 .
married      1.64e-01   2.77e-01    0.59    0.555
nodegr      -9.04e-01   3.13e-01   -2.88    0.004 **
re74        -3.16e-05   2.58e-05   -1.22    0.221
re75         6.16e-05   4.36e-05    1.41    0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 604.20  on 444  degrees of freedom
Residual deviance: 587.22  on 436  degrees of freedom
AIC: 605.2

Number of Fisher Scoring iterations: 4
```

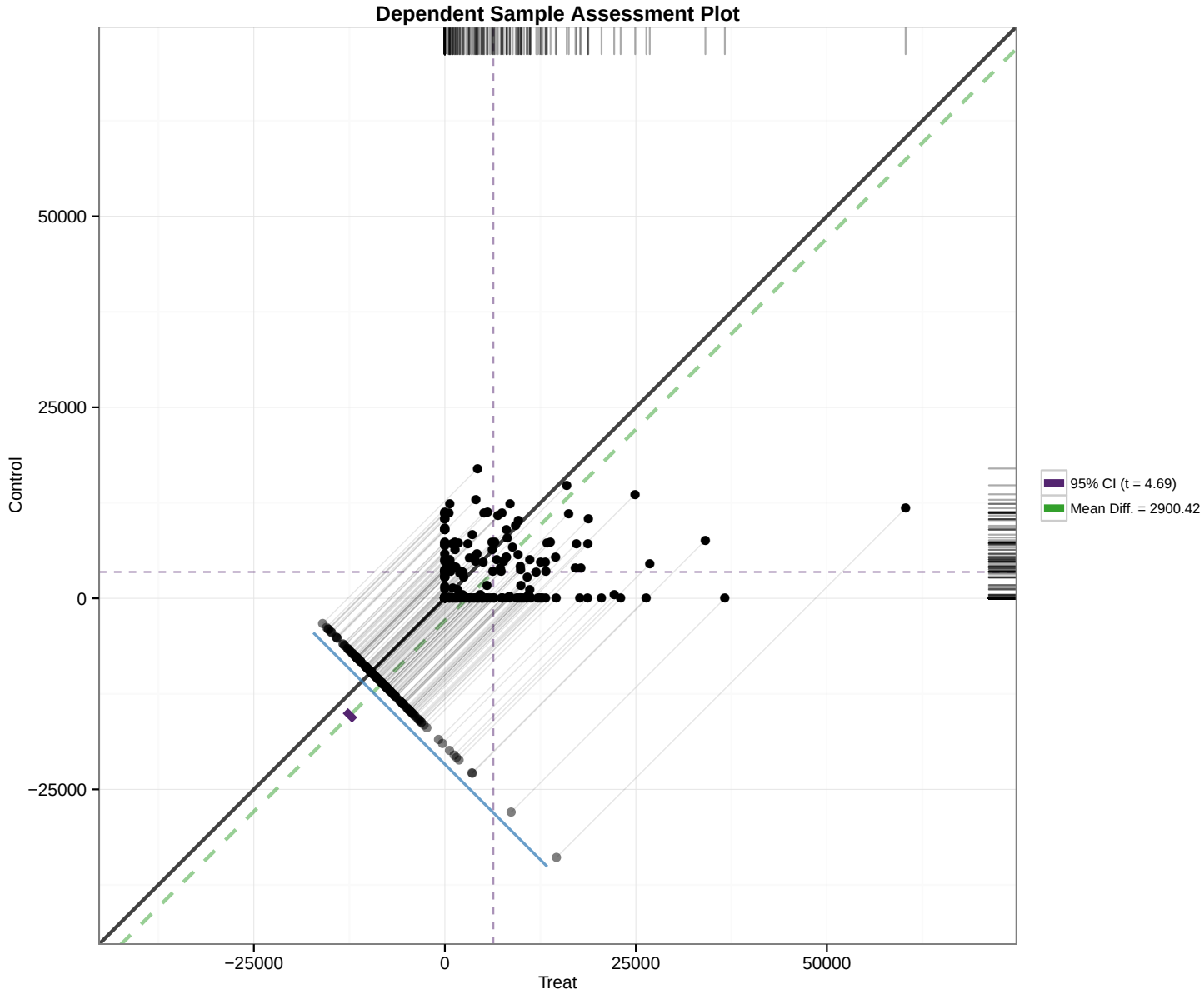
Estimating Propensity Scores

```
> ps <- fitted(glm1)  # Propensity scores
> Y  <- lalonde$re78  # Dependent variable, real earnings in 1978
> Tr <- lalonde$treat # Treatment indicator
> rr <- Match(Y=Y, Tr=Tr, X=ps, M=1, ties=FALSE)
> summary(rr) # The default estimate is ATT here

Estimate...  2900.4
SE.........  616.89
T-stat.....  4.7017
p.val......  2.5807e-06

Original number of observations..............  445
Original number of treated obs...............  185
Matched number of observations...............  185
Matched number of observations  (unweighted).  185
```

Visualizing Results

```
> matches <- data.frame(Treat=lalonde[rr$index.treated,'re78'],
   Control=lalonde[rr$index.control,'re78'])
> print(granovagg.ds(matches))
```

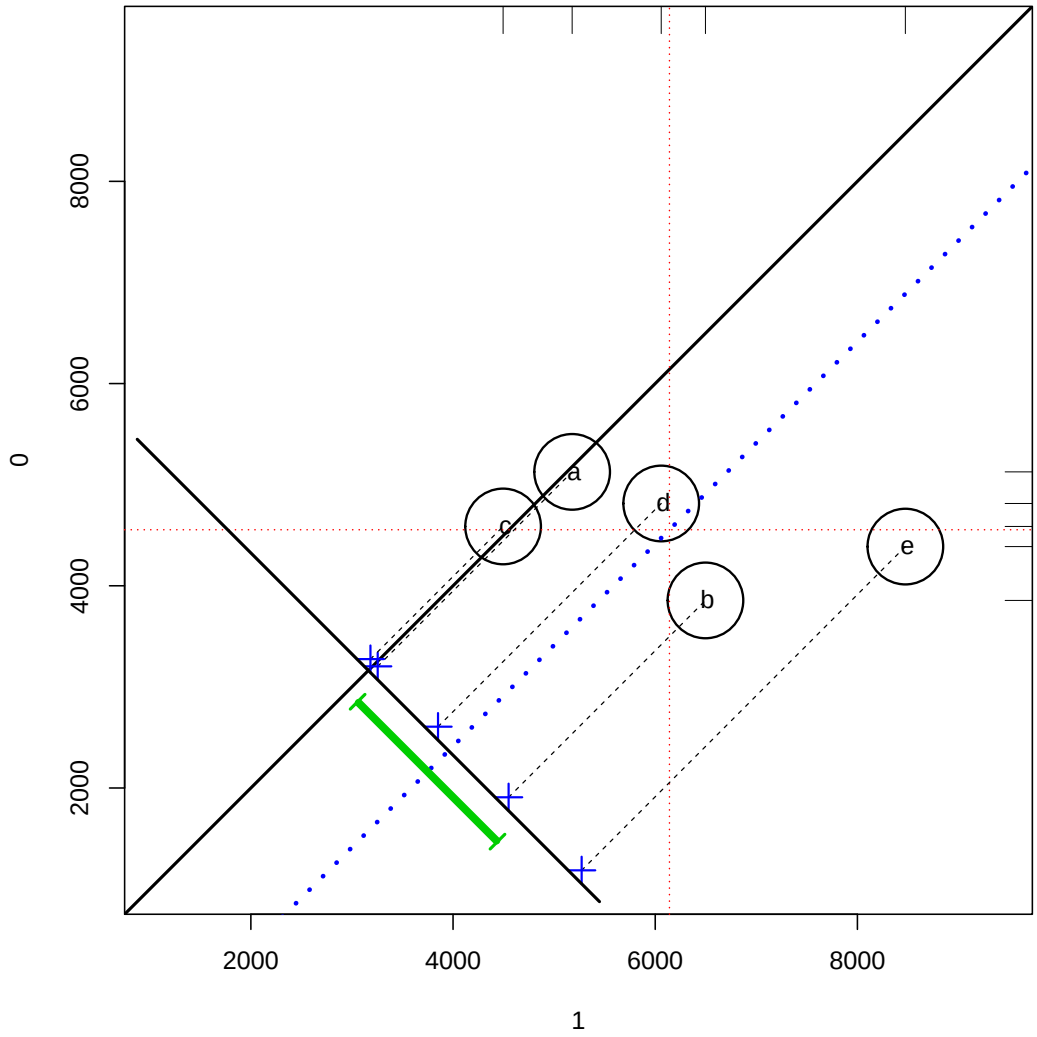
	Summary Statistics
n	185.00
Treat mean	6349.15
Control mean	3448.72
mean(D = Treat - Control)	2900.42
SD(D)	8413.44
Effect Size	0.34
r(Treat, Control)	0.12
r(Treat + Control, D)	0.58
Lower 95% Confidence Interval	1680.02
Upper 95% Confidence Interval	4120.82
t (D-bar)	4.69
df.t	184.00
p-value (t-statistic)	0.00


Dependent Sample Assessment Plot
Control
Treat
50000
25000
0
−25000
−25000
0
25000
50000
95% CI (t = 4.69)
Mean Diff. = 2900.42

Stratification (5 Strata)

```
> strata <- cut(ps, quantile(ps, seq(0, 1, 1/5)), include.lowest=TRUE, labels=letters[1:5])
> circ.psa(lalonde$re78, lalonde$treat, strata, revc=TRUE)

$summary.strata
  n.0 n.1 means.0 means.1
a  62  27    5126    5178
b  59  30    3855    6497
c  56  33    4587    4495
d  42  47    4814    6059
e  41  48    4388    8474

$wtd.Mn.1
[1] 6141

$wtd.Mn.0
[1] 4554

$ATE
[1] -1587

$se.wtd
[1] 694

$approx.t
[1] -2.3

$df
[1] 435

$CI.95
[1] -2950  -224
```

Stratification (5 Strata)

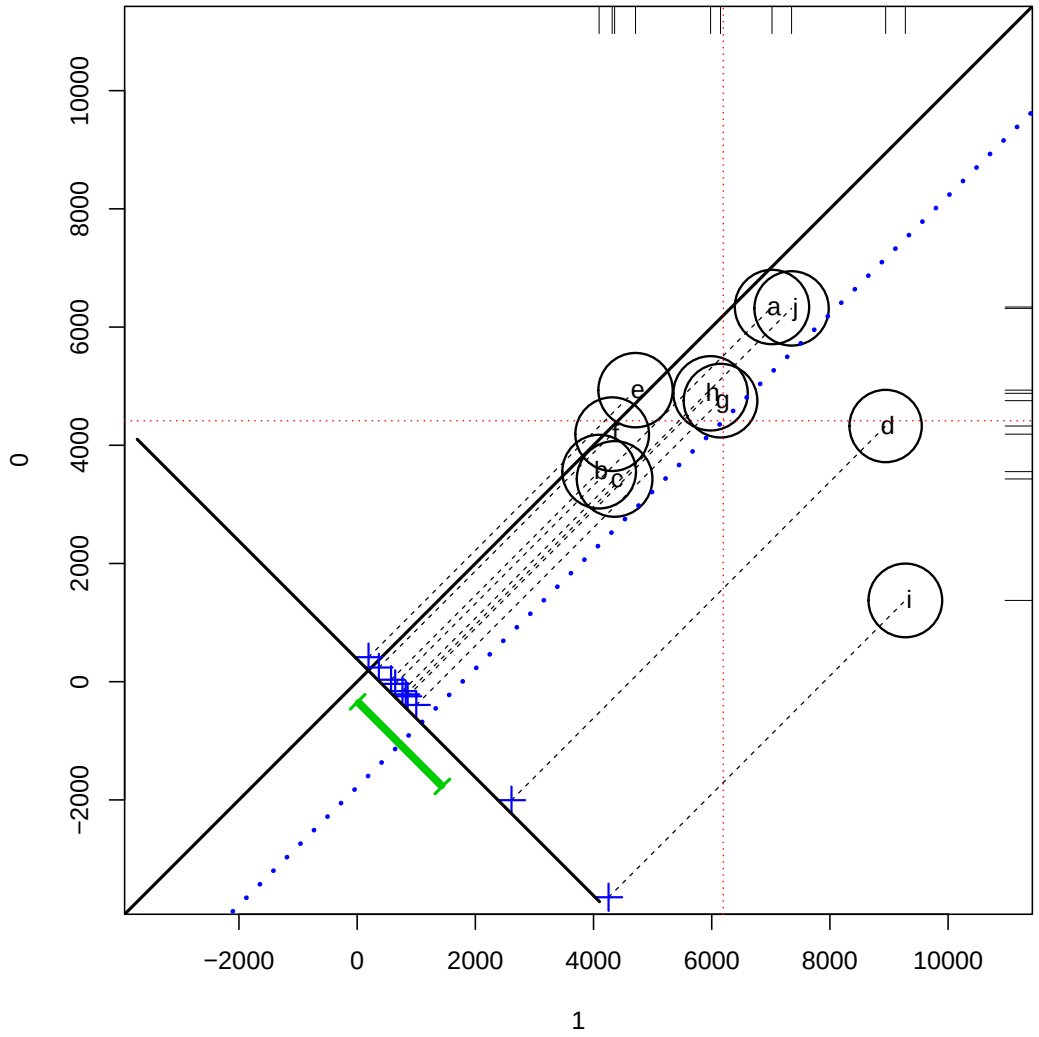


Stratification (10 Strata)

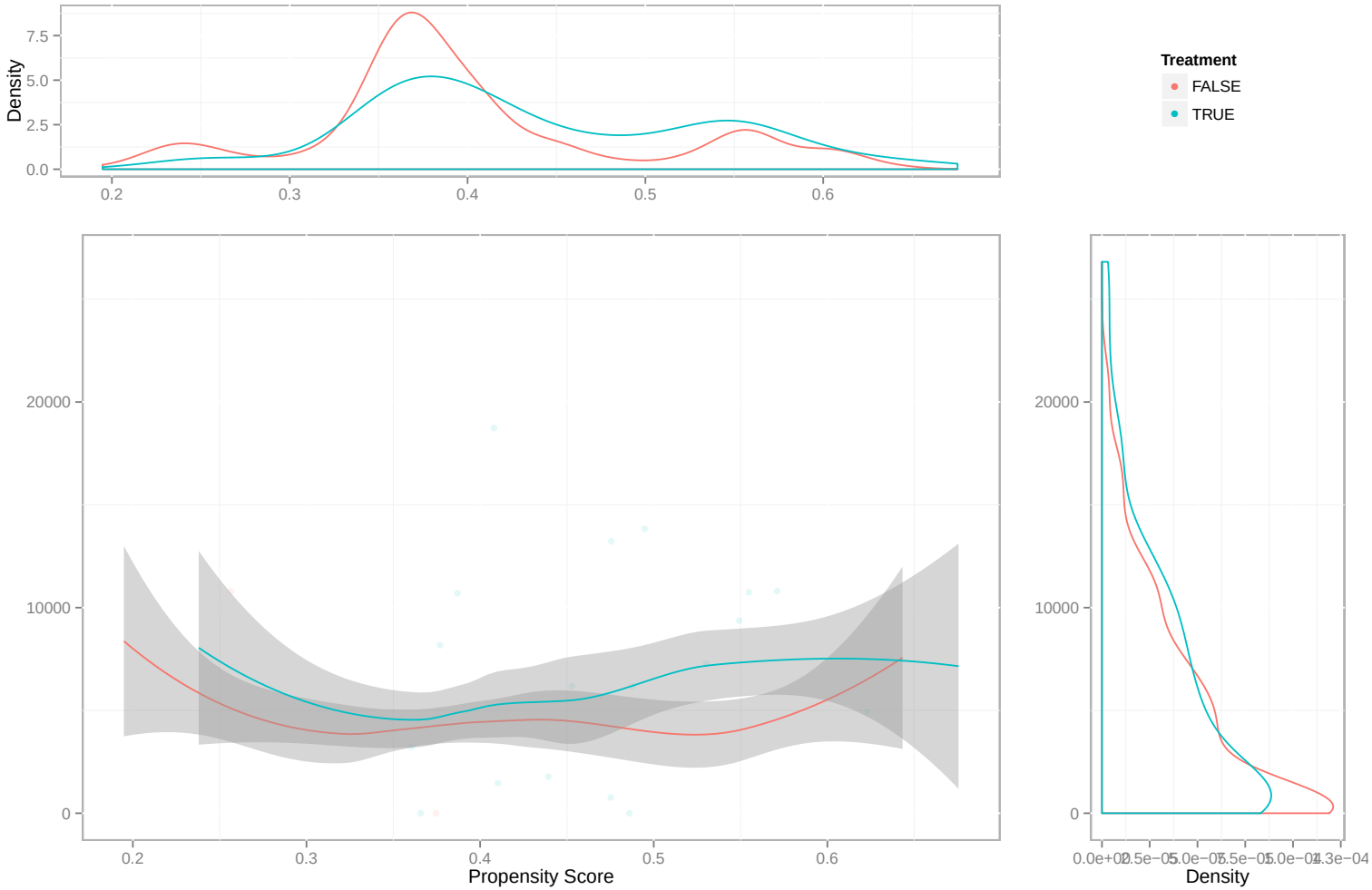
```
> strata10 <- cut(ps, quantile(ps, seq(0, 1, 1/10)), include.lowest=TRUE, labels=letters[1:10])
> circ.psa(lalonde$re78, lalonde$treat, strata10, revc=TRUE)
```

```
$summary.strata
  n.0 n.1 means.0 means.1
a  35  10    6339    7020
b  27  17    3554    4095
c  31  16    3430    4357
d  28  14    4326    8943
e  30  15    4933    4711
f  26  18    4188    4315
g  22  22    4755    6149
h  20  25    4879    5980
i  16  28    1375    9276
j  25  20    6316    7351

$wtd.Mn.1
[1] 6195

$wtd.Mn.0
[1] 4414

$ATE
[1] -1781

$se.wtd
[1] 711

$approx.t
[1] -2.5

$df
[1] 425

$CI.95
[1] -3178  -384
```

Stratification (10 Strata)

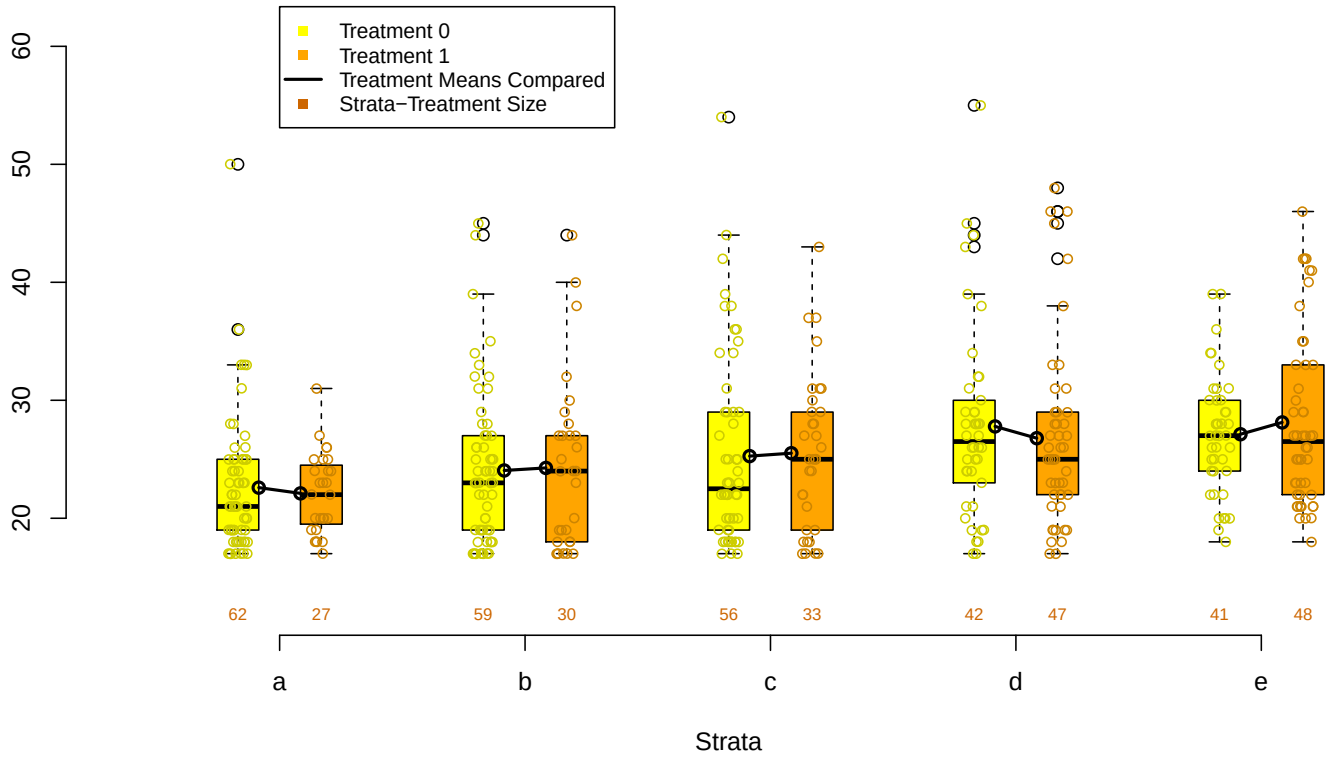


Loess Regression



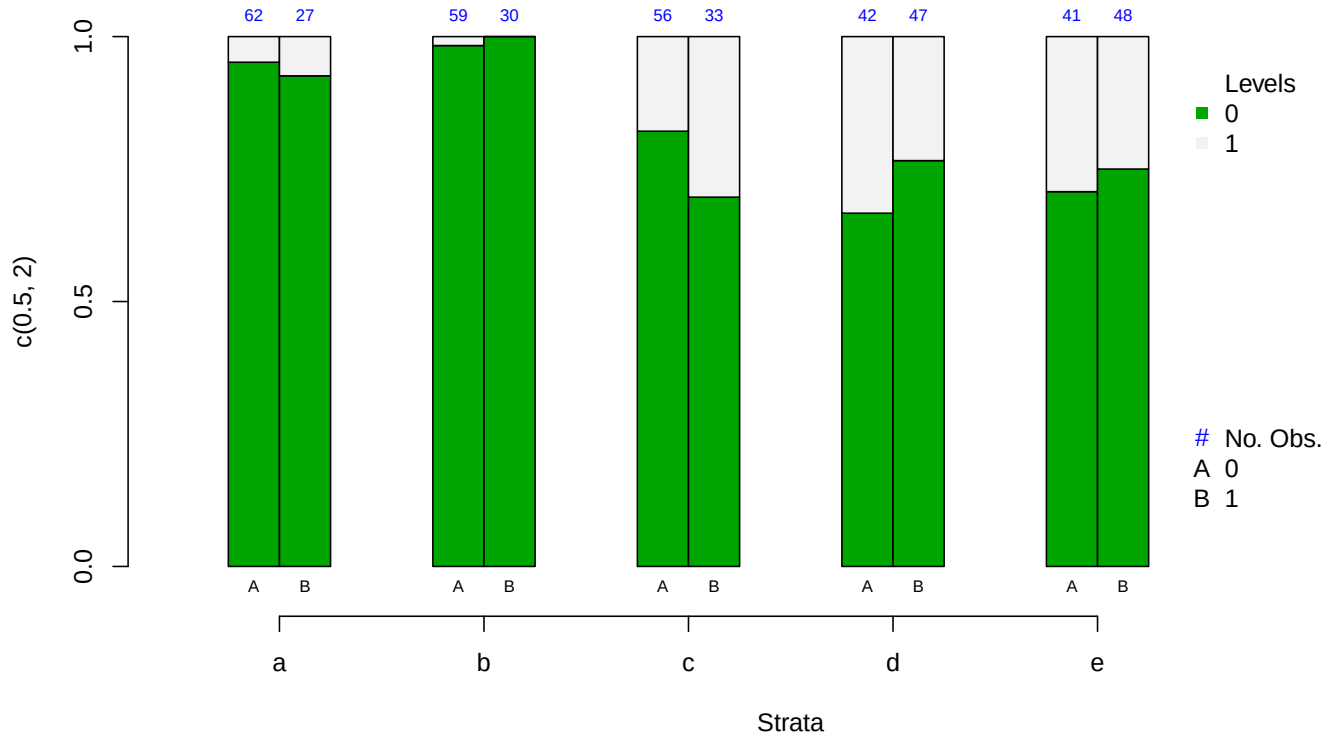
Checking Balance: Continuous Covariates

```
> box.psa(lalonde$age, lalonde$treat, strata, xlab="Strata",
    balance=FALSE)
```

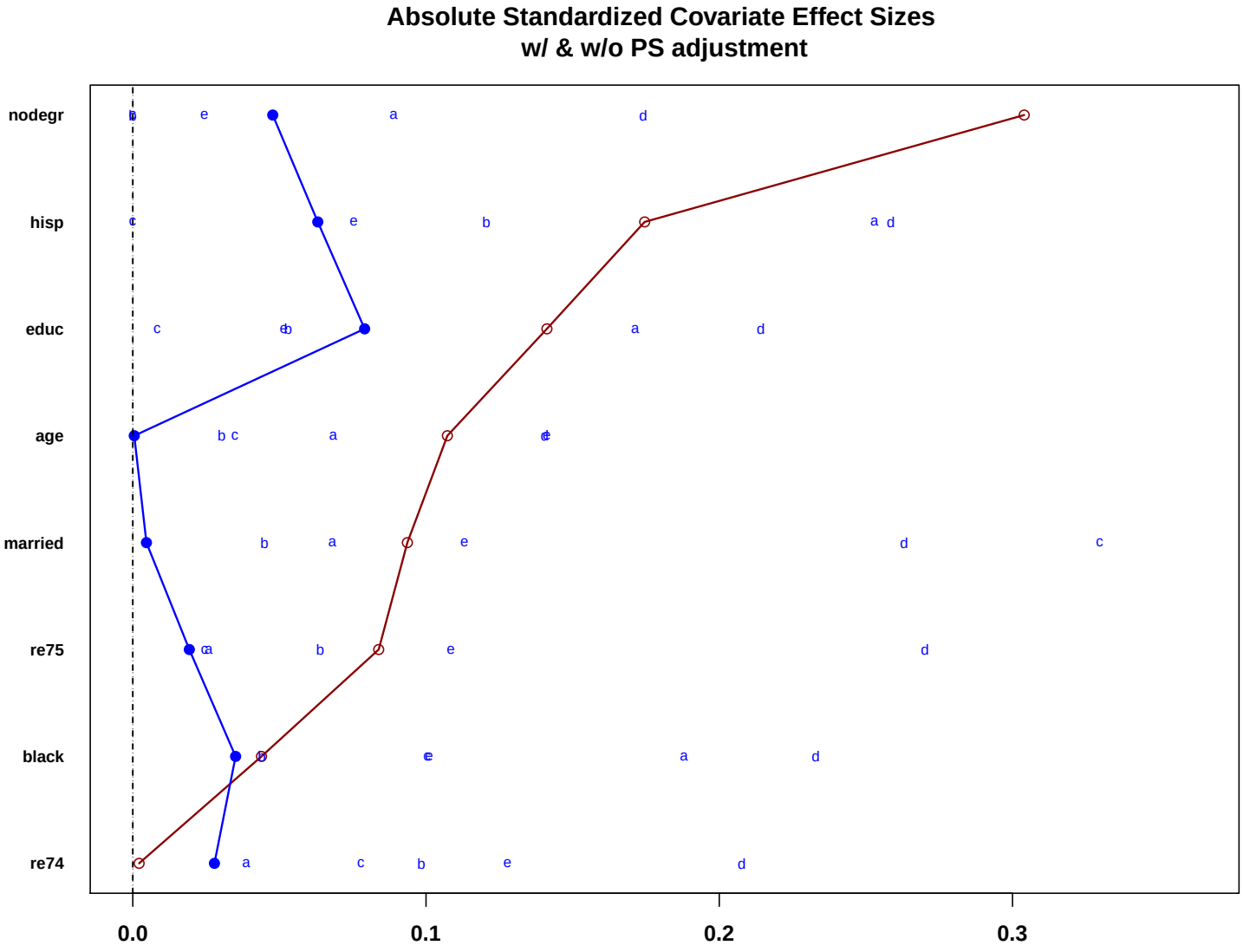


Checking Balance: Categorical Covariates

```
> cat.psa(lalonde$married, lalonde$treat, strata, xlab='Strata',
    balance=FALSE)
```



Checking Balance: Covariate Balance Plot



Agenda

Multilevel PSA

The use of PSA for clustered, or multilevel data, has been limited (Thoemmes & Felix, 2011). Bryer and Pruzek (2012, 2013) have introduced an approach to analyzing multilevel or clustered data using stratification methods and implemented in the `multilevelPSA` R package.

- Exact and partially exact matching methods implicitly adjust for clustering. That is, the covariates chosen to exactly match are, in essence, clustering variables.
- Exact matching only applies to phase I of PSA. How are the clusters related to outcome of interest.

The `multilevelPSA` uses stratification methods (e.g. quintiles, classification trees) by:

- Estimate separate propensity scores for each cluster.
- Identify strata within each cluster (e.g. leaves of classification trees, quintiles).
- Estimate ATE (or ATT) within each cluster.
- Aggregate estimated ATE to provide an overall ATE estimate.
- Several functions to summarize and visualize results and check balance.

The Programme of International Student Assessment (PISA)

- International assessment conducted by the Organization for Economic Co-operation and Development (OECD).
- Assesses students towards the end of secondary school (approximately 15-year-old children) in math, reading, and science.
- Collects a robust set of background information from students, parents, teachers, and schools.
- Assess both private and public school students in many countries.
- We will use PISA to estimate the effects of private school attendance on PISA outcomes.

Phase I of Multilevel PSA

The `multilevelPSA` provides two functions, `mlpsa.ctree` and `mlpsa.logistic`, that will estimate propensity scores using classification trees and logistic regression, respectively. Since logistic regression requires a complete dataset (i.e. no missing values), we will use classification trees in this example.

```
> data(pisana)
> data(pisa.colnames)
> data(pisa.psa.cols)
> student = pisana
> mlctree = mlpsa.ctree(student[,c('CNT','PUBPRIV',pisa.psa.cols)],
      formula=PUBPRIV ~ ., level2='CNT')
> student.party = getStrata(mlctree, student, level2='CNT')
> student.party$mathscore = apply(
   student.party[,paste0('PV', 1:5, 'MATH')], 1, sum) / 5
```

To assess what covariates were used in each tree model, as well as the relative importance, we can create a heat map of covariate usage by level.

```
> print(tree.plot(mlctree,
   level2Col=student$CNT,
   colLabels=pisa.colnames[,c('Variable','ShortDesc')]))
```

Covariate Heat Map

Phase II of Multilevel PSA

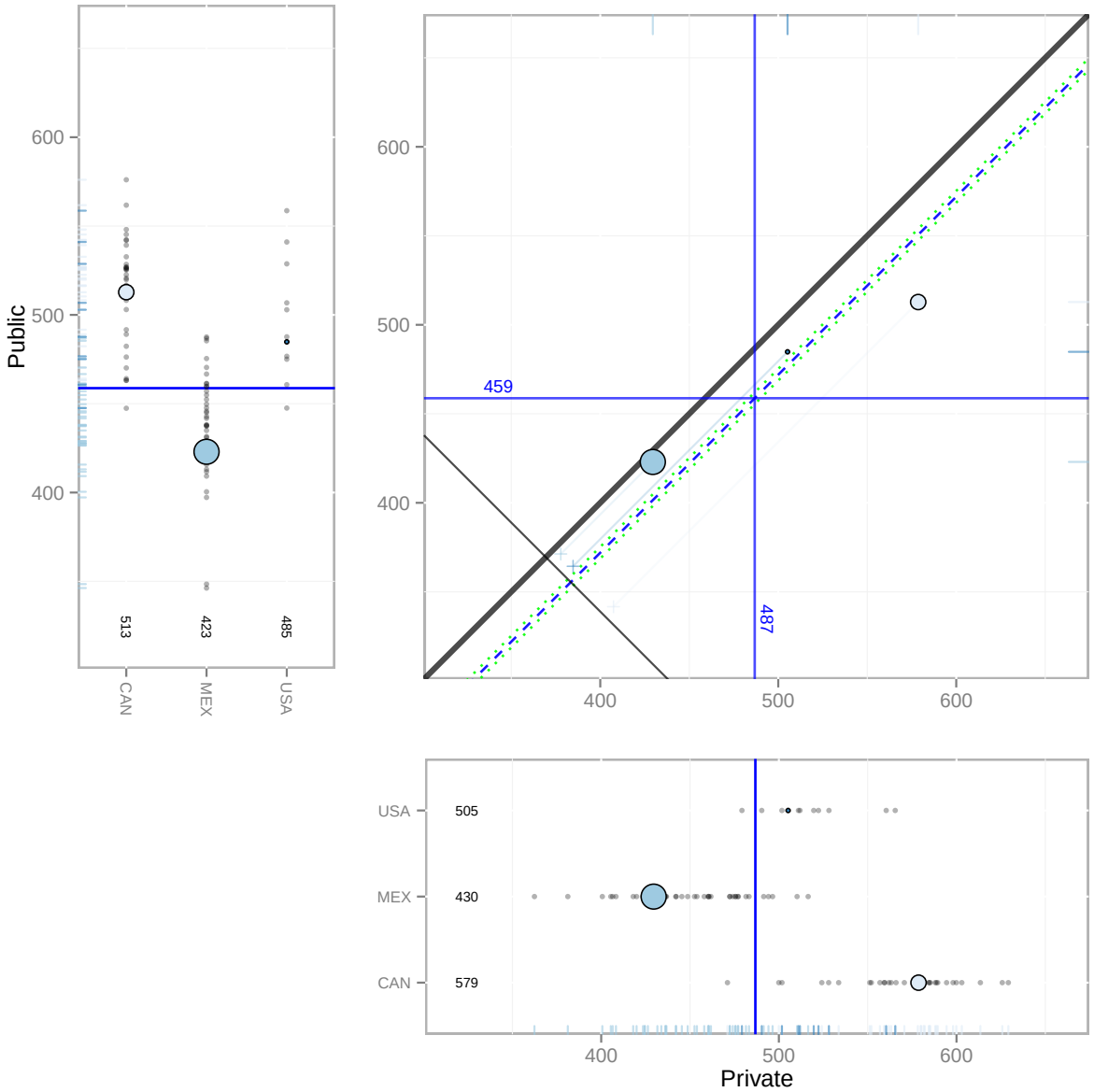
The mlpsa function will compare the outcome of interest.

```
> results.psa.math = mlpsa(response=student.party$mathscore,
    treatment=student.party$PUBPRIV, strata=student.party$strata,
    level2=student.party$CNT, minN=5)
> results.psa.math$overall.wtd
[1] -28
> results.psa.math$overall.ci
[1] -31 -25
> results.psa.math$level2.summary[,c('level2','Private','Private.n',
  'Public','Public.n','diffwtd','ci.min','ci.max')]
  level2 Private Private.n Public Public.n diffwtd ci.min ci.max
1    CAN     579      1625    513    21093   -65.8    -72  -59.6
2    MEX     430      4044    423    34090    -6.6    -10   -3.1
3    USA     505       345    485     4888   -20.4    -32   -8.8
```

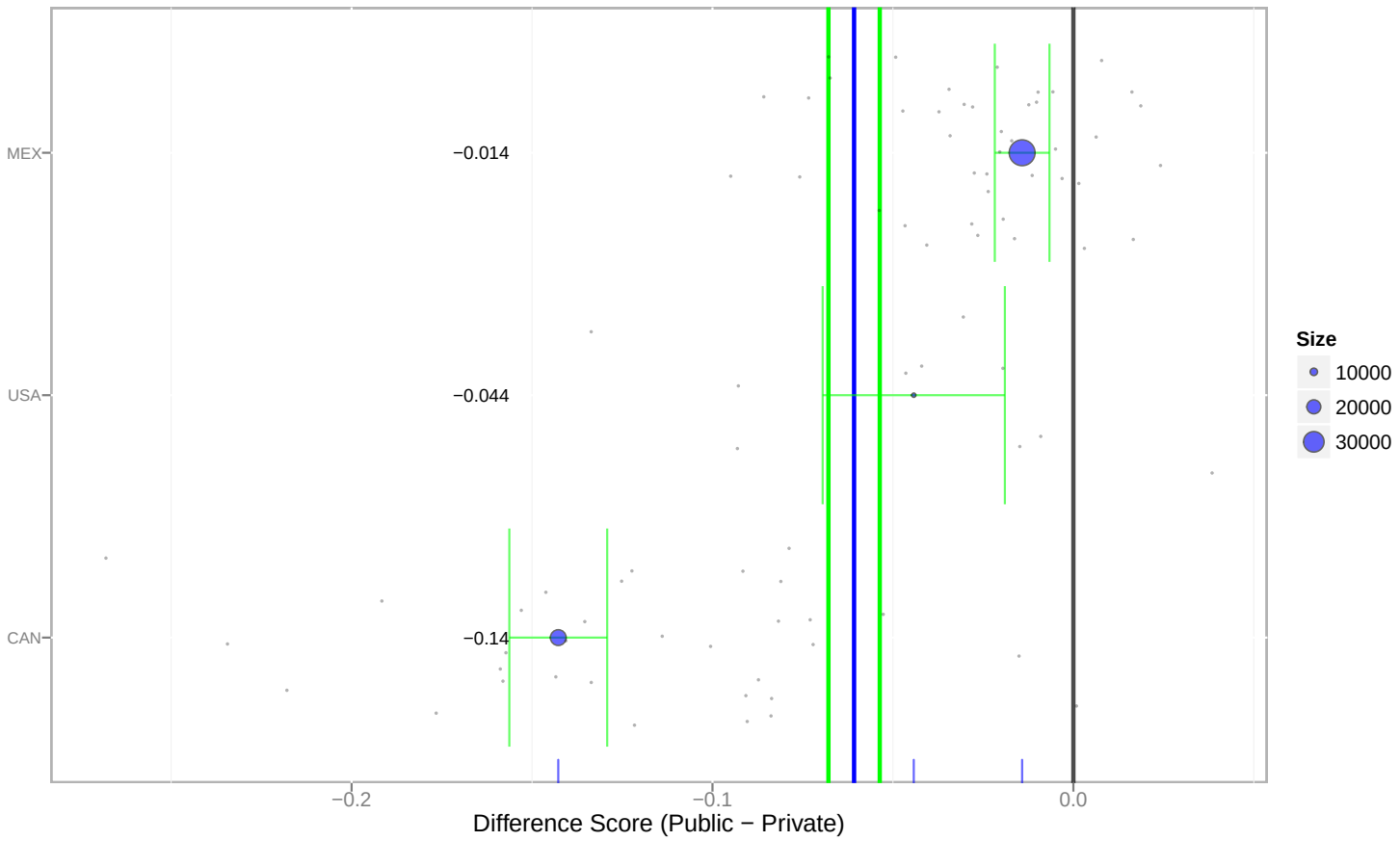
The multilevel PSA assessment plot is an extension of the circ.psa plot in PSAgraphics introduced by Helmreich and Pruzek (2009).

```
> print(plot(results.psa.math))
```

Multilevel PSA Assessment Plot



Multilevel PSA Difference Plot



Agenda

Matching of Non-Binary Treatments

- The `TriMatch` package provides functions for finding matched triplets.
- Estimates propensity scores for three separate logistic regression models (one for each pair of groups, that is, treat1-to-control, treat2-to-control, and treat1-to-treat2).
- Finds matched triplets that minimize the total distance (i.e. sum of the standardized distance between propensity scores within the three modesl). within a caliper.
- Provides multiple methods for determining which matched triplets are retained:
  - Optimal which attempts to retain all treatment units.
  - Full which retains all matched triplets within the specified caliper (.25 by default as suggested by Rosenbaum).
  - Analog of the one-to-many for matched triplets. Specify how many times each treat1 and treat2 unit can be matched.
  - Unique which allows each unit to be matched once, and only once.
- Functions for conducting repeated measures ANOVA and Freidman Ranksum Tests are provided.

Example: Tutoring

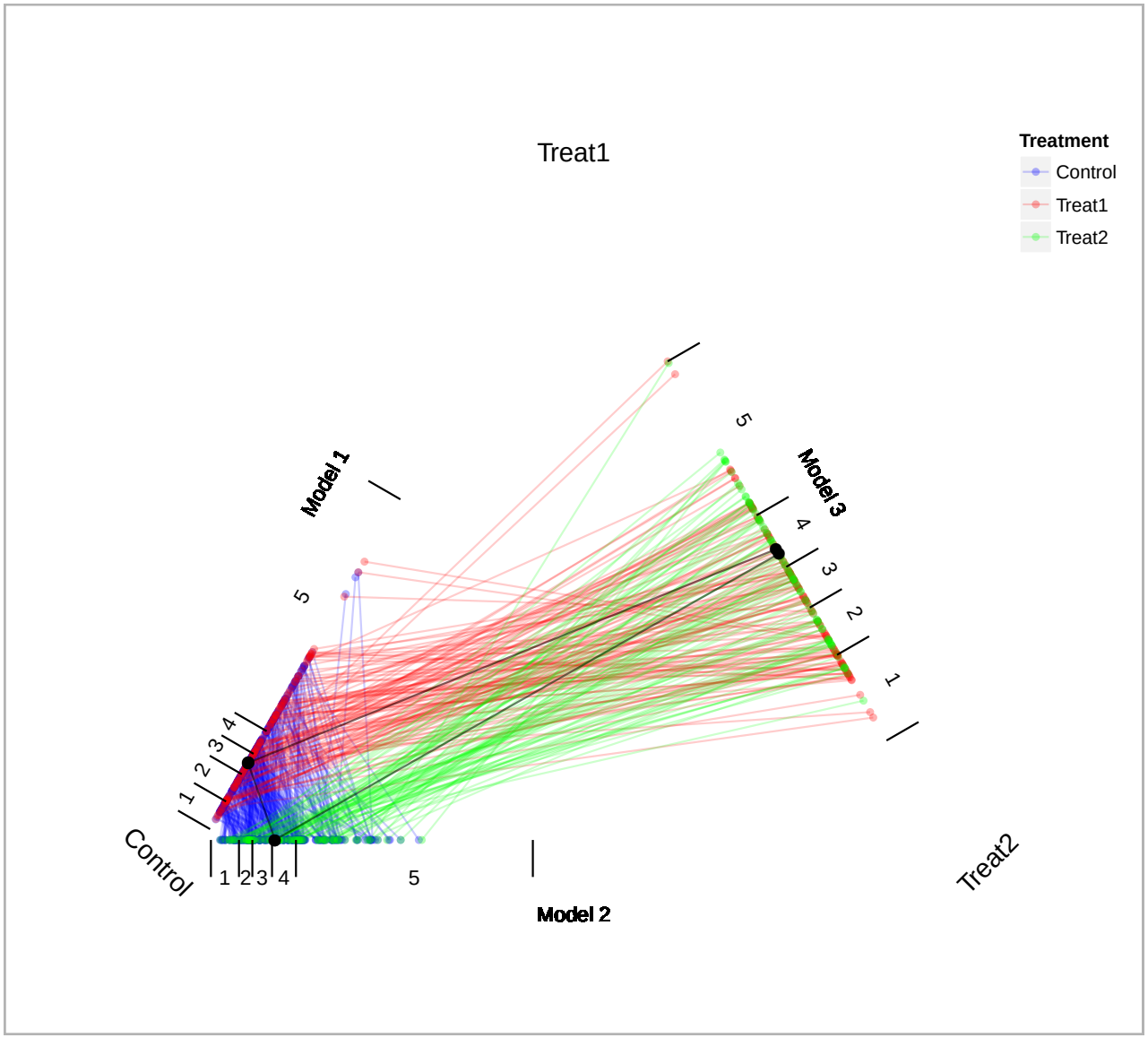
Students can opt to utilize tutoring services to supplement math courses. Of those who used tutoring services, approximately 58% of students used the tutoring service once, whereas the remaining 42% used it more than once. Outcome of interest is course grade.

Military	Active military status.
Income	Income level.
Employment	Employment level.
NativeEnglish	Is English their native language
EdLevelMother	Education level of their mother.
EdLevelFather	Education level of their father.
Ethnicity	American Indian or Alaska Native, Asian, Black or African American, Hispanic, Native Hawaiian or Other Pacific Islander, Two or more races, Unknown, White
Gender	Male, Female
Age	Age at course start.
GPA	Student GPA at the beginning of the course.

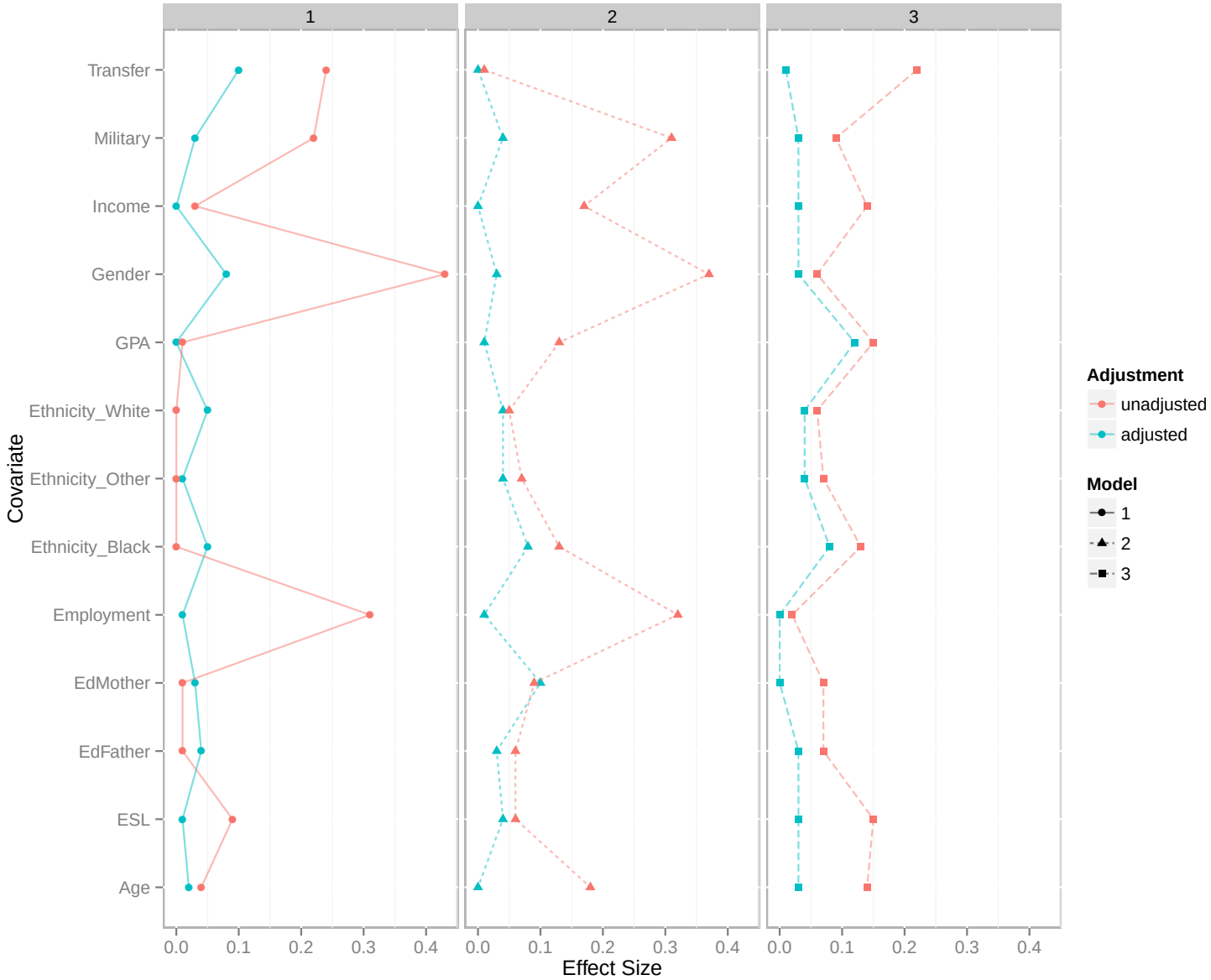
PSA for Non-Binary Treatments

- The `TriMatch` algorithm works as follows:
  1. Estimate three separate propensity score models for each pair of groups (i.e. Control-to-Treat1, Control-to-Treat2, Treat1-to-Treat2).
  2. Determine the matching order. The default is to start with the largest of two treatments, then the other treatment, followed by the control.
  3. For each unit in group 1, find all units from group 2 within a certain threshold (i.e. difference between PSs is within a specified caliper).
  4. For each unit in group 2, find all units from group 3 within a certain threshold.
  5. Calculate the distance (difference) between each unit 3 found and the original unit 1. Eliminate candidates that exceed the caliper.
  6. Calculate a total distance (sum of the three distances) and retain the smallest unique M group 1 units (by default M=2)

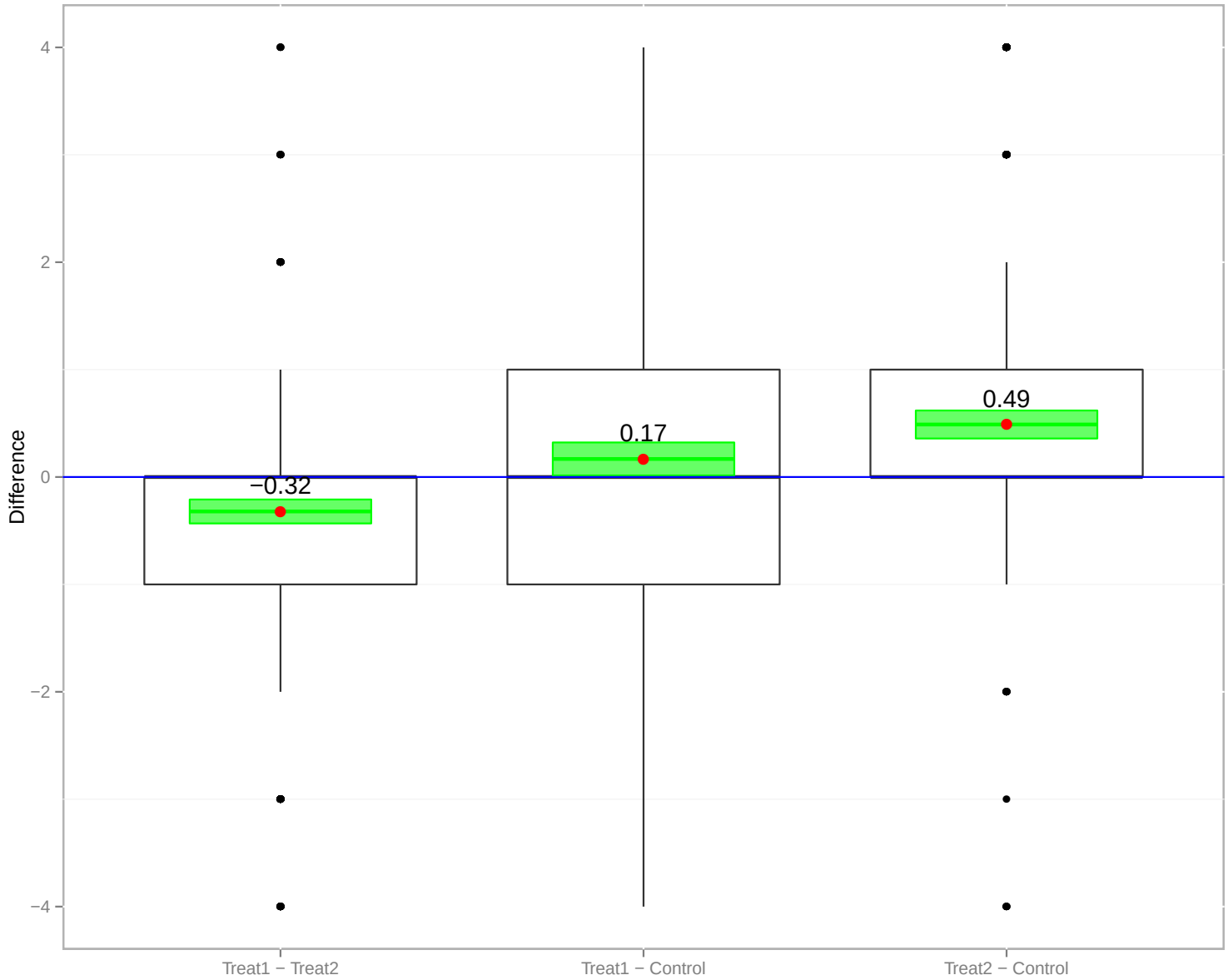
Matching Triplets



Checking Balance



Results


4
2
0
−2
−4
Difference
−0.32
0.17
0.49
Treat1 − Treat2
Treat1 − Control
Treat2 − Control

Thank You

Jason Bryer (jason@bryer.org)
http://www.bryer.org