TriMatch: An R Package for Propensity Score Matching of Non-Binary Treatments

2013 useR! Conference

Jason M. Bryer
Kimberly K. Speerschneider

http://jason.bryer.org/TriMatch
jason@bryer.org
kimkspeer@gmail.com

July 10, 2013

Propensity Score Analysis

- The use of propensity score methods for estimating causal effects has been increasing in the social sciences (Thoemmes & Kim, 2011) and in medical research (Austin, 2008) in the last decade.
- The goal of PSA is to adjust for selection bias in observational studies.
- PSA is conducted in two phases.
  - *Phase I* Calculate propensity scores using observed covariates then match or stratify rows using the propensity scores.
  - *Phase II* Compare the outcome of interest between matched pairs or within strata.
- Two approaches for conducting PSA with non-binary treatments have been suggested:
  - Perform two separate analyses comparing treatment 1 to control and treatment 2 to control.
  - Imai and van Dyk (2004) developed a propensity function that estimates propensity scores for a continuous treatment regime.

Propensity Scores

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

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

Propensity Scores

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

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

- With two treatments, Tr_1 and Tr_2, and a control, C, we estimate:

$$PS_1 = e(x_{T_1C}) = Pr(z = 1|X_{T_1C})$$

$$PS_2 = e(x_{T_2C}) = Pr(z = 1|X_{T_2C})$$

$$PS_3 = e(x_{T_2T_1}) = Pr(z = 1|X_{T_2T_1})$$

- Matching procedures are designed to minimize the distance, or difference, of propensity scores.
- For three groups we wish to minimize the sum of the distance from each of the three propensity scores.
- In Phase II, instead of estimating treatment effects, we will test for differences using repeated measures. Post hoc tests for estimating effects between groups are provided.

TriMatch Package

- The `trips` function estimates three separate propensity score models for each pair of groups (i.e. Control-to-Treat1, Control-to-Treat2, Treat1-to-Treat2).
- The `multibalance.plot` and `balance.plot` functions will create plots to help determine if sufficient balance has been achieved.
- The `trimatch` function finds matched triplets by:
  1. Determine the matching order. The default is to start with the largest of two treatments, then the other treatment, followed by the control.
  2. 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).
  3. For each unit in group 2, find all units from group 3 within a certain threshold.
  4. Calculate the distance (difference) between each unit 3 found and the original unit 1. Eliminate candidates that exceed the caliper.
  5. Calculate a total distance (sum of the three distances).
  6. *Optionally* retain the matched triplets with the smallest distances.
- The `trimatch` function also allows for partial exact matching.
- The `summary`, `boxdiff.plot`, `loess3.plot`, and `parallel.plot` functions are used for estimating differences between matched triplets.

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.

Phase I: Estimating Propensity Scores

```
> require(TriMatch)
> data(tutoring)
> table(tutoring$treat)
```

```
Control  Treat1  Treat2
    918     134      90
```

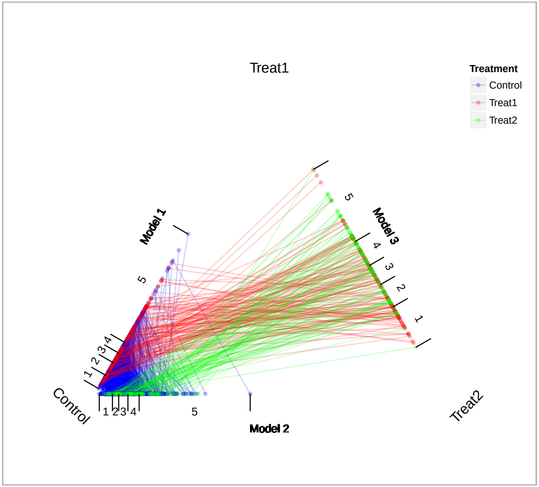
```
> formu <- ~ Gender + Ethnicity + Military + ESL + EdMother +
   EdFather + Age + Employment + Income + Transfer + GPA
```

Note that the dependent variable (i.e. treatment indicator) is not specified in the formula above. This will be set by the `trips` separately for each pair of groupings.

```
> tutoring.tpsa <- trips(tutoring, tutoring$treat, formu)
> head(tutoring.tpsa)
```

Triangle Plot

```
> plot(tutoring.tpsa, draw.segments=TRUE)
```



Finding Matched Triplets

```
> tutoring.matched <- trimatch(tutoring.tpsa,
   method=maximumTreat)
```

The `method` parameter is a function that determines how to reduce the number of matched triplets returned. Specifying `NULL` will not reduce the number of matched triplets and is equivalent to caliper matching in the two-group case. That is, all matches where the distances between propensity scores is less than the caliper will be retained. Other methods include:

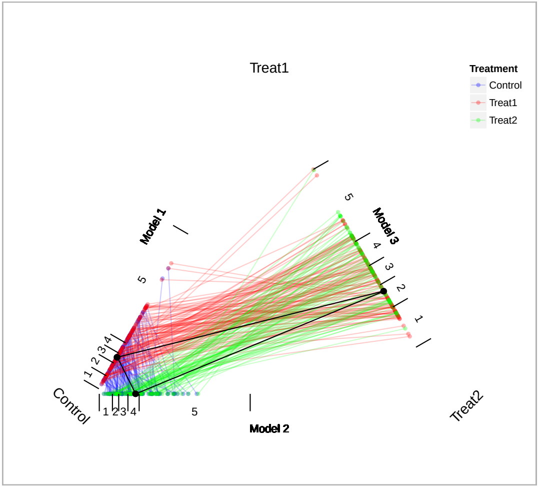
maximumTreat This method will attempt to use all treatments once. However, in the case where Tr_2 units would not be included because they were matched to a Tr_1 unit matched with a different Tr_2 unit with a smaller overall distance, those rows will be retained. Therefore, Tr_1 units may be represented more than once in the returned dataset.

OneToN This method requires two additional parameters `M1` and `M2` that specify how many Tr_1 and Tr_2 units, respectively, can be present in the returned dataset. This is the three group equivalent to one-to-n matching.

Of course, you can implement your own reduction method.

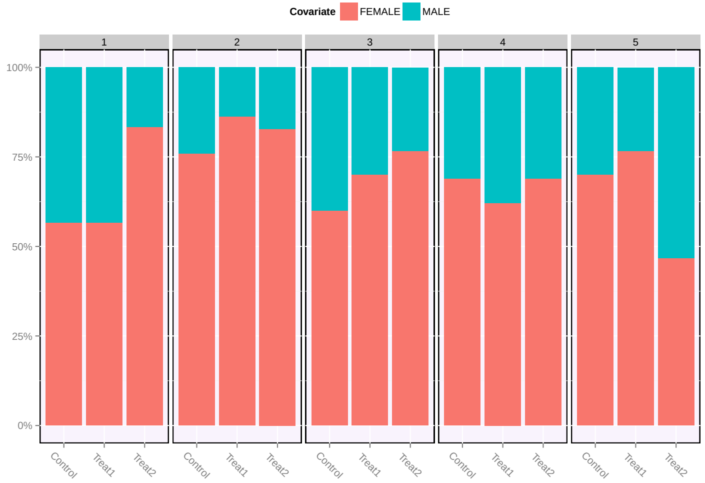
Triangle Plot Revisited

```
> plot(tutoring.matched, rows=c(1), line.alpha=1, draw.segments=TRUE
```



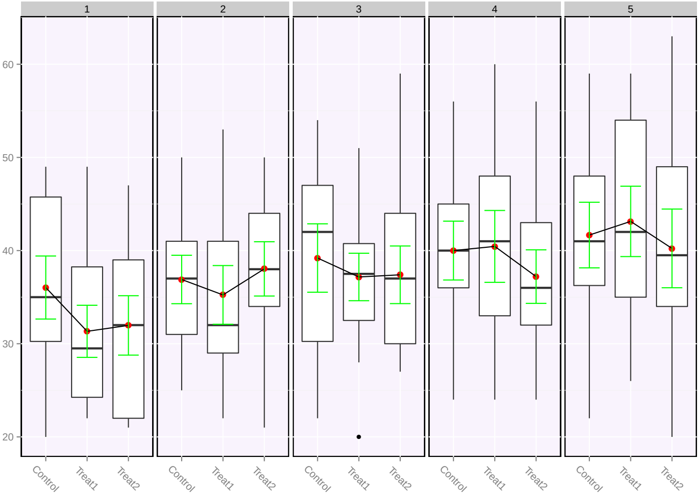
Checking Balance: Categorical Covariate

```
> balance.plot(tutoring.matched, tutoring$Gender)
```



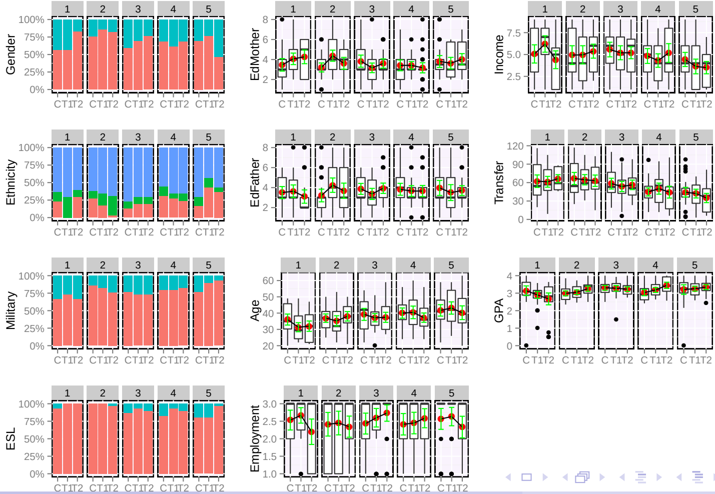
Checking Balance: Continuous Covariate

```
> balance.plot(tutoring.matched, tutoring$Age, label="Age")
```



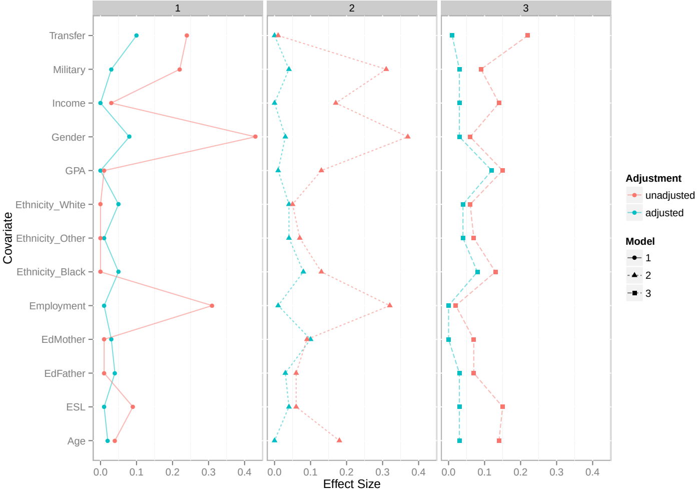
Checking Balance: All Covariates

```
> bplots <- balance.plot(tutoring.matched, tutoring[,all.vars(formu
    legend.position="none", x.axis.labels=c("C","T1","T1"), x.axis.a
> plot(bplots, cols=3, byrow=FALSE)
```



Checking Balance: Covariate Effect Size Plot

```
> multibalance.plot(tutoring.tpsa, grid=TRUE)
```



Phase II

With the matched triplets estimated, we can now compare out outcome of interested, course grade in this example.

```
> sout <- summary(tutoring.matched, tutoring$Grade)
> ls(sout)
```

```
[1] "PercentMatched"       "friedman.test"
[3] "pairwise.wilcox.test" "rmanova"
[5] "t.tests"
```

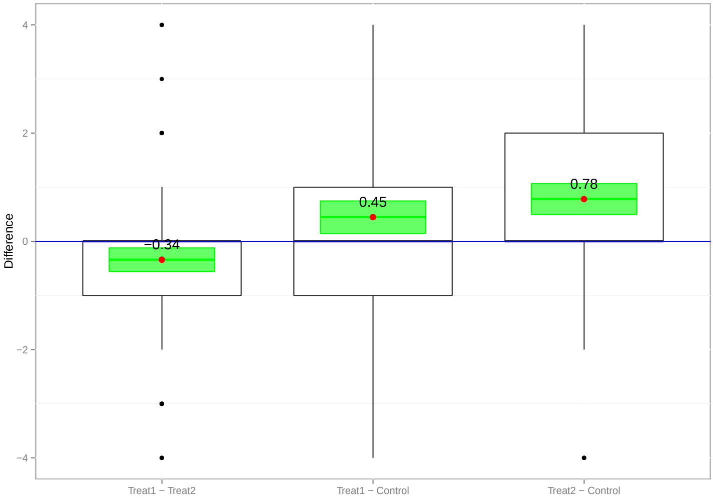
```
> sout$friedman.test
```

```
Friedman rank sum test


data:  Outcome and Treatment and ID
Friedman chi-squared = 17, df = 2, p-value = 0.0001663
```

The summary function performs, at minimum, a Friedman Rank Sum test and a repeated measures ANOVA. If either of those have a p value less than 0.5 (the default but configurable), then a Wilcoxon signed rank test and individual paired t-tests will also be performed.

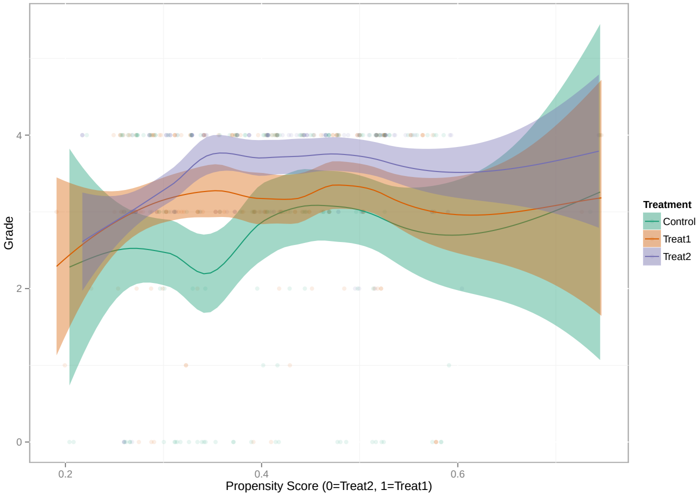
Boxplot of Differences

```
> boxdiff.plot(tutoring.matched, tutoring$Grade)
```



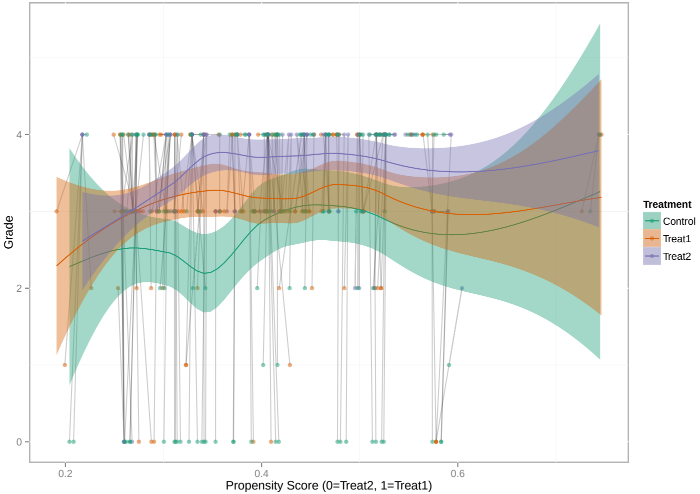
Loess Plot

```
> loess3.plot(tutoring.matched, tutoring$Grade, ylab="Grade")
```



Loess Plot

```
> loess3.plot(tutoring.matched, tutoring$Grade, ylab="Grade",
    points.alpha=.5, plot.connections=TRUE)
```



merge Function

A custom merge function is provided to merge a vector from the original data frame to the matched triplets.

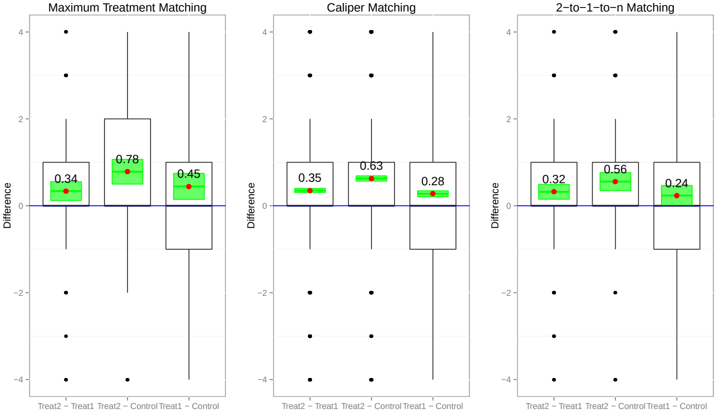
```
> matched.out <- merge(tutoring.matched, tutoring$Grade)
> head(matched.out)
```

```
  Treat1 Treat2 Control   D.m3    D.m2    D.m1 Dtotal Treat1.out
1    368     39     331 0.0071 0.00179 0.01039  0.019          4
2    800   1088    1105 0.0185 0.00074 0.00018  0.019          4
3    286    655     853 0.0169 0.00424 0.00195  0.023          2
4    158    279     365 0.0034 0.00953 0.01071  0.024          4
5    899    209     100 0.0019 0.01363 0.00918  0.025          4
6   1034    791     484 0.0105 0.00854 0.00924  0.028          4
  Treat2.out Control.out
1          4           0
2          4           3
3          4           4
4          4           4
5          3           4
6          4           4
```

Comparing the Three Match Methods

```
> nrow(tutoring.matched)

[1] 148

> tutoring.matched.2to1 <- trimatch(tutoring.tpsa, method=OneToN,
   M1=2, M2=1)
> nrow(tutoring.matched.2to1)

[1] 221

> tutoring.matched.caliper <- trimatch(tutoring.tpsa, method=NULL)
> nrow(tutoring.matched.caliper)

[1] 2981

> s1 <- summary(tutoring.matched, tutoring$Grade)
> s2 <- summary(tutoring.matched.caliper, tutoring$Grade)
> s3 <- summary(tutoring.matched.2to1, tutoring$Grade)
> print("MaxTreat"=s1, "Caliper"=s2, "2-to-1"=s3)

    Method Friedman.chi2 Friedman.p     rmANOVA.F rmANOVA.p
1 MaxTreat            17    1.7e-04 ***        17   1.4e-07 ***
2  Caliper           182    3.0e-40 ***       222   8.3e-94 ***
3   2-to-1            17    2.4e-04 ***        14   8.0e-07 ***
```

Comparing the Three Match Methods



Example II: National Medical Expenditure Study

- Dataset originally used by Johnson et al. (2003) then later used by Imai and van Dyk (2004) to estimate the effects of smoking on medical expenditures.
- Imai and van Dyk developed pscore function to provide propensity scores of a continuous variable. Their treatment variable, *packyear*, is defined as:

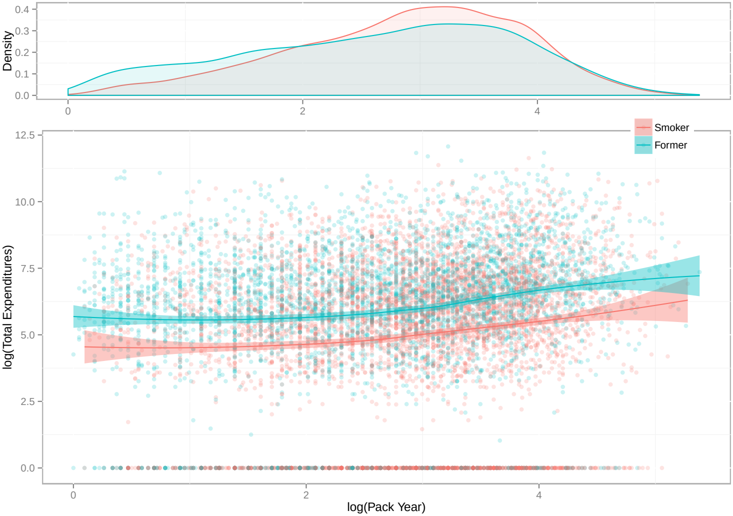
$$packyear = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked}$$

- We will estimate two separate propensity score models using different treatment regimes:
  - Smoking status (i.e. former smoker, current smoker, never smoked)
  - Lifetime smoking (i.e. heavy smoker, moderate smoker, never smoked)
- Covariates available: age when surveyed, gender, ethnicity, seat belt use, education level, marital status, census region, poverty status.

Example II: National Medical Expenditure Study

```
> table(nmes$smoke, useNA="ifany")

 Never Smoker Former
  9802   5472   4078

> (medPY <- median(nmes[nmes$smoke != "Never",]$packyears))

[1] 17

> nmes$smoke2 <- ifelse(nmes$smoke == "Never", "Never",
      ifelse(nmes$packyears > medPY, "Heavy", "Moderate"))
> table(nmes$smoke2, useNA="ifany")

   Heavy Moderate    Never
    4770     4780     9802

> table(nmes$smoke, nmes$smoke2, useNA="ifany")

         Heavy Moderate Never
  Never      0        0  9802
  Smoker  2901     2571     0
  Former  1869     2209     0
```

Relationship Between Pack Year and Total Expenditures by Current Smoking Status

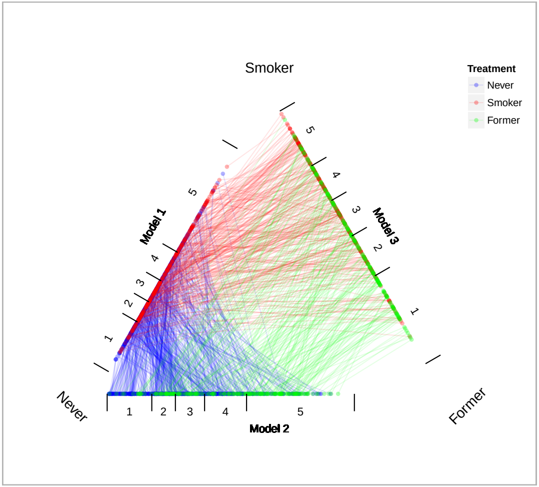


Estimate Propensity Scores

```
> formu <- ~ LASTAGE + MALE + RACE3 + beltuse + educate +
   marital + SREGION + POVSTALB
> tpsa.smoke <- trips(nmes, nmes$smoke, formu)
> tpsa.packyears <- trips(nmes, nmes$smoke2, formu)
```

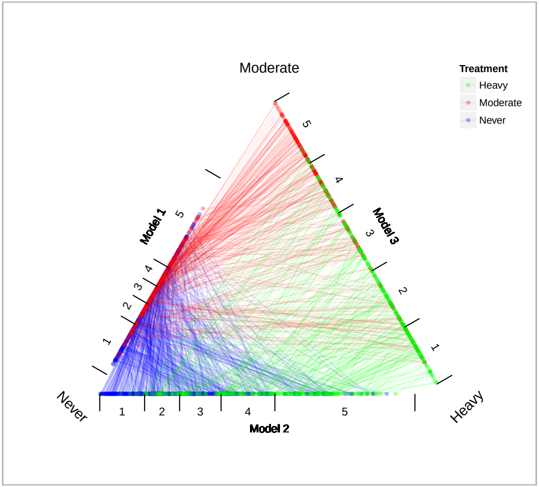
Triangle Plot: Smoking Status

```
> plot(tpsa.smoke, sample=c(.05), edge.alpha=.1)
```



Triangle Plot: Pack Years

```
> plot(tpsa.packyears, sample=c(.05), edge.alpha=.1)
```



Find Matched Triplets

We will categorize the age of being surveyed so that we can use to perform partial exact matching.

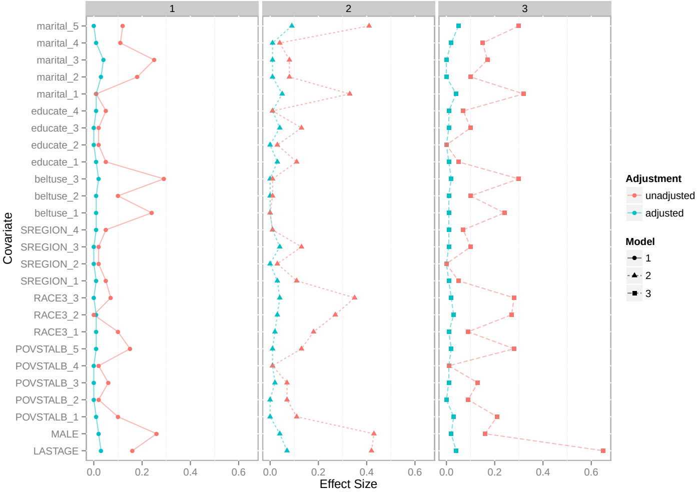
```
> nmes$LastAge5 <- cut(nmes$LASTAGE,
   breaks=quantile(nmes$LASTAGE, probs=seq(0,1,1/5)),
   include.lowest=TRUE, orderd_result=TRUE)
```

Find matched triplets.

```
> tmatch.smoke <- trimatch(tpsa.smoke,
   exact=nmes[,c("LastAge5","MALE","RACE3")], nmatch=10)
> tmatch.packyears <- trimatch(tpsa.packyears,
   exact=nmes[,c("LastAge5","MALE","RACE3")], nmatch=10)
```

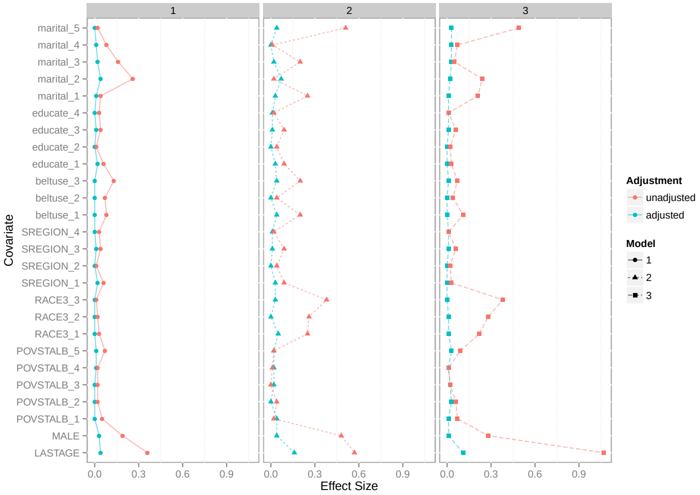
Balance Plot: Smoking Status

```
> multibalance.plot(tpsa.smoke)
```



Balance Plot: Pack Years

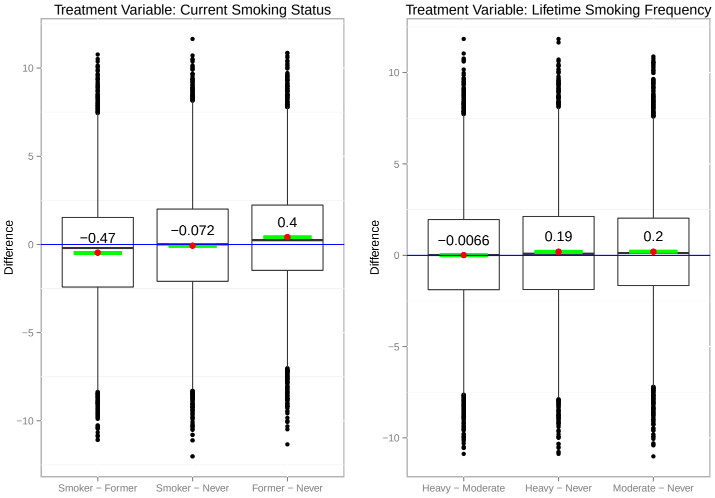
```
> multibalance.plot(tpsa.packyears)
```



Results

```
> sum.smoke <- summary(tmatch.smoke, nmes$LogTotalExp,
   ordering=c("Smoker","Former","Never"))
> sum.packyears <- summary(tmatch.packyears, nmes$LogTotalExp,
   ordering=c("Heavy","Moderate","Never"))
> print("Current Smoking Status"=sum.smoke,
   "Smoking Frequency"=sum.packyears)

                  Method Friedman.chi2 Friedman.p     rmANOVA.F
1 Current Smoking Status            98    5.5e-22 ***        74
2      Smoking Frequency            37    7.6e-09 ***        14
  rmANOVA.p
1   9.1e-33 ***
2   5.9e-07 ***
```

Boxplot of Differences



More Information

- Package vignette `vignette("TriMatch")`
- Tutoring Demo (`demo("tutoring")`)
- National Medical Expenditure Study Demo (`demo("nmes")`)

Thank You

Jason Bryer (jason@bryer.org)
Kimberly Speerschneider (kimkspeer@gmail.com)

`http://jason.bryer.org/TriMatch`
`http://github.com/jbryer/TriMatch`