Big_match - Testing and Demo

Rachael 'Rocky' Aikens, Voight Lab

August 24, 2018

This R markdown document is used for testing and demoing the current functionality of bigmatch. We'll use the sample data from the MatchIt package for basic testing.

```r
library(MatchIt)
library(ggplot2)
library(ggpubr)
library(dplyr)

data("lalonde")
source('big_match.R')
source('class_functions.R')

# adding a binary outcome
lalonde$outcome <- lalonde$re78 > 15000
lalonde$re78 <- NULL
```

Stratify

Manual Stratify

Testing errors and warnings

This call should return an error because “educ” is a continuous variable.

```
# manual stratification with a continuous variable
m.strat <- manual_stratify(lalonde, "treat", "outcome",
                           covariates = c("black", "hispan", "educ", "nodegree"))

# Error in warn_if_continuous(data[, covariates[i]], covariates[i]) :
#   There are 19 distinct values for educ. Is it continuous?
```

Testing Functionality with Valid Inputs

This call should return six strata (since black and hispanic seem to be mutually exclusive categories in this dataset).

```
# allowable manual stratification
m.strat <- manual_stratify(lalonde, "treat", "outcome",
                           covariates = c("black", "hispan", "nodegree"))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
m.strat$strata_table
```

```
## # A tibble: 6 x 5
## # Groups:   black, hispan [?]
##   black hispan nodegree stratum  size
##   <int>  <int>    <int>   <dbl> <int>
## 1     0      0        0       1   136
## 2     0      0        1       2   163
## 3     0      1        0       3    17
## 4     0      1        1       4    55
## 5     1      0        0       5    74
## 6     1      0        1       6   169
```

Auto Stratify

Testing Errors and Warnings

First, testing error handling. These should fail and/or give warnings.

```
# auto stratification with missing arguements
a.strat <- auto_stratify(lalonde, "treat", "outcome")

# Error in auto_stratify(lalonde, "treat", "outcome") :
#   At least one of covariates and prog_scores should be specified.
```

```
# auto stratification with covariates and prog scores specified, and prog_scores invalid
a.strat <- auto_stratify(lalonde, "treat", "outcome", c("age", "educ"), prog_scores = 1:4)

# covariates and prog_scores are both specified. Using prog_scores; ignoring covariates.
# Error in auto_stratify(lalonde, "treat", "outcome", c("age", "educ"), :
#   prog_scores must be the same length as the data
```

Testing Functionality with Valid Inputs

These should give valid results.

```
# auto stratification with pre-specified prognostic score
myprogscore <- runif(n = dim(lalonde)[1])
a.strat1 <- auto_stratify(data = lalonde, "treat", "outcome",
                          prog_scores = myprogscore, size = 100)

# auto stratification
a.strat2 <- auto_stratify(lalonde, "treat", "outcome",
                          covariates = c("age", "educ", "hispan", "nodegree", "black"),
                          size = 100)

# auto stratification with a non-continuous prognostic score
a.strat3 <- auto_stratify(lalonde, "treat", "outcome",
                          covariates = c("hispan", "nodegree", "black"), size = 100 )
```

Diagnostics

Most of this is implemented with the generic functions `print`, `summary`, and `plot`.

Print

```r
print(m.strat)
```

```
## manual_strata object from package big_match.
##
## Strata Definition Table:
## # A tibble: 6 x 5
## # Groups:   black, hispan [?]
##   black hispan nodegree stratum  size
##   <int>  <int>    <int>   <dbl> <int>
## 1     0      0        0       1   136
## 2     0      0        1       2   163
## 3     0      1        0       3    17
## 4     0      1        1       4    55
## 5     1      0        0       5    74
## 6     1      0        1       6   169
##
## Strata summary:
##   stratum treat_mean age_mean educ_mean black_mean hispan_mean
## 1       1 0.06617647 29.75735 12.786765          0           0
## 2       2 0.05521472 28.01840  8.773006          0           0
## 3       3 0.11764706 25.52941 12.235294          0           1
## 4       4 0.16363636 26.03636  8.018182          0           1
## 5       5 0.58108108 26.12162 12.567568          1           0
## 6       6 0.66863905 25.96450  9.213018          1           0
##   married_mean nodegree_mean re74_mean re75_mean stratum_size
## 1    0.5441176             0  7836.767  2858.656          136
## 2    0.5828221             1  4945.338  2228.511          163
## 3    0.4705882             0  4946.764  2471.928           17
## 4    0.4363636             1  4272.401  2824.681           55
## 5    0.2297297             0  3853.371  1809.523           74
## 6    0.2189349             1  1906.608  1528.063          169
```

```r
print(a.strat1)
```

```
## auto_strata object from package big_match.
##
## Prognostic scores prespecified.
##
## Strata summary:
##   stratum treat_mean age_mean educ_mean black_mean hispan_mean
## 1       1  0.3181818 25.75000  10.52273  0.4090909  0.10227273
## 2       2  0.3068182 28.56818  10.03409  0.3977273  0.14772727
## 3       3  0.3068182 27.02273  10.05682  0.3863636  0.14772727
## 4       4  0.2873563 26.85057  10.57471  0.3333333  0.09195402
## 5       5  0.2386364 28.59091  10.20455  0.4318182  0.14772727
## 6       6  0.3295455 27.28409  10.32955  0.3750000  0.07954545
## 7       7  0.3218391 27.47126  10.16092  0.4367816  0.10344828
##   married_mean nodegree_mean re74_mean re75_mean stratum_size
## 1    0.3068182     0.6136364  4369.674  2252.626           88
## 2    0.4204545     0.6477273  5219.705  2405.886           88
## 3    0.3863636     0.6477273  4437.516  2321.808           88
## 4    0.4712644     0.5517241  4961.191  1756.087           87
## 5    0.4090909     0.6931818  4996.131  2548.535           88
## 6    0.4772727     0.6250000  4369.167  1951.788           88
## 7    0.4367816     0.6321839  3542.496  2051.447           87
```

```
print(a.strat2)
```

```
## auto_strata object from package big_match.
## 
## Prognostic Score Model:
## 
## Call:  glm(formula = formula(formula_str), family = "binomial", data = data0)
## 
## Coefficients:
## (Intercept)          age         educ       hispan     nodegree  
##    -5.00887      0.04639      0.19183      0.09823      0.14705  
##       black  
##    -0.53706  
## 
## Degrees of Freedom: 428 Total (i.e. Null);  423 Residual
## Null Deviance:       388.2 
## Residual Deviance: 361.1     AIC: 373.1
## 
## Strata summary:
##   stratum treat_mean age_mean educ_mean black_mean hispan_mean
## 1       1 0.45454545 20.44318  7.590909 0.78409091  0.06818182
## 2       2 0.42045455 22.32955  9.284091 0.61363636  0.13636364
## 3       3 0.44318182 24.39773  9.829545 0.54545455  0.10227273
## 4       4 0.29885057 24.72414 10.517241 0.35632184  0.12643678
## 5       5 0.25000000 28.26136 10.647727 0.22727273  0.14772727
## 6       6 0.18181818 30.69318 11.602273 0.14772727  0.15909091
## 7       7 0.05747126 40.81609 12.436782 0.09195402  0.08045977
##   married_mean nodegree_mean re74_mean re75_mean stratum_size
## 1    0.1363636     1.0000000  936.6031  1064.158           88
## 2    0.2840909     0.8181818 2471.3263  1572.605           88
## 3    0.3636364     0.7500000 2288.4294  1594.332           88
## 4    0.3678161     0.7011494 3951.9516  2573.689           87
## 5    0.5454545     0.5340909 5674.6884  2803.897           88
## 6    0.5795455     0.3181818 7637.6392  3221.502           88
## 7    0.6321839     0.2873563 8985.6253  2472.065           87
```

Summary

```
# TODO: implement summary methods
summary(m.strat)
```

```
##              Length Class      Mode
## data         11     tbl_df     list
## strata_table  5     grouped_df list
```

```
summary(a.strat1)
```

```
##             Length Class      Mode
## data         11    data.frame list
## prog_scores 614    -none-     numeric
## prog_model    0    -none-     NULL
```

```
summary(a.strat2)
```

```
##             Length Class      Mode
## data         11    data.frame list
## prog_scores 614    -none-     numeric
## prog_model   30    glm        list
```

Plot

There are three types of plots: "scatter", "residual", and "hist". Any other plot options for a strata object will throw an error.

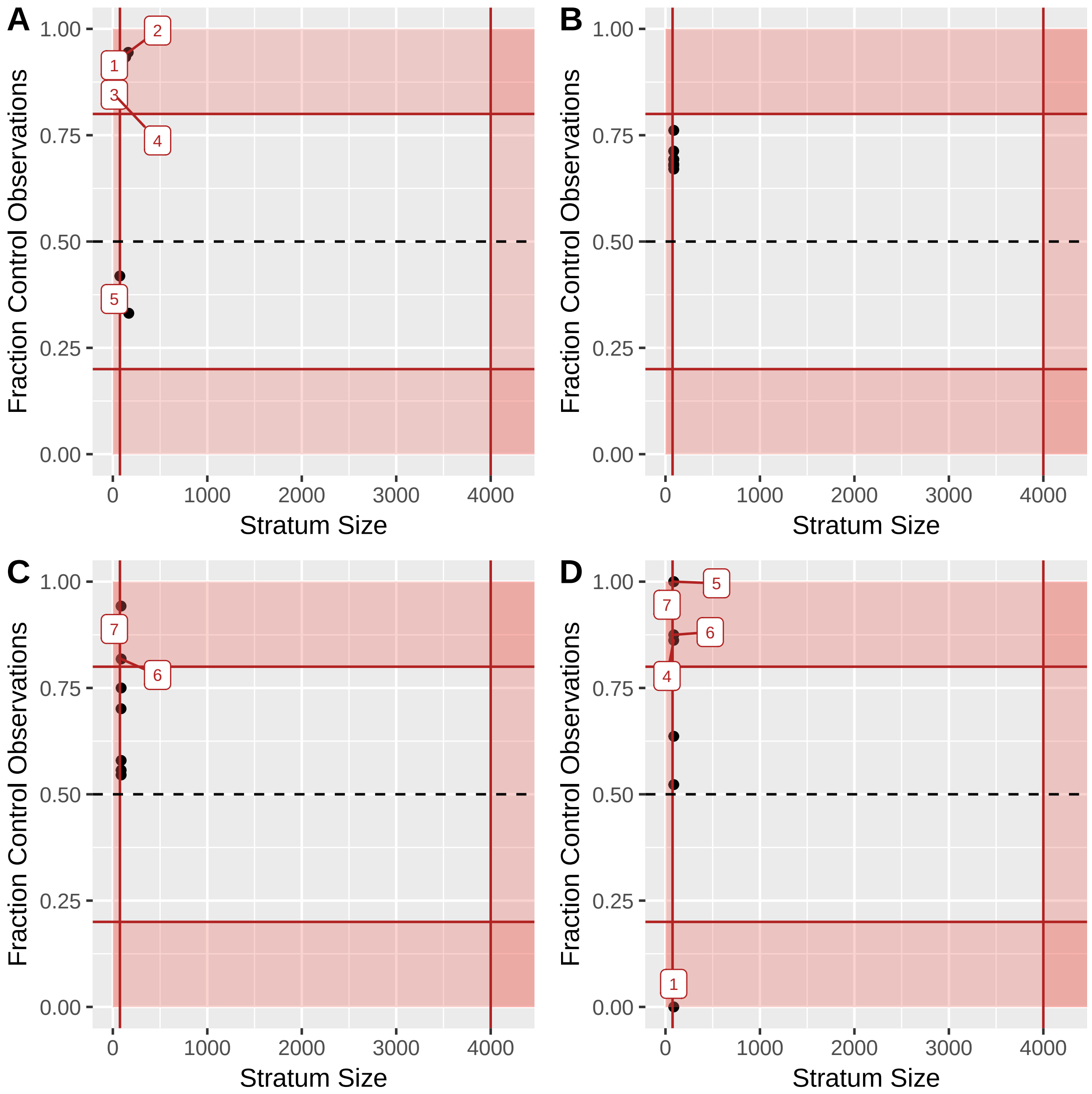
```
plot(m.strat, type = "QQ")
# Error in plot.strata(m.strat, type = "QQ") :
#   Not a recognized plot type.
```

Auto and Manual Stratify : Size-Balance Scatterplot

Below are the basic size × control fraction scatterplots for (A) manual stratification by black, hispan and nodegree, (B) auto-stratification with a uniform random prognostic score, (C) auto-stratification with a prognostic score that is relatively continuous, and (D) auto-stratification with a prognostic score that is discontinuous (built solely from discrete variables with few distinct values). As you can see, this sample data contains a relatively small number of examples to begin with, so most strata are quite small.

```
a <- plot(m.strat) # equivalently: plot(m.strat, type = "scatter")
b <- plot(a.strat1)
c <- plot(a.strat2)
d <- plot(a.strat3)

ggarrange(a, b, c, d, ncol = 2, nrow = 2,
          labels = "AUTO")
```

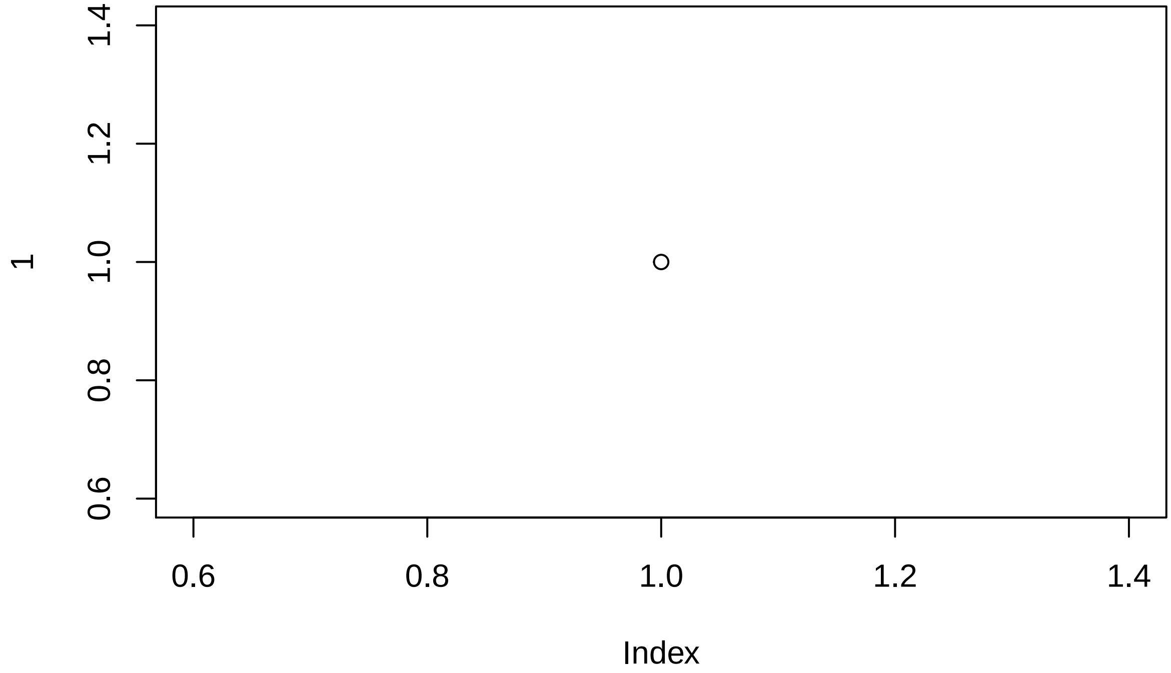


Auto Stratify: Prognostic Score Residual Plot

This function is only meant for auto-stratified data. Running it on a `manual_strata` object will throw an error.

```
plot(m.strat, type = "residual")
# Error in plot.strata(m.strat, type = "residual") :
#   Prognostic score residual plots are only valid for auto-stratified data.
```

```
# TODO: implement this plot
plot(a.strat1, type = "residual")
plot(a.strat2, type = "residual")
```



Auto Stratify: Prognostic Score Histograms

This function is only meant for auto-stratified data. Running it on a `manual_strata` object will throw an error.

```
plot(m.strat, type = "hist")

# Error in plot.strata(m.strat, type = "hist"):
#  Prognostic score histograms are only valid for auto-stratified data.
```

```
# uniformly generated prognostic score. Nicely continuous from 0 to 1
a <- plot(a.strat1, type = "hist")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

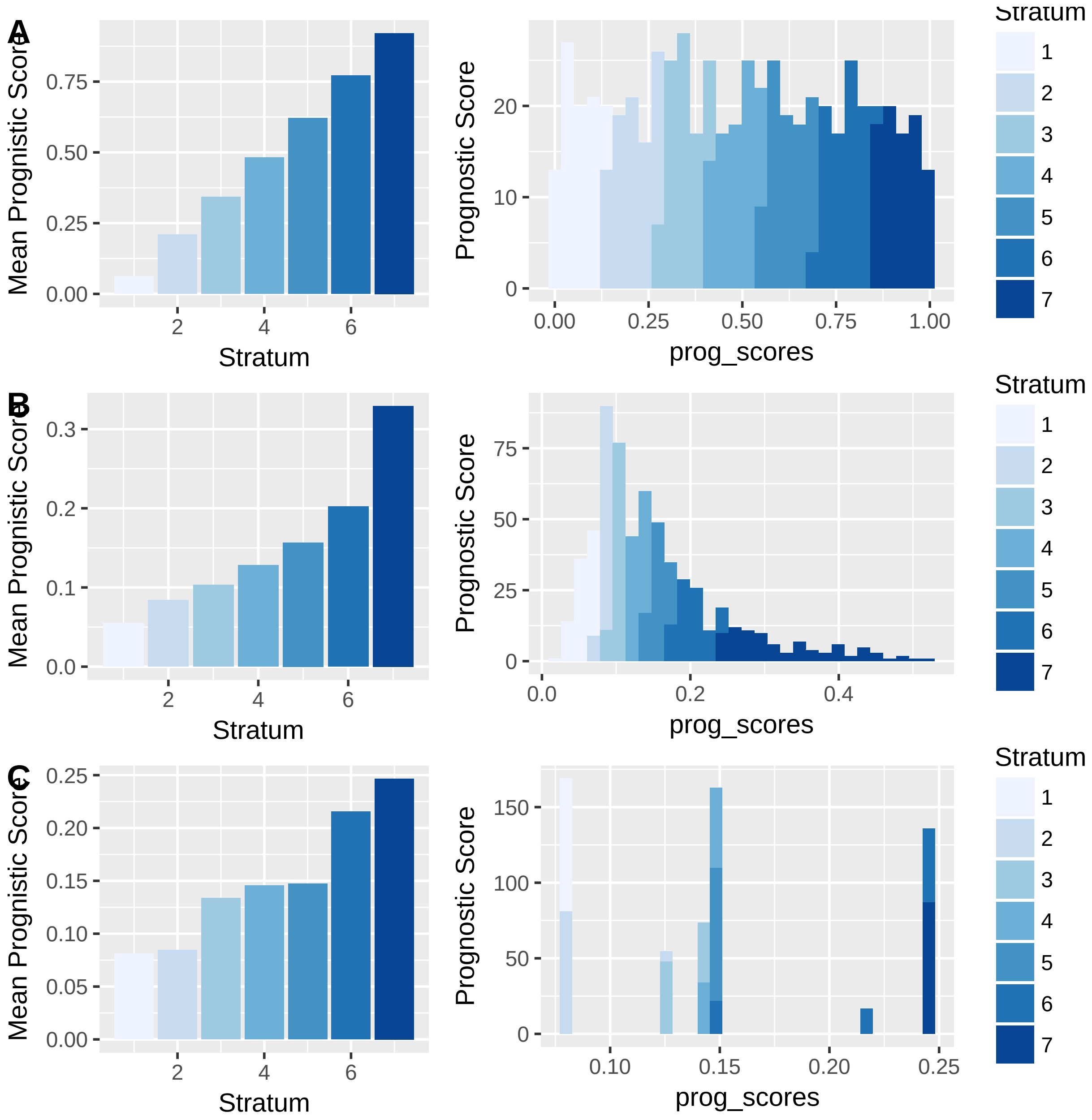
```
# prognostic score generated from some continuous and some distcrete variables.
# Fairly continuous
b <- plot(a.strat2, type = "hist")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
# prognostic score generated from only a few discrete variables.
# Since prog_score only takes on a few different values,
# strata quantiles are less evenly distributed from 0 to 1
c <- plot(a.strat3, type = "hist")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```r
ggarrange(a, b, c, ncol = 1, nrow = 3, labels = "AUTO")
```



Matching