# 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.

```
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 - should fail
m.strat <- manual_stratify(lalonde, "treat", "outcome",
                           covariates = c("black", "hispan", "educ", "nodegree"))
```

### 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
# throws error
a.strat <- auto_stratify(lalonde, "treat", "outcome")
```

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

### Testing Functionality with Valid Inputs

These should give valid results.

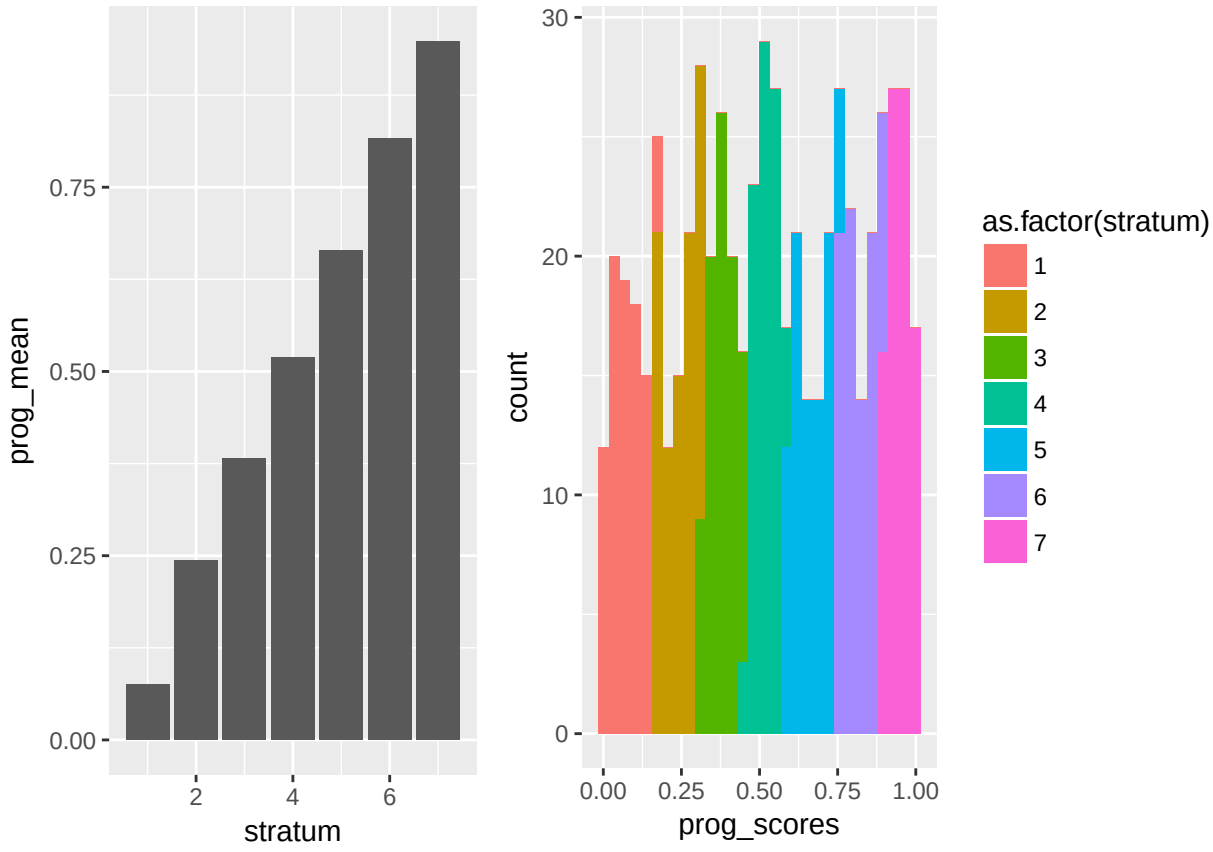
```
# auto stratification with pre-specified prognostic score
myprogscore <- runif(n = dim(lalonde)[1])
a.strat1 <- auto_stratify(data = lalonde, "treat", "outcome",
                          prog_scores = myprogscore, size = 100)

# 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 )
```

Basic visualization of our prognostic score strata, with outcome

Below are plots of prognostic score by strata. The plot on the left shows the average prognostic score by stratum, the plot on the right gives a histogram of the prognostic scores of all samples, colored by stratum.

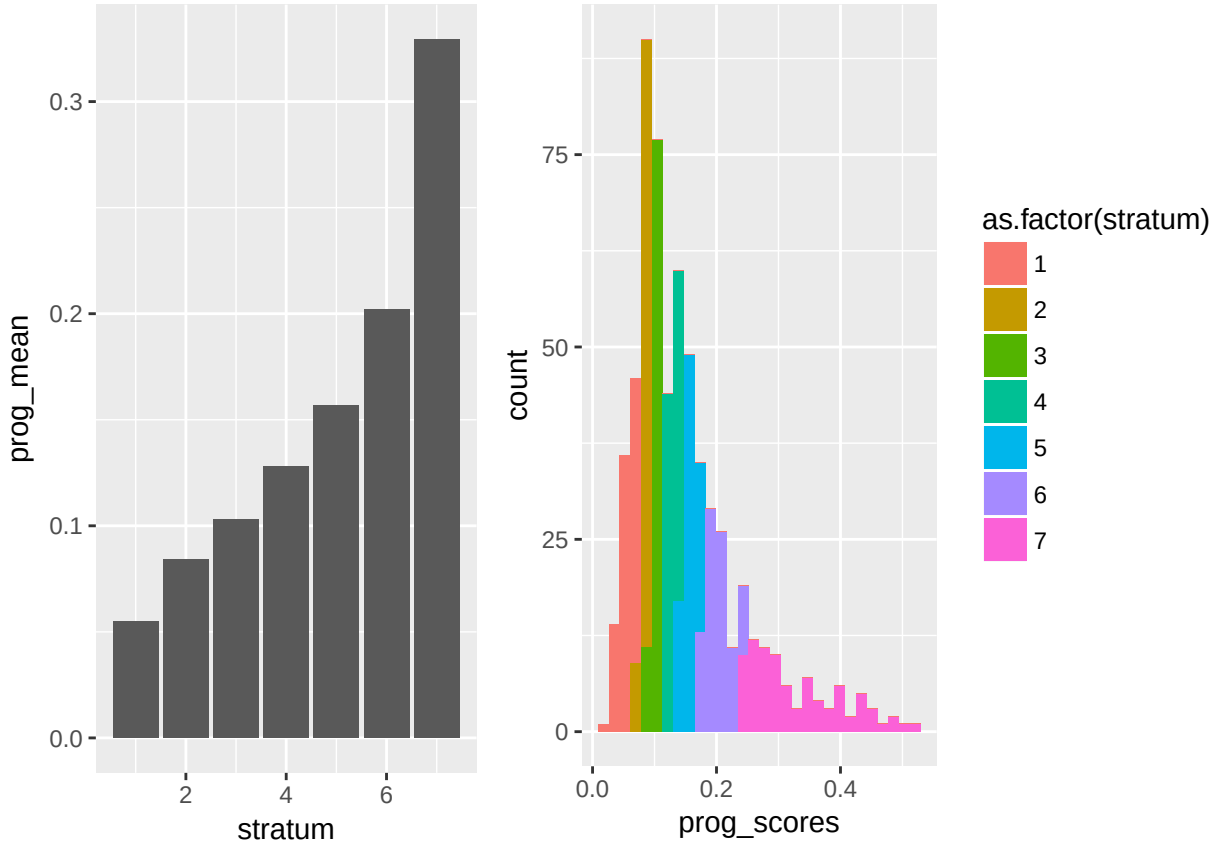
```
# uniformly generated prognostic score. Nicely continuous from 0 to 1
basic_viz(a.strat1)
```

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



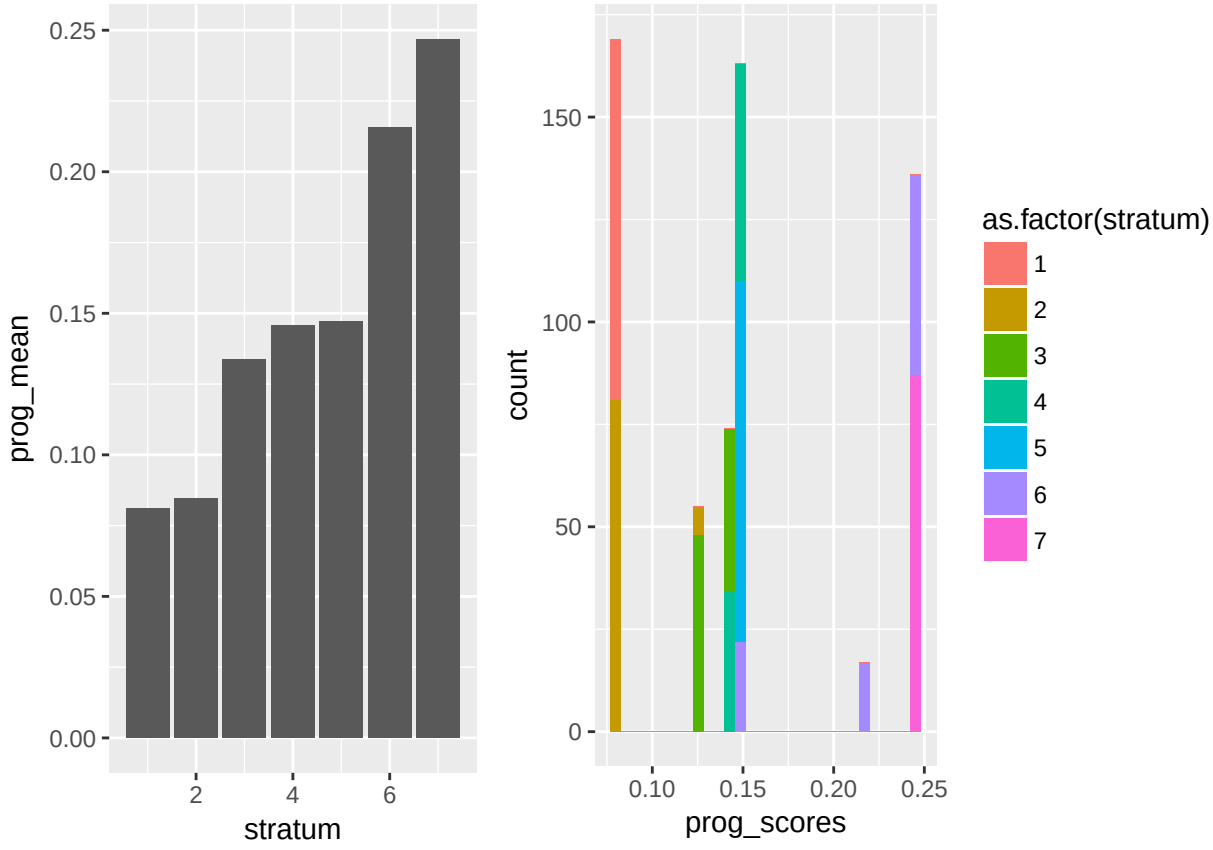

```
# prognostic score generated from some continuous and some distcrete variables.
# Fairly continuous
basic_viz(a.strat2)
```

```
## `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
basic_viz(a.strat3)
```

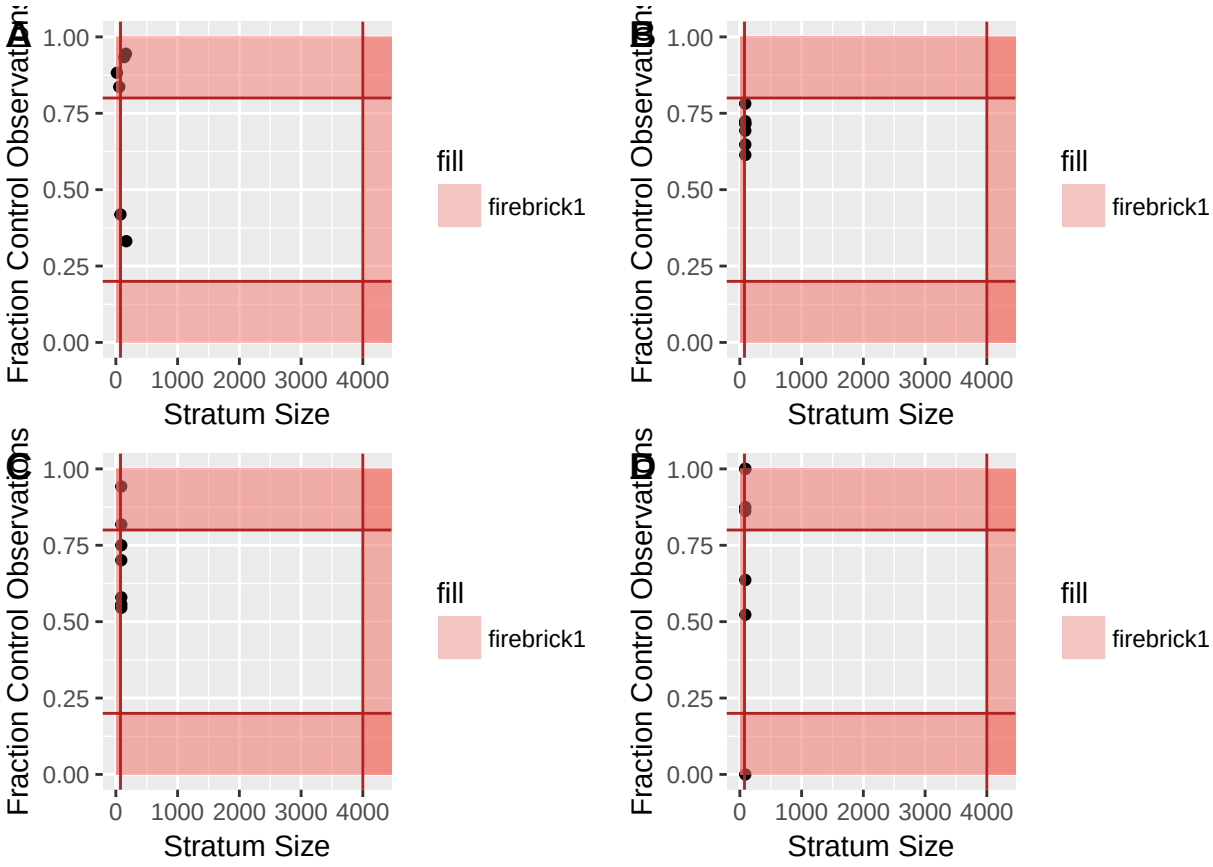
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




## Plots

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)
b <- plot(a.strat1)
c <- plot(a.strat2)
d <- plot(a.strat3)

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


A
B
C
D
Fraction Control Observations
Stratum Size
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
fill
firebrick1