SS154 Final Project

Joyce Gu, Nguyen Nguyen, and Minh Nguyen

04/23/2021

Import and format

```
# Import the data
data = read.csv('data.csv')
# Create treated variable (big city if pop > 250000)
treated <- ifelse (data$population_in_2010 >= 250000, 1, 0)
# Create a new column
data$treated = treated
# Tier is categorical variable
data$tier.f = factor(data$tier)
head(data)
```

```
##   X                      name starting_median_salary   zip tier
sat_avg_2013
## 1 0       Stanford University                  70400 94305    1
1475
## 2 1       Harvey Mudd College                  71800 91711    2
1494
## 3 2        Occidental College                  51900 90041    2
1300
## 4 3            Pomona College                  48600 91711    2
1460
## 5 4 Humboldt State University                  42600 95521    5
985
## 6 5        Seattle University                  48300 98122    6
1170
##   asian_or_pacific_share black_share hisp_share alien_share        city
## 1             0.20999239  0.07329445 0.08648238 0.040451434    Stanford
## 2             0.21059972  0.00418410 0.04044630 0.034867503   Claremont
## 3             0.18085732  0.06165590 0.14034058 0.041691132 Los Angeles
## 4             0.13722999  0.03748412 0.07623889 0.019695044   Claremont
## 5             0.03107126  0.02411501 0.08239295 0.004482918      Arcata
## 6             0.20133212  0.04874357 0.05146836 0.101422950     Seattle
##   population_in_2010 par_median treated tier.f
## 1              13809     172600       0      1
## 2              34926     139800       0      2
## 3            9818605     122400       1      2
## 4              34926     161600       0      2
## 5              17231      96000       0      5
## 6             608660     105700       1      6
```

Matching by matchit

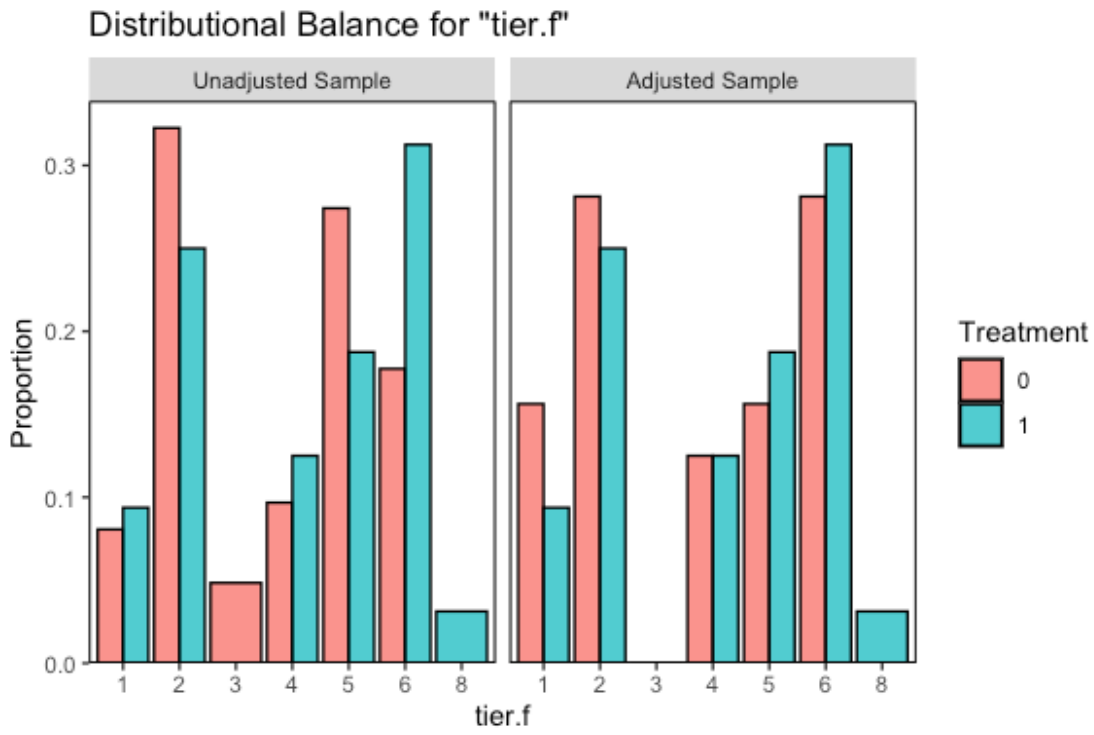
```
# Use matchit to match based
# Use nearest method to match
matched_output <- matchit(treated ~ tier.f + sat_avg_2013 +
asian_or_pacific_share + black_share + hisp_share + alien_share + par_median,
                          data = data, method="nearest", ratio=1)

# Summary of matching process
summary(matched_output)

##
## Call:
## matchit(formula = treated ~ tier.f + sat_avg_2013 + asian_or_pacific_share
+
##     black_share + hisp_share + alien_share + par_median, data = data,
##     method = "nearest", ratio = 1)
##
## Summary of Balance for All Data:
##                          Means Treated Means Control Std. Mean Diff. Var.
Ratio
## distance                        0.4848        0.2659          0.8644
2.4159
## tier.f1                         0.0938        0.0806          0.0450
.
## tier.f2                         0.2500        0.3226         -0.1676
.
## tier.f3                         0.0000        0.0484         -0.2753
.
## tier.f4                         0.1250        0.0968          0.0853
.
## tier.f5                         0.1875        0.2742         -0.2221
.
## tier.f6                         0.3125        0.1774          0.2914
.
## tier.f8                         0.0312        0.0000          0.1796
.
## sat_avg_2013                 1227.6094     1219.8961          0.0458
0.9171
## asian_or_pacific_share          0.0787        0.0585          0.3482
0.9807
## black_share                     0.1063        0.0531          0.3493
14.3833
## hisp_share                      0.0504        0.0460          0.1320
0.3494
## alien_share                     0.0511        0.0306          0.4234
4.7839
## par_median                 123418.7500   126779.0323         -0.0785
1.1299
##                          eCDF Mean eCDF Max
## distance                    0.2576   0.4133
```

```
## tier.f1                    0.0131   0.0131
## tier.f2                    0.0726   0.0726
## tier.f3                    0.0484   0.0484
## tier.f4                    0.0282   0.0282
## tier.f5                    0.0867   0.0867
## tier.f6                    0.1351   0.1351
## tier.f8                    0.0312   0.0312
## sat_avg_2013               0.0418   0.1310
## asian_or_pacific_share     0.1210   0.2369
## black_share                0.1789   0.3609
## hisp_share                 0.0767   0.2117
## alien_share                0.1230   0.2319
## par_median                 0.0571   0.2016
##
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var.
Ratio
## distance                       0.4848        0.3789          0.4183
3.4488
## tier.f1                        0.0938        0.1562         -0.2144
.
## tier.f2                        0.2500        0.2812         -0.0722
.
## tier.f3                        0.0000        0.0000          0.0000
.
## tier.f4                        0.1250        0.1250          0.0000
.
## tier.f5                        0.1875        0.1562          0.0801
.
## tier.f6                        0.3125        0.2812          0.0674
.
## tier.f8                        0.0312        0.0000          0.1796
.
## sat_avg_2013                1227.6094     1251.0175         -0.1390
0.7752
## asian_or_pacific_share         0.0787        0.0766          0.0356
0.6846
## black_share                    0.1063        0.0711          0.2314
10.6436
## hisp_share                     0.0504        0.0451          0.1598
1.5135
## alien_share                    0.0511        0.0381          0.2687
4.3706
## par_median                123418.7500   131321.8750         -0.1846
1.1154
##                         eCDF Mean eCDF Max Std. Pair Dist.
## distance                   0.0745   0.3125          0.4201
## tier.f1                    0.0625   0.0625          0.6433
## tier.f2                    0.0312   0.0312          1.0825
```

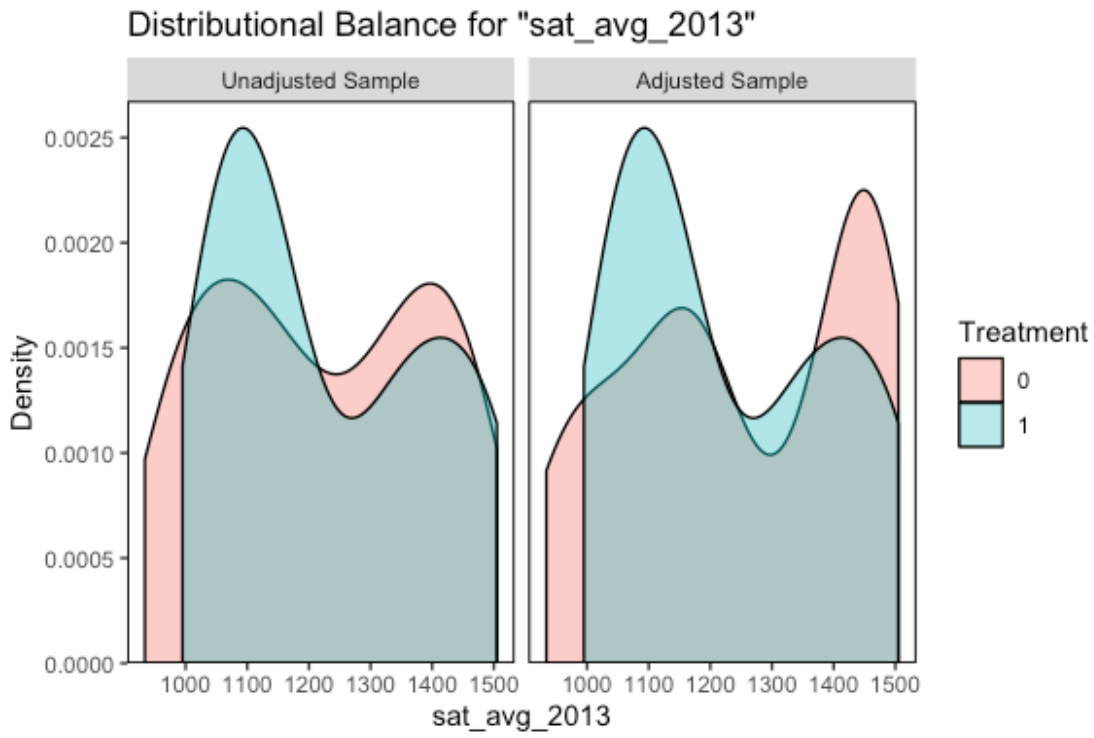
```
## tier.f3                    0.0000   0.0000          0.0000
## tier.f4                    0.0000   0.0000          0.2500
## tier.f5                    0.0312   0.0312          0.8807
## tier.f6                    0.0312   0.0312          1.0113
## tier.f8                    0.0312   0.0312          0.1796
## sat_avg_2013               0.0808   0.1562          1.3684
## asian_or_pacific_share     0.0615   0.1562          1.2616
## black_share                0.0489   0.1250          0.4781
## hisp_share                 0.0665   0.2188          0.9677
## alien_share                0.0432   0.1875          0.7140
## par_median                 0.0780   0.2188          1.2405
##
## Percent Balance Improvement:
##                          Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                                  51.6      -40.4      71.1     24.4
## tier.f1                                 -376.9          .    -376.9   -376.9
## tier.f2                                   56.9          .      56.9     56.9
## tier.f3                                  100.0          .     100.0    100.0
## tier.f4                                  100.0          .     100.0    100.0
## tier.f5                                   64.0          .      64.0     64.0
## tier.f6                                   76.9          .      76.9     76.9
## tier.f8                                    0.0          .       0.0      0.0
## sat_avg_2013                            -203.5     -194.3     -93.2    -19.2
## asian_or_pacific_share                    89.8    -1841.1      49.2     34.0
## black_share                               33.8       11.3      72.7     65.4
## hisp_share                               -21.1       60.6      13.3     -3.3
## alien_share                               36.5        5.8      64.9     19.1
## par_median                              -135.2       10.6     -36.6     -8.5
##
## Sample Sizes:
##           Control Treated
## All            62      32
## Matched        32      32
## Unmatched      30       0
## Discarded       0       0
```

Compare the balance before and after matching

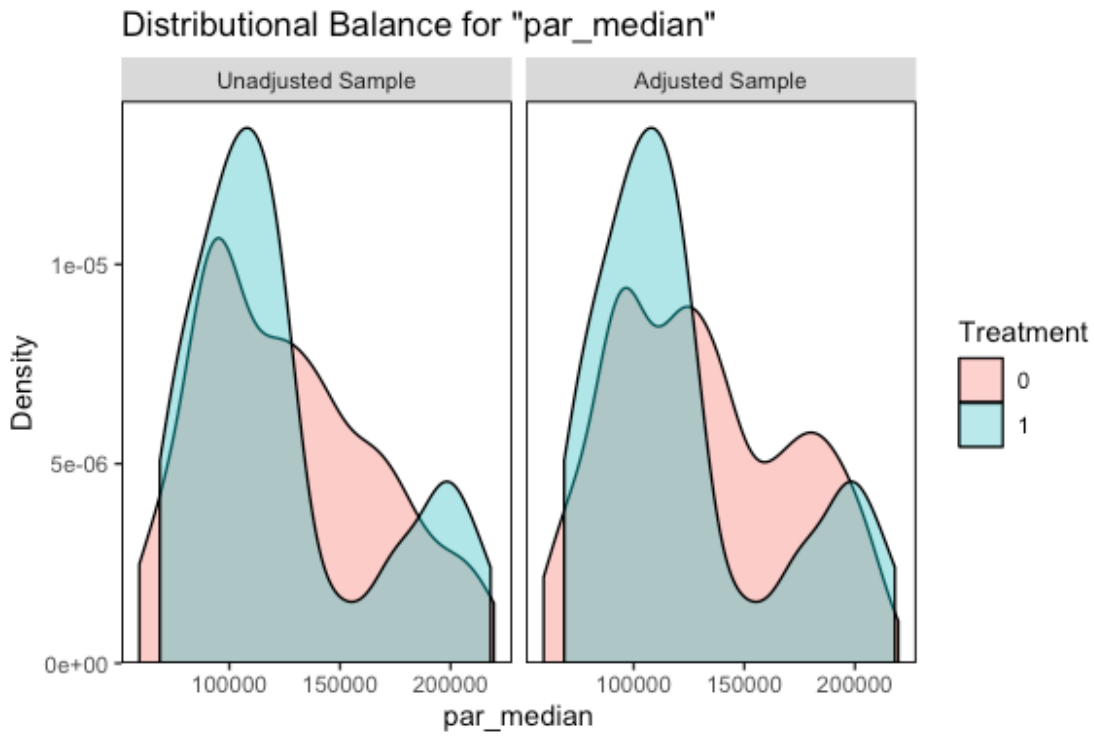
```
# Plot of each covariate before and after matching
bal.plot(matched_output, var.name = "tier.f", which = "both")
```



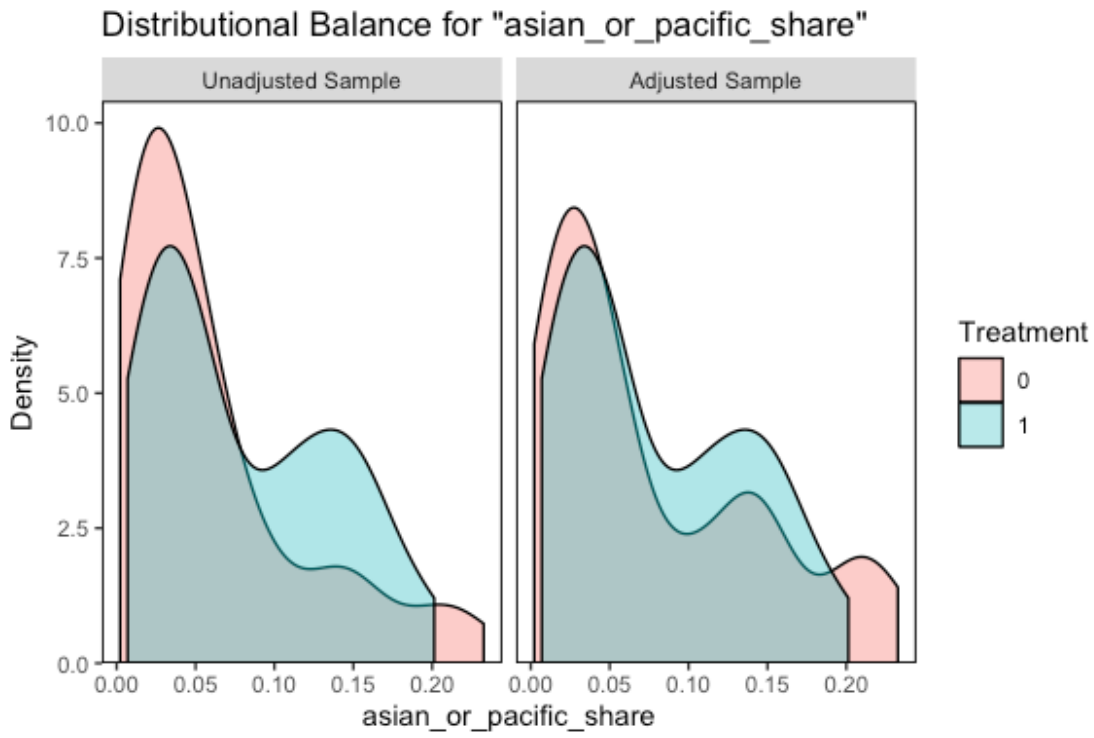

```
bal.plot(matched_output, var.name = "sat_avg_2013", which = "both")
```



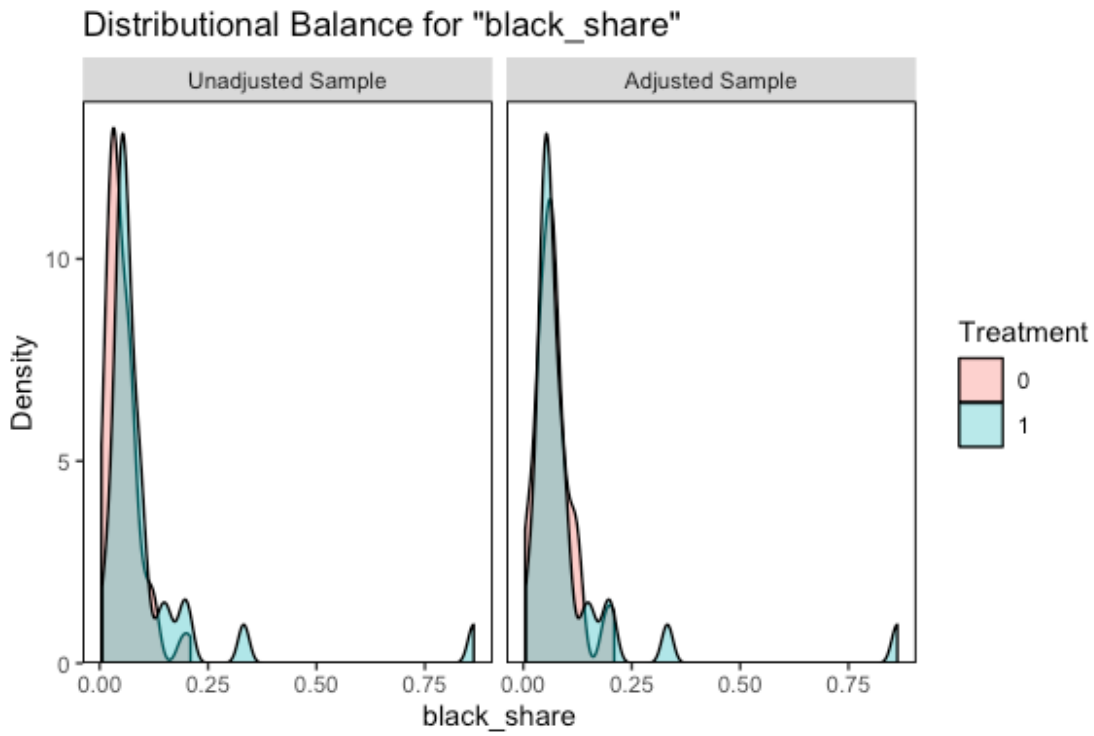

```
bal.plot(matched_output, var.name = "par_median", which = "both")
```



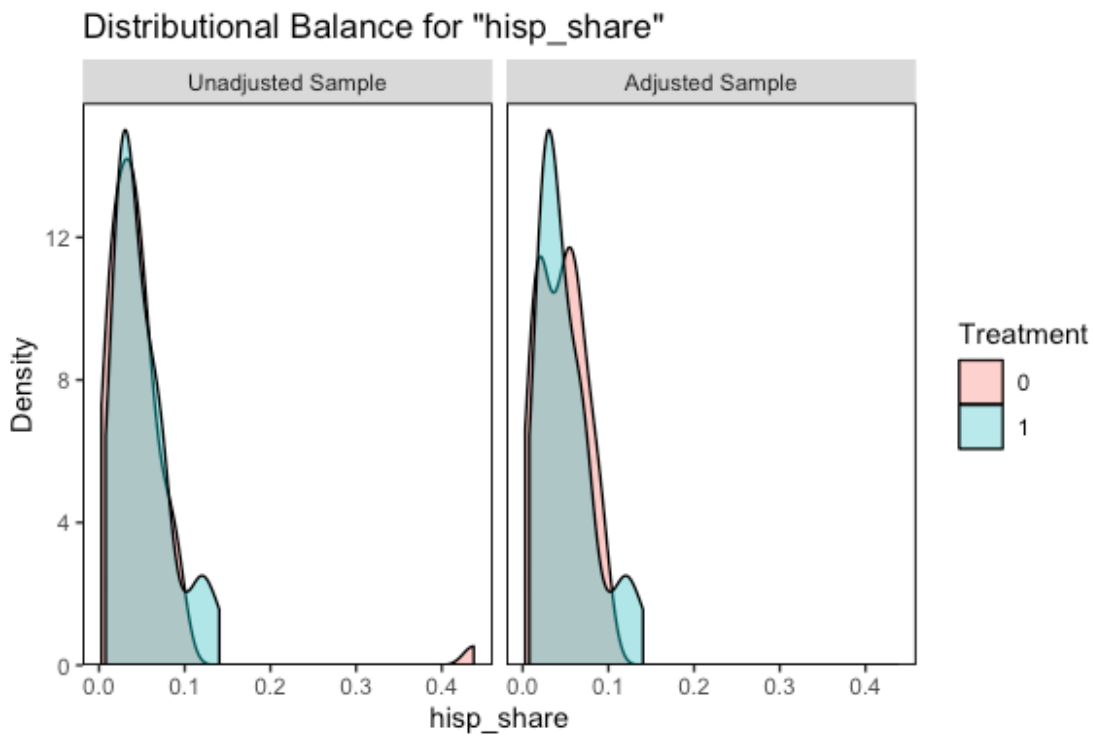

```r
bal.plot(matched_output, var.name = "asian_or_pacific_share", which = "both")
```



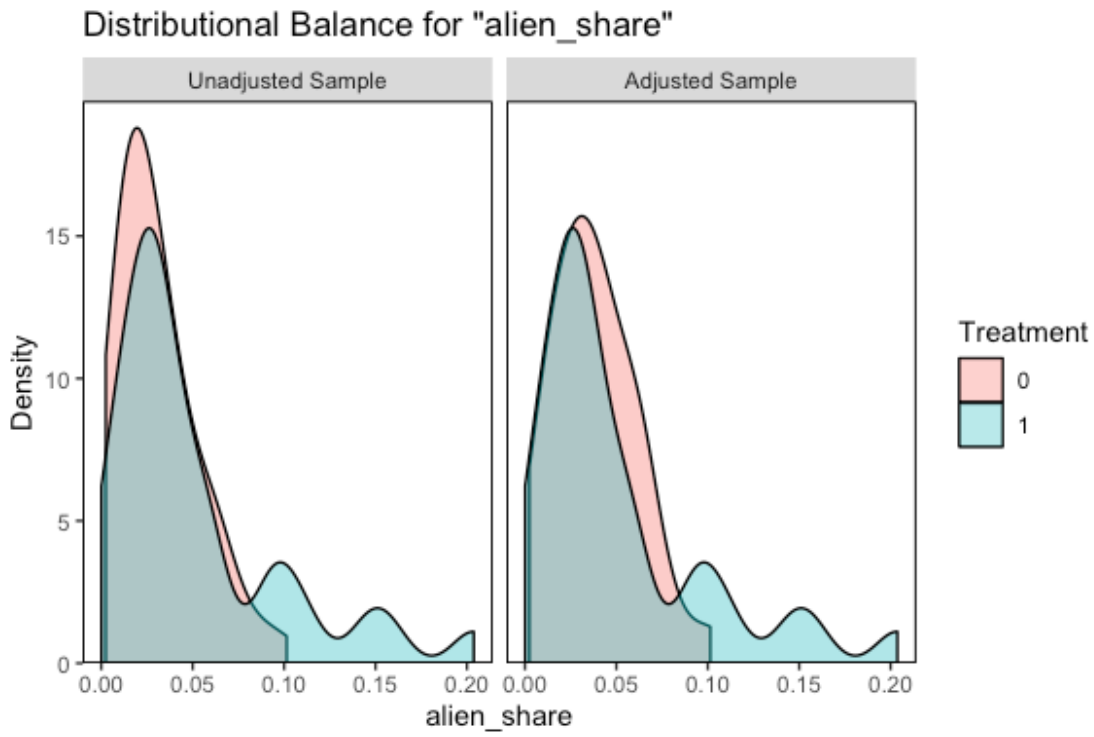

```r
bal.plot(matched_output, var.name = "black_share", which = "both")
```




```
bal.plot(matched_output, var.name = "hisp_share", which = "both")
```

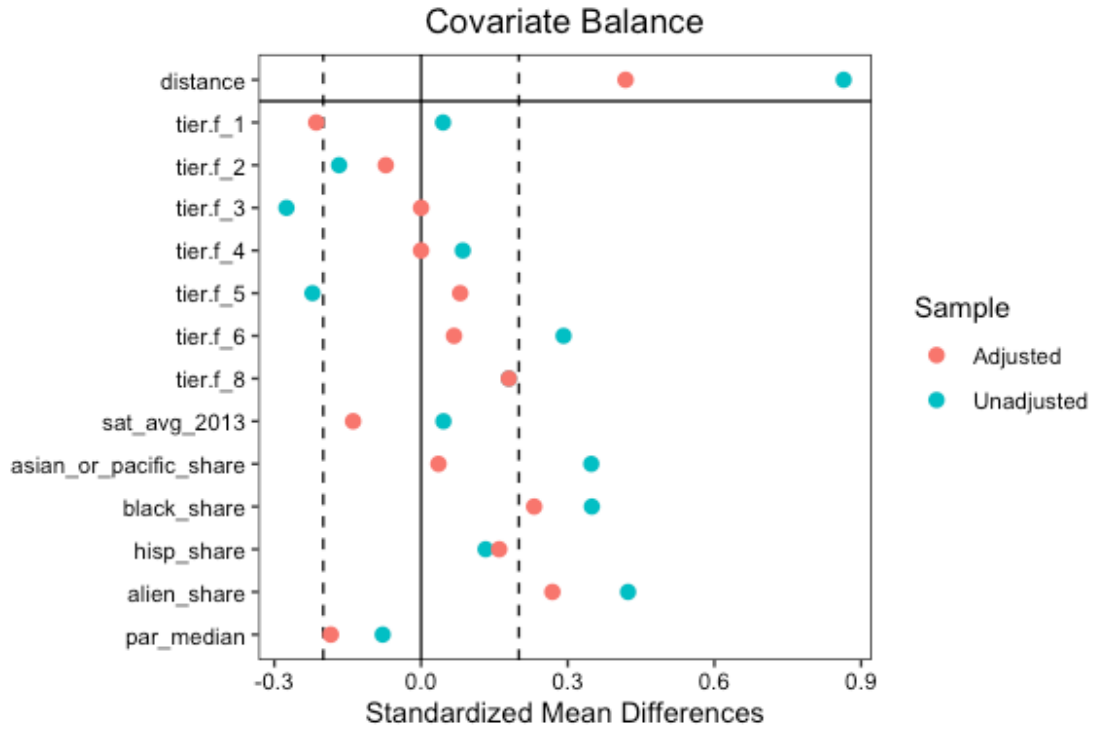



```
bal.plot(matched_output, var.name = "alien_share", which = "both")
```



Covariate Balance in different plot

```
love.plot(matched_output, binary = "std", thresholds = c(m = .2))
```



###

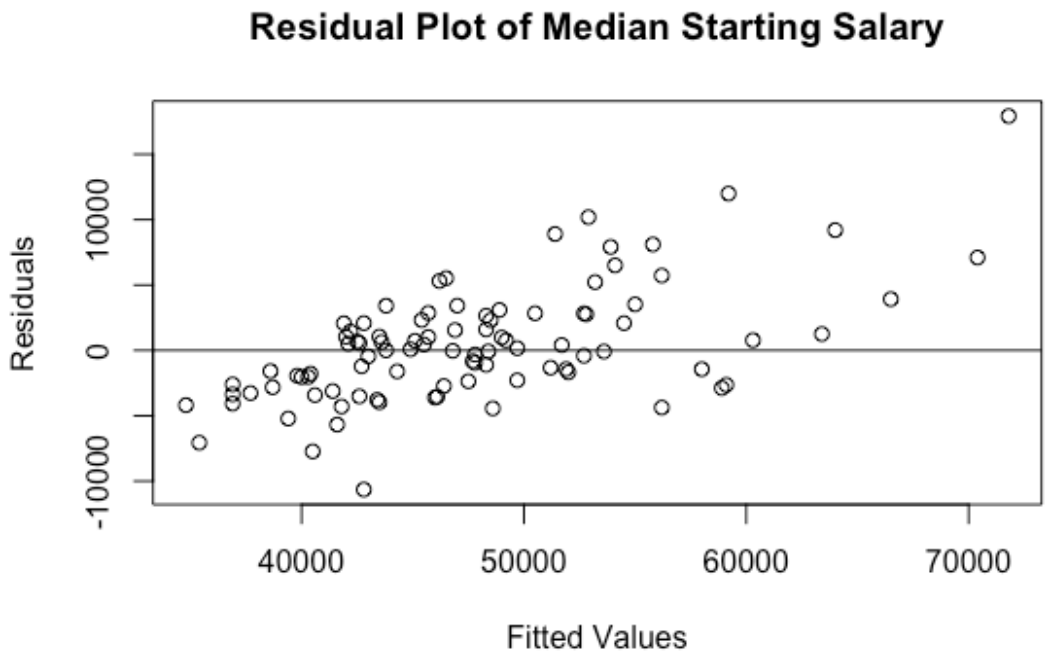
Regression with matched_data

```r
# Drop unmatched subject and create matched data
matched_data <- match.data(matched_output, data=data, group="all", distance
="pscore")
# Run regression based on the new data- matched data
fit_matched <- lmrob(starting_median_salary ~ treated + tier.f + sat_avg_2013
+ asian_or_pacific_share + black_share + hisp_share + alien_share +
par_median, data = matched_data)
# Summary the regression result
summary(fit_matched)

##
## Call:
## lmrob(formula = starting_median_salary ~ treated + tier.f + sat_avg_2013 +
##     asian_or_pacific_share + black_share + hisp_share + alien_share +
par_median,
##     data = matched_data)
##  \--> method = "MM"
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9452.0 -2293.6   135.8  2174.1 19666.1
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.866e+04  1.369e+04   2.093   0.0414 *
## treated                 9.729e+02  1.157e+03   0.841   0.4044
## tier.f2                -7.881e+03  1.836e+03  -4.294 7.87e-05 ***
## tier.f4                -5.680e+03  3.520e+03  -1.614   0.1128
## tier.f5                -5.481e+03  5.590e+03  -0.981   0.3314
## tier.f6                -5.389e+03  4.510e+03  -1.195   0.2376
## tier.f8                -7.628e+03  3.927e+03  -1.942   0.0576 .
## sat_avg_2013            1.111e+01  8.993e+00   1.236   0.2222
## asian_or_pacific_share  3.113e+04  2.026e+04   1.536   0.1307
## black_share             6.953e+03  3.925e+03   1.771   0.0825 .
## hisp_share              4.098e+04  2.949e+04   1.389   0.1707
## alien_share            -1.615e+04  1.452e+04  -1.113   0.2710
## par_median              5.062e-02  4.524e-02   1.119   0.2684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 3639
## Multiple R-squared:  0.744,  Adjusted R-squared:  0.6837
## Convergence in 26 IRWLS iterations
##
## Robustness weights:
##  observation 2 is an outlier with |weight| = 0 ( < 0.0016);
##  7 weights are ~= 1. The remaining 56 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3192  0.8816  0.9576  0.8882  0.9796  0.9972
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
```

```
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         1.563e-03
##             eps.x warn.limit.reject warn.limit.meanrw
##         3.967e-07         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s           k.max
##            500             50              2              1             200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0            2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

Residuals with unmatched data

```
lm.01 <- lmrob(starting_median_salary ~ treated + tier.f + sat_avg_2013 +
asian_or_pacific_share + black_share + hisp_share + alien_share + par_median,
data = data)
lm.res = resid(lm.01)
plot(data$starting_median_salary, lm.res,  ylab="Residuals", xlab="Fitted
Values", main="Residual Plot of Median Starting Salary")
abline(0, 0)
```



Regression without matching

```
# Regression without matching
fit_unmatched <- lmrob(starting_median_salary ~ treated + tier.f +
```

```
sat_avg_2013 + asian_or_pacific_share + black_share + hisp_share +
alien_share + par_median, data = data)

# Summary of the regression
summary(fit_unmatched)

##
## Call:
## lmrob(formula = starting_median_salary ~ treated + tier.f + sat_avg_2013 +
##     asian_or_pacific_share + black_share + hisp_share + alien_share +
par_median,
##     data = data)
##  \--> method = "MM"
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10646.94  -2544.16     40.28   2316.21  17923.55
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.480e+04  1.102e+04   3.157  0.00225 **
## treated                 1.104e+03  1.025e+03   1.078  0.28433
## tier.f2                -7.556e+03  1.660e+03  -4.551 1.88e-05 ***
## tier.f3                -4.013e+03  3.094e+03  -1.297  0.19837
## tier.f4                -7.980e+03  3.342e+03  -2.388  0.01930 *
## tier.f5                -8.534e+03  4.440e+03  -1.922  0.05815 .
## tier.f6                -7.418e+03  4.061e+03  -1.827  0.07144 .
## tier.f8                -8.809e+03  3.733e+03  -2.360  0.02071 *
## sat_avg_2013            1.049e+01  7.254e+00   1.446  0.15208
## asian_or_pacific_share  3.100e+04  1.418e+04   2.187  0.03169 *
## black_share             5.469e+03  3.497e+03   1.564  0.12184
## hisp_share              1.689e+04  5.943e+03   2.842  0.00568 **
## alien_share            -2.061e+04  1.055e+04  -1.953  0.05430 .
## par_median              3.179e-02  3.377e-02   0.941  0.34932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 3746
## Multiple R-squared:  0.7088, Adjusted R-squared:  0.6615
## Convergence in 19 IRWLS iterations
##
## Robustness weights:
##  observation 2 is an outlier with |weight| = 0 ( < 0.0011);
##  7 weights are ~= 1. The remaining 86 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2841  0.8966  0.9535  0.9005  0.9867  0.9989
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         1.064e-03
```

```
##             eps.x warn.limit.reject warn.limit.meanrw
##         3.995e-07         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                    psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                   "SM"
## seed : int(0)
```