Propensity Score Matching Homework Using Data on Schools

This assignment is due **WEDNESDAY** April 24. Our final exam is **FRIDAY** April 26.

Introduction

In this homework assignment, you will apply propensity score matching techniques using the `MatchIt` package in R, utilizing the dataset 'newyork.csv'. The key variable for our analysis is `stw`, which indicates whether a school was designated a "school to watch".

The variables are as follows

- school: Name of the school
- tot: school size (enrollment)
- min: percentage of minority students in the school
- dis: percentage of students receiving free and reduced lunch
- stw: whether a school was designated a "Schools to Watch" school.
  - Awarded to "outstanding middle schools across the country".

This data is real and comes from the paper *Falbe, K. (2014). The relationship between Schools to Watch designation and academic achievement.* Unfortunately, I could not find the data that had the outcome variable of "math scores", so I have simulated the math scores variable to have similar properties to the one in her paper.

- Assume "math" is average score on a standardized math test given to all the schools..

Questions

Question 1: Why would a researcher use propensity score matching in this case? (2pt)

A. To ensure that the math score data from all schools are normally distributed.

B. To eliminate the need for collecting data on school characteristics such as teacher quality or student demographics.

C. To compare "school to watch" schools with similar non-designated schools to control for confounding variables that might affect math scores and designation.

D. To increase the sample size of "school to watch" schools by matching them with multiple non-designated schools.

C

Question 2: Matching (2pt)

Perform propensity score matching using the nearest neighbor method. Store it as `match_out`. Please use all three covariates.

```r
match_out <- matchit(stw ~ tot + min + dis, data = data, method = 'nearest')
```

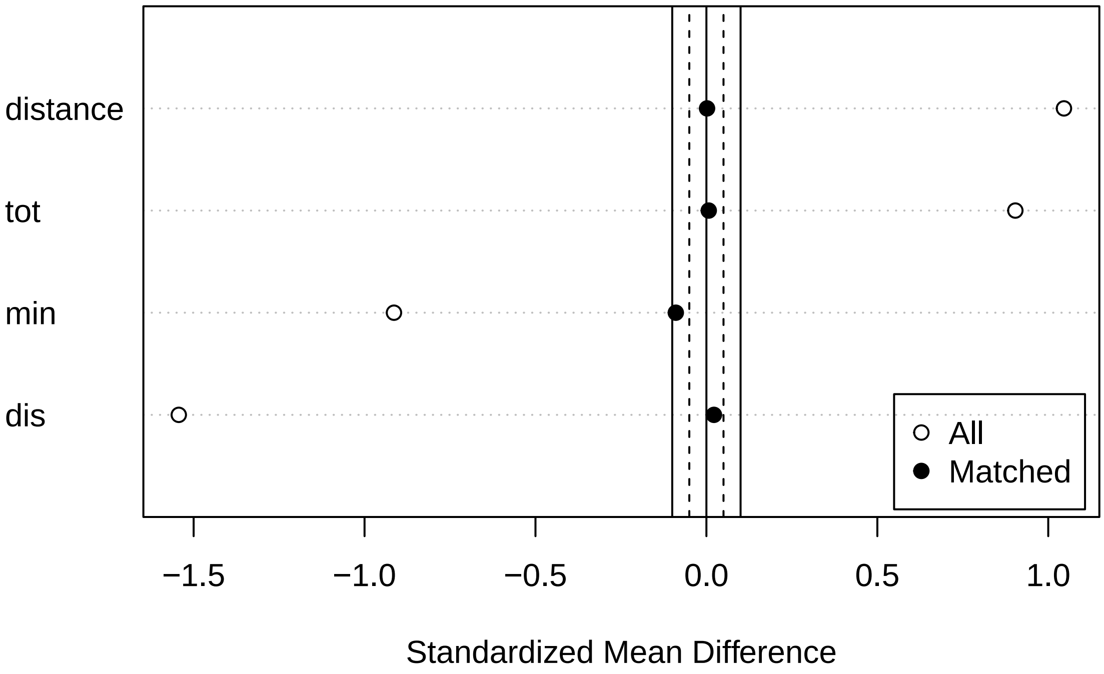
Question 3: Displaying Covariate Balance (2pt)

Provide the numerical summary of balance, the love plot, and the density plot. Be sure to set `abs = FALSE` in your `plot()` command when you make the love plot.

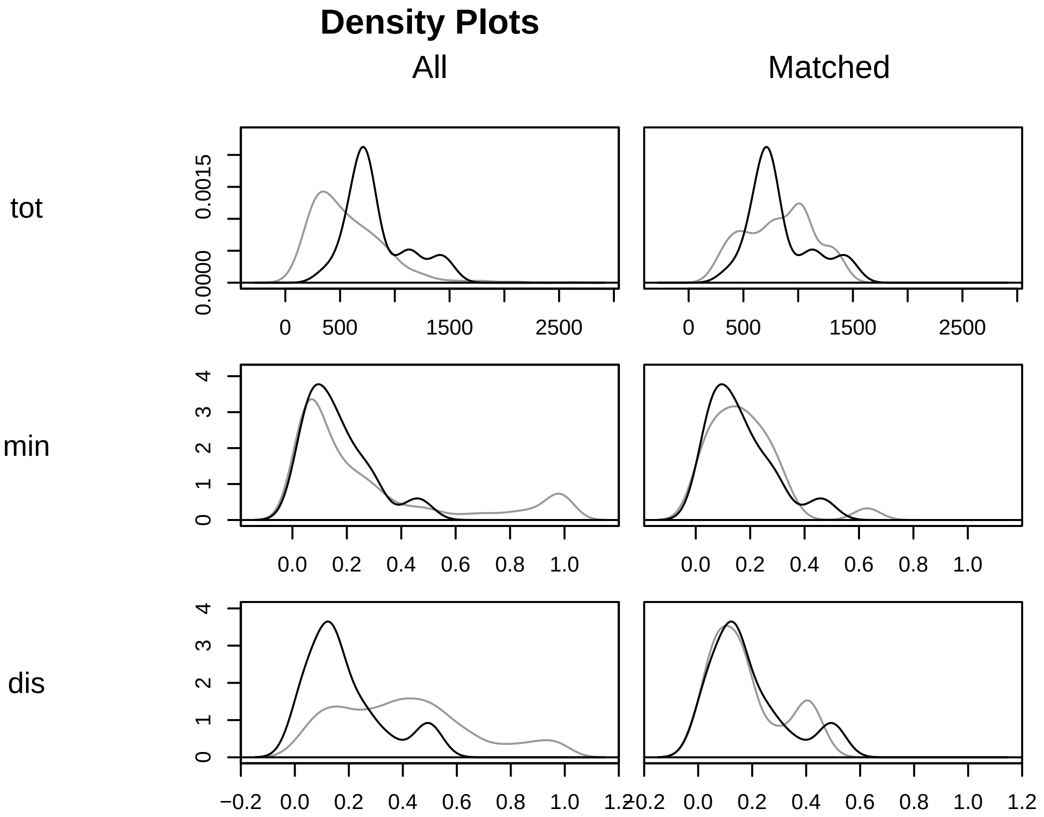
```r
summary(match_out)
```

```
##
## Call:
## matchit(formula = stw ~ tot + min + dis, data = data, method = "nearest")
##
## Summary of Balance for All Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.0943        0.0405          1.0459     1.0425    0.3159
## tot           832.6400      568.8998          0.9037     0.7650    0.2434
## min             0.1664        0.2767         -0.9140     0.1605    0.1115
## dis             0.1840        0.4079         -1.5435     0.3368    0.2283
##          eCDF Max
## distance   0.5185
## tot        0.4996
## min        0.1896
## dis        0.4744
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.0943        0.0942          0.0019     1.0016    0.0017
## tot           832.6400      830.6400          0.0069     0.8546    0.0902
## min             0.1664        0.1772         -0.0895     0.8234    0.0277
## dis             0.1840        0.1808          0.0221     1.1359    0.0264
##          eCDF Max Std. Pair Dist.
## distance     0.04          0.0098
## tot          0.24          1.1790
## min          0.12          1.2233
## dis          0.12          0.6177
##
## Sample Sizes:
##           Control Treated
## All           559      25
## Matched        25      25
## Unmatched     534       0
## Discarded       0       0
```

```r
plot(summary(match_out),abs=F)
```



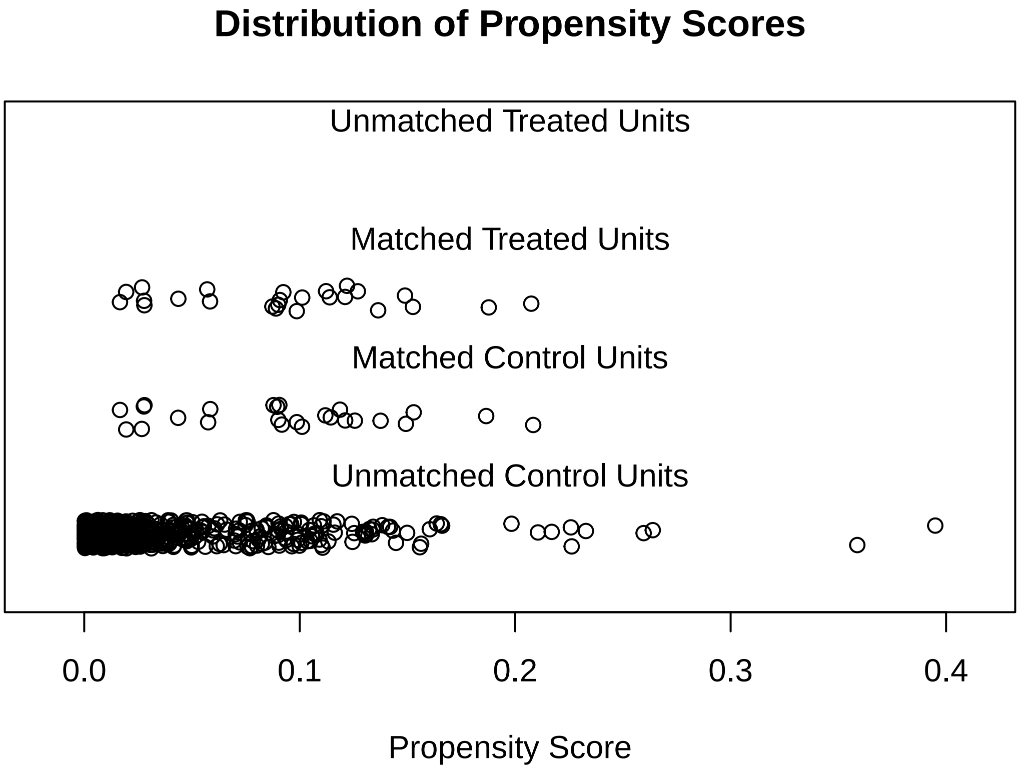
```
plot(match_out, type = 'density')
```



Question 4: Displaying Propensity Score Match (2pt)

Display a jitter plot.

```
plot(match_out,type='jitter')
```



```
## To identify the units, use first mouse button; to stop, use second.
```

Question 5: Assess the balance and match (2pt)

Briefly assess the quality of the balance achieved based on your results from question 3 and 4. Be sure to talk about the love plot, the density plots, and the jitter plot.

The balance achieved by the match is good. The summary shows that the average standard mean difference in absolute value across all covariates matched is 0.0301, implying that the mean difference across the treatment and control group is small. The love plot visualizes this by showing the matched units have a mean difference of essentially zero. Notably, the density plot shows that the disribution of the covariates in the treated and control are not exactly the same even though their average difference is small. The jitter plot shows that the distribution of the propensity scores across treated and control is quite similar. Where there is greater density of units in the treated group; a similar density of units is observed at the same propesnity score interval for the control group. One conern may be the low propesnity scores in the treated and control groups; however it is prefered to have a better match than to obtain high propensity scores. This is preferred because having a better match allows us to better control for potential confounders across the treated and control group that might affect math scores and designation.

Question 6: Analyzing Matched Data (2pt)

Estimate the following four regression models:

1. $Math = \beta_0 + \beta_1 Stw + \epsilon$ on the full data.
2. $Math = \beta_0 + \beta_1 Stw + +\beta_2 Tot + \beta_3 Min + \beta_4 Dis + \epsilon$ on the full data.
3. $Math = \beta_0 + \beta_1 Stw + \epsilon$ on the matched data
4. $Math = \beta_0 + \beta_1 Stw + +\beta_2 Tot + \beta_3 Min + \beta_4 Dis + \epsilon$ on the matched data.

Store them as reg1 to reg4 or mod1 to mod4, these are a common naming scheme for models.

```
match_data <- match.data(match_out)
```

```
reg1 <- lm(math ~ stw,data = data)
reg2 <- lm(math ~ stw + tot + min + dis, data = data)
reg3 <- lm(math ~ stw, data = match_data)
reg4 <- lm(math ~ stw + tot + min + dis, data = match_data)
```

Question 7: Displaying them in a table (2pt)

Fully replicate the stargazer table we made in the code-along.

```
stargazer(reg1,reg2,reg3,reg4, type = 'text',
          omit.stat = c("f", "ser"),
          omit = c("mom", "income"),
          add.lines = list(c("Sample", "Full", "Full", "Matched", "Matched"),
                           c("Controls?", "No", "Yes", "No", "Yes")))
```

```
##
## ======================================================
##                          Dependent variable:
##               -----------------------------------------
##                                 math
##                  (1)       (2)        (3)       (4)
## ------------------------------------------------------
## stw             2.269**    -0.944    -0.963     -0.843
##                (1.146)   (0.932)    (1.447)   (1.322)
##
## tot                         0.001                -0.002
##                          (0.001)              (0.003)
##
## min                         -0.322               5.965
##                          (0.896)              (6.167)
##
## dis                     -13.241***           -16.453***
##                          (1.123)              (4.790)
##
## Constant     79.584*** 84.615***  82.816*** 85.981***
##                (0.237)   (0.592)    (1.023)   (2.238)
##
## ------------------------------------------------------
## Sample         Full       Full     Matched   Matched
## Controls?       No         Yes         No        Yes
## Observations    584       584         50        50
## R2             0.007      0.375      0.009     0.226
## Adjusted R2    0.005      0.371     -0.012     0.158
## ======================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Question 8: Interpretation of unconditional estimate. (2pt)

Explain what the estimate in column 1 means, in literal terms. Consider how nothing is being controlled for. Think back to what I said in the code along above the "questions" section.

The estimate in column 1 is explaining the difference in the means of math scores across schools that are "Schools to Watch" versus schools that are not designated as "Schools to Watch".

Question 9: Explain what it means. (2pt)

Explain the implication of how column 1 has a positive, statistically significant result, but all the other columns have a null estimate, meaning that we would not claim a non-zero relationship between being designated as a "school to watch".

The null estimates imply that the confounding factors of school size, percentage of minority students in the school, and percentage of students receiving free and reduced lunch are introducing positive bias into the column 1 estimate and making the estimate statistically significant. When in reality, the designation does not have a statistically signficiant effect on the math scores of the students.

Question 10: Causal interpretation. (2pt)

According to our matching analysis, does being designated a "School to Watch" affect math test scores?

No, being designated as a "School to Watch" does not affect math test scores. There are confounding factors that can affect math scores such as school size, percentage of minority students in the school, and percentage of students receiving free and reduced lunch. Once those are matched or controlled for in the regression models, there is a null effect of the "School to Watch" designation on math test scores.

Disclaimer:

Since I do not have access to the researcher's test score data, I don't know what her results would have been had she just used regression adjustment, like you did in model 2. For us, models 2 through 4 tell us the same thing, which means matching may not have been necessary. However, matching is probably *at least as good* as regression adjustment, in general. So, here, it looks like we didn't "need" to match, and just adjusting for some characteristics of the schools nullifies the "effect", but it's hard to find good replicable data without just fully simulating it. When possible, I like to use real data based on real research. I hope you have found matching to be interesting!