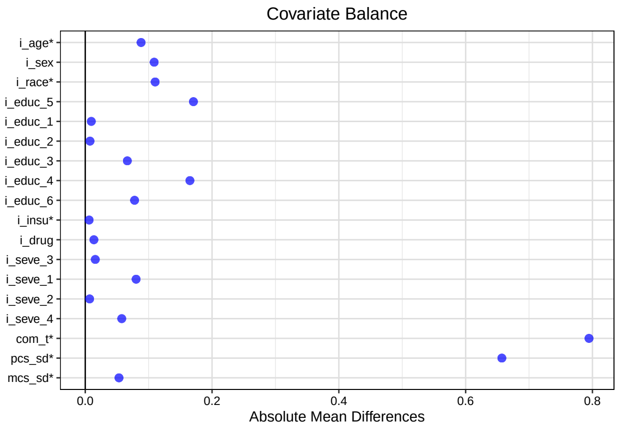
Methods and Analysis 5

Ashwini Marathe

11/19/2019

Question 1

Are the covariates in this data balanced between the two groups? If no, which covariates are not? How did you assess balance?



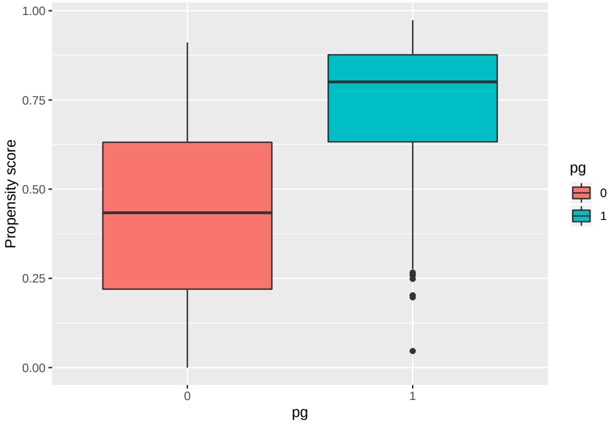
Checked the balance between the covariates using absolute standardized difference (ASD). Absolute values of ASD above 0.1 are assumed to be imbalanced. The plot above shows the balance between variates and we can see that certain covariates are highly imbalanced (*pcs_sd*, *com_t*). Using a thresholding value of 0.1, 6 covariates were found to imbalanced. The list of the covariates and the corresponding ASD values are represented in the table below.

Variables	ASD
i_sex	-0.1087256
i_race	0.1095029
i_educ_5	0.1706028
i_educ_4	-0.1650427
com_t	-0.9871509
pcs_sd	0.7536967

Question 2

Estimate the propensity score e using a logistic regression with all pre-treatment variables entering in the model as main effects.

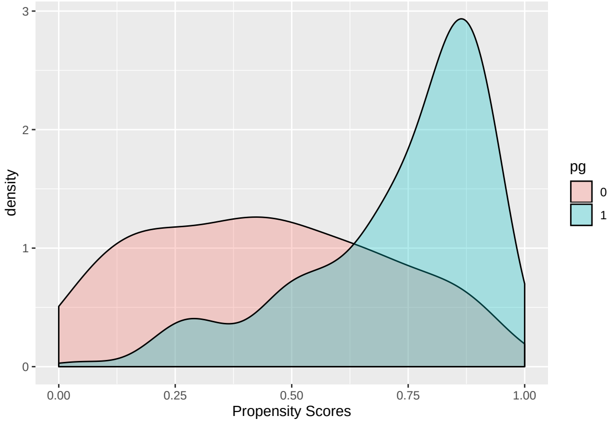
From the graph below we can see that the two groups are probably not balanced.



Question 2a:

Are there any observations with an estimated propensity score e that is out of the range of e in the other group? If there are only a few such outliers (less than 5), keep them; If many, discard them and report the number of the discarded observations. This is to ensure overlap!

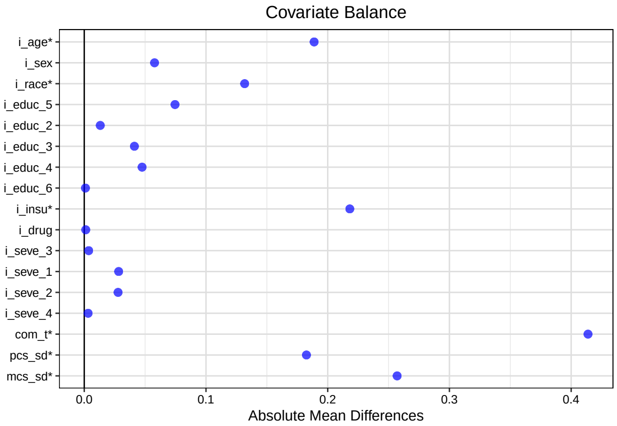
From the density plot below we see that the plots do not overlap much. Checking the overlap between the points we observe that a total of points (4 on right and 21 on left) do not overlap. Since this is a significant number we discard these non-overlapping observations.



Question 2b:

Using one-to-one, nearest neighbor matching on the estimated propensity scores, check balance again. Are the covariates balanced now? If no, which ones are not?

Matched the data using nearest neighbor 1:1 matching using the propensity scores obtained from logistic regression. After the matching 6 variables remain unbalanced. The plot below shows the ASD values for the covariates after matching.



Question 2c:

Estimate the average causal effect Q using the matched sample obtained above. Also, report a standard error for your estimate. Construct a 95% confidence interval and interpret your findings.

Using the matched sample the average causal effect Q is -0.1898148. The standard error for the estimate is 0.0658501 and the confidence interval for average causal effect is [-0.3188809, -0.0607487].

Interpretation: Since the confidence interval of the average causal effect Q does not include 0, we can say that 18.9814815% people served by the patient group 1 are less satisfied with the treatment as compared to patient group 0.

Question 2d:

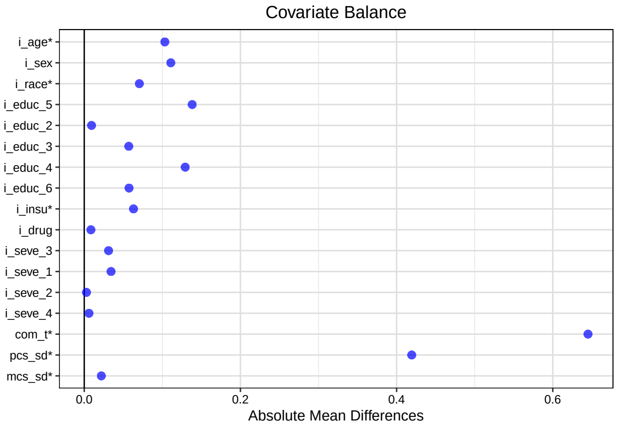
Fit a logistic regression to the response variable using the main effects of all pre-treatment variables. Also include the propensity score e as a predictor. Report the estimated causal odds ratio. If it is significant, interpret the effect in context of the problem. Note that this estimated effect is not an estimate of Q = p1 - p2 but intuitively, it still makes sense to look at it.

The variable *pg* is significant in the logistic regression model and the causal odds ratio is 0.31 and the confidence interavl for it is [0.12, 0.701]. Since this value is less than 1 and the confidence interval does not include 1, we can draw conclusion similar to Question 1c, that people served by the patient group 1 are satisfied less with the treatment as compared to patient group 0.

Question 2e:

Repeat parts (b) to (d) using one-to-many (five) nearest neighbor matching with replacement, instead of one-to-one nearest neighbor matching. How do your results compare to what you had before?

5 covariates are imbalanced in this case. However the imbalance is less than 1:1 matching. Below is the plot of ASD values:



Using the matched sample from one to many matching, the average causal effect Q is -0.1390449. The standard error for the estimate is 0.0597291 and the confidence interval for average causal effect is [-0.2561139, -0.021976].

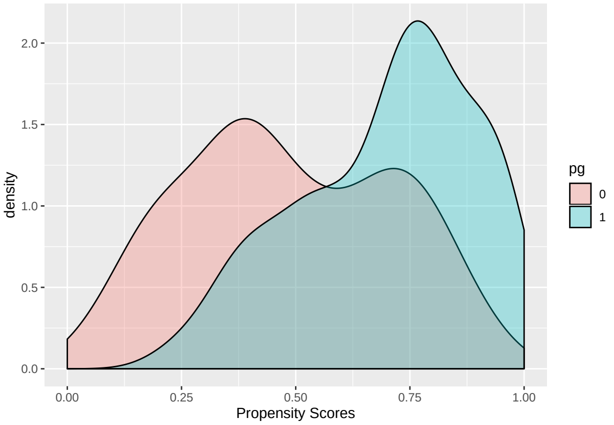
Interpretation: Since the confidence interval of the average causal effect Q does not include 0, we can make meaningful interpretaions form the value of treatment effect obtained like in the 1:1 matching. People in group 1 are 13.9044944 less satisfied than people in group 0.

The variable *pg* is significant in the logistic regression model and the causal odds ratio is 0.42 and the confidence interavl for it is [0.207, 0.856]. Since this value is less than 1 and the confidence interval does not include 1, we can draw conclusion similar to Question 1c, that people served by the patient group 1 are satisfied less with the treatment as compared to patient group 0.

Question 3

Question 3a

Are there any observations with an estimated propensity score e that is out of the range of e in the other group? If there are only a few such outliers (less than 5), keep them; If many, discard them and report the number of the discarded observations. This is to ensure overlap!

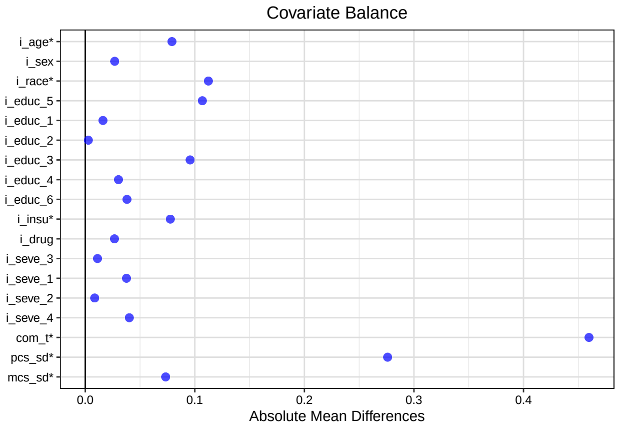


From the density plot above we see that the plots overlap better than in logistic regression Checking the overlap between the points we observe that a total of 17 points (3 on right and 14 on left) do not overlap. Since this is a significant number we discard these non-overlapping observations.

Question 3b:

Using one-to-one, nearest neighbor matching on the estimated propensity scores, check balance again. Are the covariates balanced now? If no, which ones are not?

Matched the data using nearest neighbor 1:1 matching using the propensity scores obtained from random forest. After the matching 3 variables are not matched. But the overall match looks better than logistic regression.



Question 3c:

Estimate the average causal effect Q using the matched sample obtained above. Also, report a standard error for your estimate (use the formula for computing standard error for difference in proportions; if you are not familiar with this, check page 280 of the OIS book we used for the online summer review). Construct a 95% confidence interval and interpret your findings.

Using the matched sample the average causal effect Q is -0.1152688. The standard error for the estimate is 0.0681754 and the confidence interval for average causal effect is [-0.2488926, 0.018355].

Interpretation: Since the confidence interval of the average causal effect Q does not include 0, we can say that people served by the patient group 1 are 18.9814815% less satisfied with the treatment as compared to patient group 0.

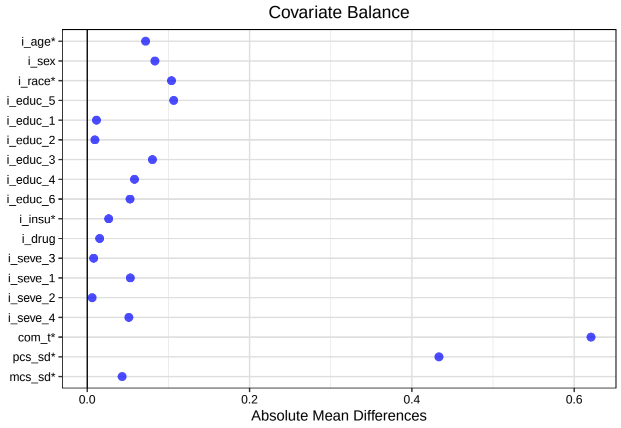
Question 3d:

Fit a logistic regression to the response variable using the main effects of all pre-treatment variables. Also include the propensity score e as a predictor. Report the estimated causal odds ratio. If it is significant, interpret the effect in context of the problem. Note that this estimated effect is not an estimate of Q = p1 - p2 but intuitively, it still makes sense to look at it.

The variable *pg* is significant in the logistic regression model and the causal odds ratio is 0.35 and the confidence interavl for it is [0.15, 0.81]. Since this value is greater than 0 and the confidence interval does not include 0, we can draw conclusion similar to Question 3c, that people served by the patient group 0 are satisfied more with the treatment as compared to patient group 1.

Question 3e:

A lot of covariates are balanced in this case. Below is the table of the imbalanced covariates:



Using the matched sample from one to many matching, the average causal effect Q is -0.1460606. The standard error for the estimate is 0.059627 and the confidence interval for average causal effect is [-0.2629296, -0.0291916].

Interpretation: Since the confidence interval of the average causal effect Q does not include 0, we can comment on the treatment effect. People in group 1 are 14.6060606% less satisfied than people in group 0.

The variable *pg* is significant in the logistic regression model and the causal odds ratio is 0.42 and the confidence interavl for it is [0.207, 0.8521]. Since this value is less than 1 and the confidence interval does not include 1, we can draw conclusion similar to Question 3c, that people served by the patient group 1 are less satisfied with the treatment as compared to patient group 0. However, in 1:many matching the confidence intervals are narrower than the 1:1 matching.

Question 4

Which of the methods do you consider most reliable (or feel most comfortable with) for estimating the causal effect? Why?

The 1:n matching using propensity score from random forest balanced the data very well. More the balance more can we rely on the treatment effect values that we get. Also the propensity score density plot for both the groups is more balanced than that in logistic regression. The confidence intervals for treatment efect are narrower in 1:n matching using propensity score from random forest.