IDS 702 - Data Analysis Assignment 5

Asthma Patients in California

Q1. Are the covariates in this data balanced between the two groups? If no, which covariates are not? How did you assess balance?

On the basis of the love plot and balance table, we look for the variables whose absolute value is greater than 0.1 (deviation from the mean difference of the distribution) to consider them unbalanced. We have 6 variables:
i) i_sex (sex of the asthma patient)
ii) i_race_1 (race category 1)
iii) i_race_2 (race category 2)
iv) i_educ_5 (education category 5)
v) com_t [centered] (total number of comorbidity)
vi) pcs_sd [centered] (standard physical comorbidity scale)

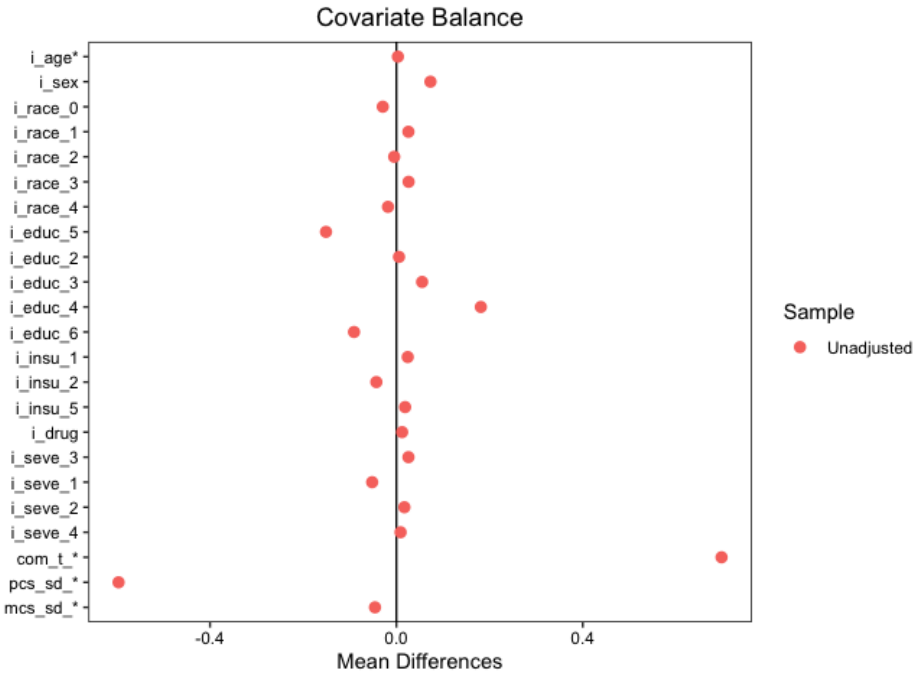

Figure 1: Covariates Balance Graph

Q2. Estimate the propensity score e using a logistic regression with all pre-treatment variables entering in the model as main effects

To estimate the propensity scores, we fit a logistic model on our data with pg variable (treatment assignment) as the response variable since it is binary. The pg variable represents the physician assigned to the patient (1 or 2). For our analysis, we have considered

Q2(a). 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. Note that this is to ensure overlap

Based on the histogram we can see the distribution of the propensity scores. The density plots show us overlap between the treatment and control groups. However we do have outliers. There are 8 outliers on the left tail and 40 on the right tail. Since the outliers are greater than 5 in number we drop these values.

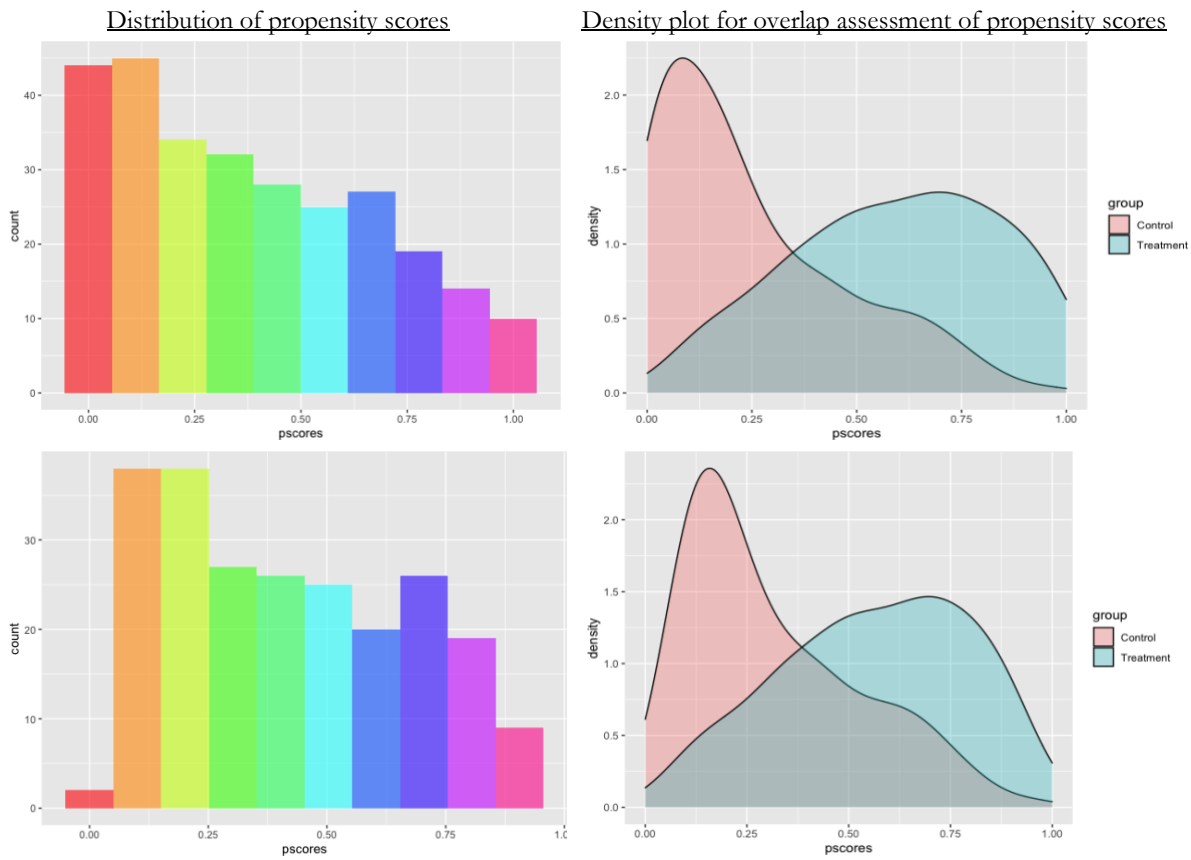

Figure 2: Comparison of plots before and after removal of outliers

The top row in Figure 2 represent the plots before the removal of outliers. The bottom row is after the outliers were removed. We can there is a change in the bars present at each end of the histogram. We can also observe that there is an improve in the overlap between treatment and control group.

Q2(b). 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?

We match on the estimated propensity scores using the matchit function. We then check the percent balance improvement. We still have some unbalanced covariates. They are:

i) i_age (sex of the asthma patient)

ii) i_race_2 (race category 2)
iii) i_race_4 (race category 4)
iv) i_insu_1 (insurance status 1)
v) mcs_sd [centered] (standard mental comorbidity scale)

We also observed 36 observations that were unmatched in the control group after using the one-one nearest neighbour matching.

Q2(c). Estimate the average causal effect Q "directly" using the matched sample obtained above. Also, report a standard error for your estimate. Construct a 95% confidence interval and interpret your findings.

The average causal effect Q is 0.155. The standard error estimate for the sample is 0.065. The confidence interval at the 95% confidence level is [0.028,0.281]. As our confidence interval does not contain the value 0, we can conclude that the treatment effect is statistically significant to determine the asthma patient satisfaction.

Q2(d). Fit a logistic regression to the response variable using the main effects of all pre-treatment variables on the matched data. Also include the treatment variable and the propensity score e as predictors. 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 = p2 – p1 but intuitively, it still makes sense to look at it

Based on our model, the estimated causal log odds ratio for the treatment variables(pg) is 0.920. With a p-value of 0.0256, it is statistically significant at the 95% confidence level.

Keeping all else constant, the odds of a positive satisfaction level for asthma patients who consulted physician1, 2.51 times more likely than asthma patients who consulted physician 2.

Q2(e). 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?

We repeat the process again with a one to many approach (5 in this case) and observe that in this case we only get 31 unmatched observations.
The Average Causal Effect Q on matched data with one to many matching is 0.167. The standard error estimate is 0.064. The confidence interval at the 95% confidence level is [0.043,0.295]. As our confidence interval does not contain the value 0, we can conclude that the treatment effect is statistically significant to determine the asthma patient satisfaction.

Based on our model, the estimated causal log odds ratio for the treatment variables(pg) is 0.880. With a p-value of 0.0285, it is statistically significant at the 95% confidence level. We can observe that this p-value is slightly higher than the value we obtained in the model run on one-one matched data.

Keeping all else constant, the odds of a positive satisfaction level for asthma patients who consulted physician1, 2.41 times more likely than asthma patients who consulted physician 2. The causal odds ratio is almost the same as what we obtained in the previous question.

Q3. Which of the methods do you consider most reliable (or feel most comfortable with) for estimating the causal effect? Why?

The one to many method seems more reliable since it matches one person in the treatment group to many individuals in the control group. This helps to average out the balance of covariates. However, we must keep in mind that while doing one to many matches we cannot ascertain how many people from the control group will be repeated while carrying out the match.