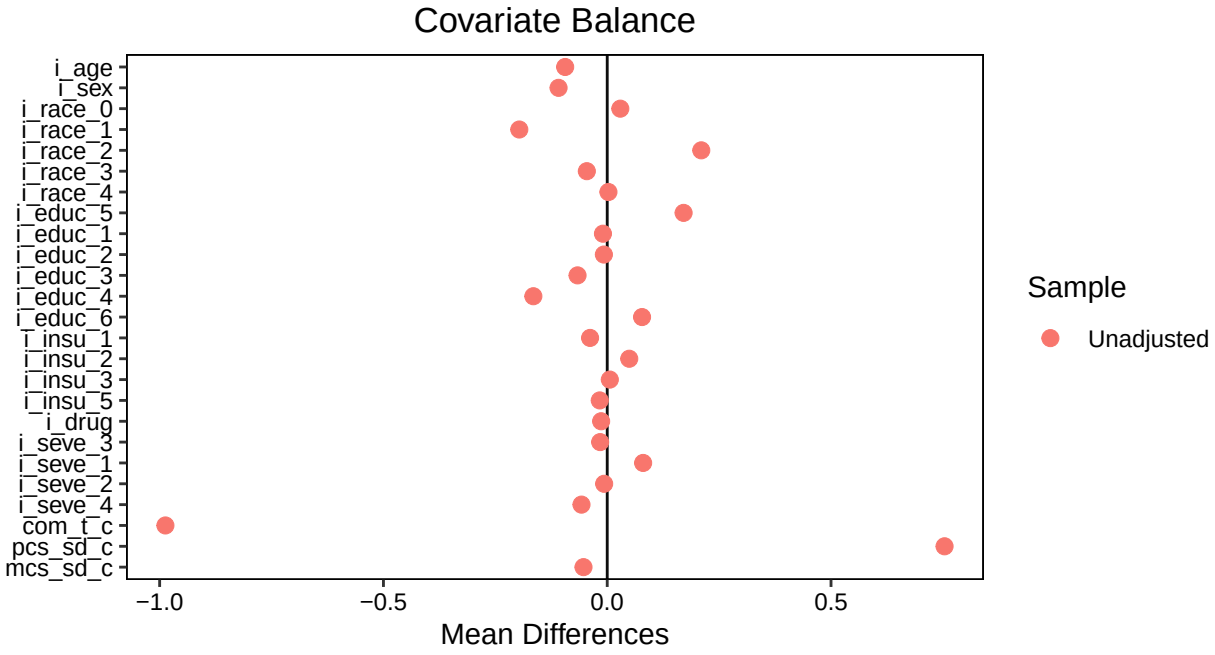
Assignment 5

Mohammad Anas

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

To check if the covariates are balanced or not between the control and the treatment group, we calculate absolute standardized difference between the two groups for all the covariates. The variables sex, race1, race2, educ4, educ5, number of comorbidities, physical comorbidity scale and mental comorbidity scale were imbalanced as the absolate value of the ASD for these variables was above 0.1. Before doing this analysis, we converted the respective variables to factors and also mean center the continuous variables.

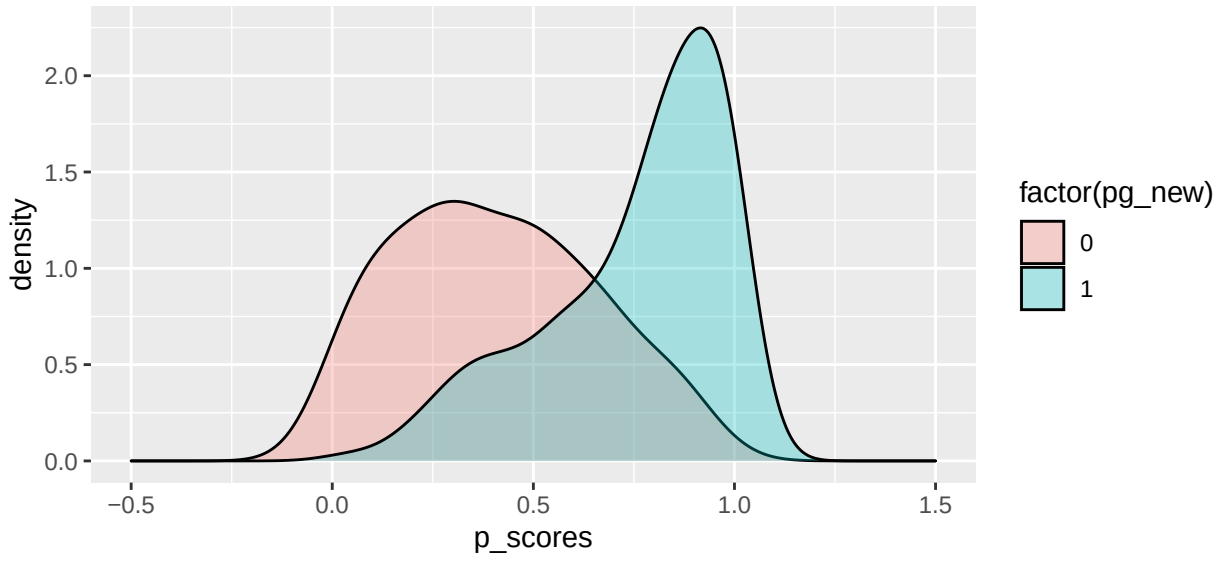


Estimate the propensity score e using a logistic regression with all pre-treatment variables entering in the model as main effects.

We estimate the propensity scores by running a logistic regression model with treatment variable as a response variable. We use the model to predict propensity scores. We use all the variables in the datasets we have in the model as overfitting is not a concern. It should be noted that the treatment variable here is where pg is 2 and control group has pg equal to 1.

(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.

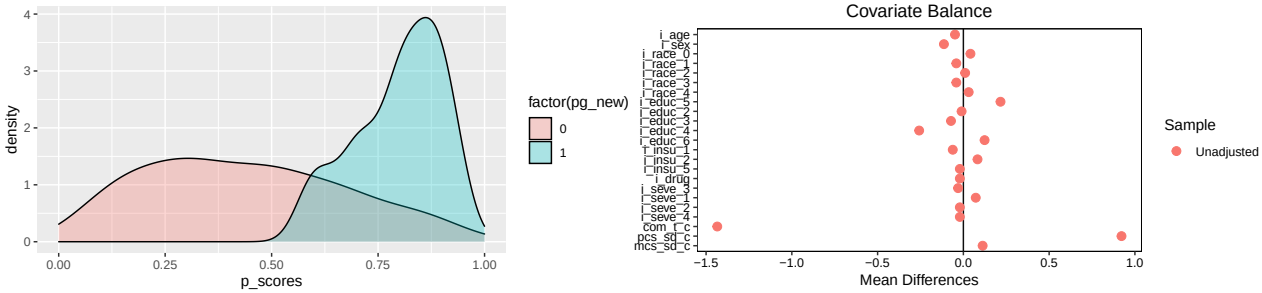
We do see a lot of outliers in regards to propensity scores. Some of the predicted values of the propensity score are outside the range zero and one. The density plot shown below shows us not only the overlap of covariates between the 2 groups but also shows us the propensity scores for one variable which are out of the range for the other group.



We see on the left side where the propensity scores are close to zero, the control group has scores which are out of range for the treatment group. On the other hand, on the right side there are a lot of data entries in the treatment group where estimated propensity score are greater than one. As these data entries are extreme and will be difficult to match we remove them from the data. We remove 8 entries where the propensity scores are at the lower side and 40 entries on the right side in such the a way that the upper and lower bound for propensity score is equal for both groups.

(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 data entries on the nearest neighbour propensity scores. We use one to one matching and extract 97 entries for each group that were balanced on the propensity scores. We check the covariate balance and see that the balance has not improved. The variables sex, educ4, educ5, educ6, number of comorbidities, physical comorbidity scale and mental comorbidity scale are still imbalanced as they have an ASD of above 0.01. The percentage change in balance has gone down for all the variables with age, race2, insu1 and seve4 showing huge drops in mean difference between the two groups. We create density plots of the covariates for both the groups and see that the overlapping area has gone down significantly



(c) Estimate the average causal effect Q "directly" 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 third edition of the OIS book we used for the online summer review). Construct a 95% confidence interval and interpret your findings.

We take the mean of the outcome variable for both the groups in the matched data and take the difference of the mean and not that the outcome variable (i_aqoc) has lower average for the treatment group as compared the the control group. The mean difference for the outcome variable comes out to be -0.20 (mean(i_aqoc|treatment) - mean(i_aqoc|control)). We calculate standard error and construct the confidence

intervals. The lower bound of our 95% confidence interval comes out to be -0.32 and upper bound is -0.06. As the confidence intervals do not contain zero, we can say that the causal effect is significant.

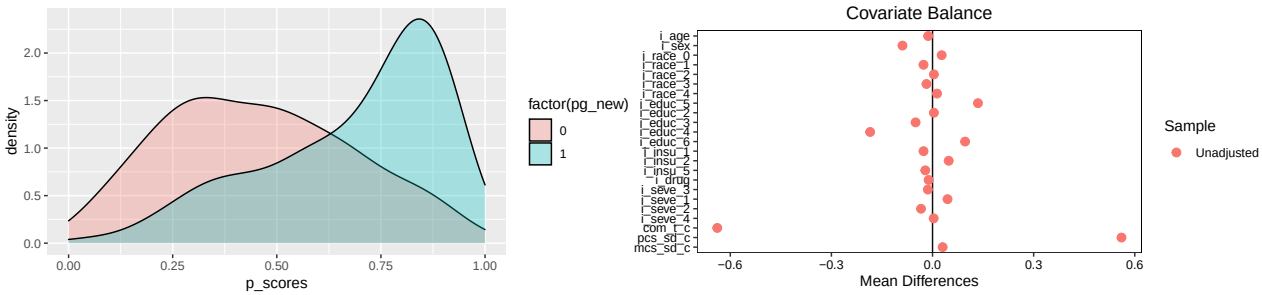
(d) Fit a logistic regression to the response variable using the main effects of all pre-treatment variables on the matched data. Also include the treatment variable and the propensity score e as predictors. Report the estimated causal odds ratio. If it is significant, interpret the effect in context of the problem.

Now that we our matched dataset, we use logistic regression to estimate the effect of treatment on our outcome variable. We include all the covariates as predictors and also control for the propensity scores in our model. The effect of the treatment group comes out to be significant at 5% significance level. The causal odds ratio are -0.86. The odds of patient being satisfied are 0.85 times the odds of not being satisfied if he belongs to physician group 2 as compared to the physician group 1.

(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 then use one to 5 nearest neighbor matching on the propensity scores to create another matched dataset. Given that we have allowed multiple data entries to be matched from control group on each entry in the treatment group, our data now comprises of 133 treatment data entries and 91 entries for control. We see that the mean differences has improved for most variables. However, it has gone down drastically for age. It has also gone down for other variables like race1, race2, race4, drug0, drug1,insu1, seve3, seve4 and mcs_sd_sc.

When we look at covariate balance using ASD almost all the variables seem balanced now except educ5, educ4,com_t and pcs_sd_c. Looking at the plots shown below we can see that the match has improved overall. (The scale for covariate balance plot here is different as compared to the ones above.)



We now estimate the difference in the average of our outcome variable for both groups. To calculate standard errors we used the respective number of data entries in each group from our match dataset (133 for treatment and 91 for control) and then calculate the confidence intervals. The upper bound for our 95% confidence interval comes out to be -0.03 and the lower bound is -0.27. The mean difference is -0.15 (mean(i_aqoc|treatment) - mean(i_aqoc|control)). The confidence interval does not contain zero and hence the causal effect is significant.

To confirm whether the two groups are actually significant we also run a logistic model, with our outcome variable as aqoc and all the covariates and propensity scores as predictors. The causal odds ratio for treatment comes out to be 0.47 and the effect is significant at 5% significance level. The odds of patient being satisfied are 0.47 times less if he belongs to physician group 2 as compared to the physician group 1.

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

I feel like in this case using one to many matching is more reliable. There are 2 reasons for this.

1. We were able to attain a better covariate balance.
2. Secondly, all the data entries in the treatment group were included in the analysis. Overall, we used 224 data entries in one to many match as compared to 194 observations in one to one matching. Hence, we wasted less data.

However, I think that this might not be the case every time. We can always tweak our analysis using different models and different matching techniques. We can iterate over different combinations of these until we get a good covariate balance.