Methods and Data Analysis 5

Srishti Saha (ss1078)

20 November, 2019

Question 1

Part 1

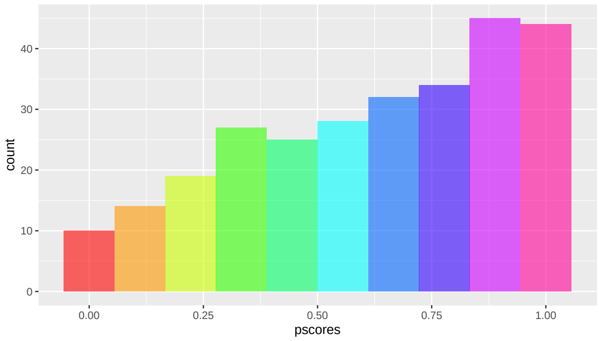
Table 1: Unbalanced covariates

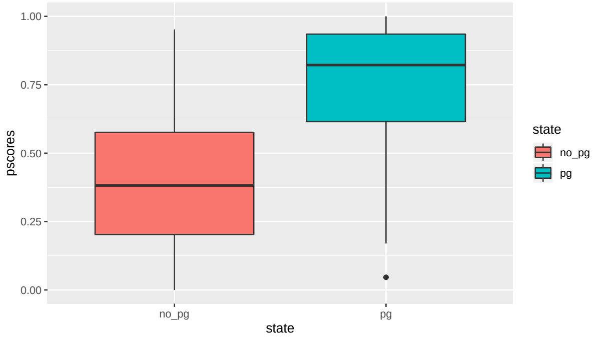
	Type	Diff.Un
i_sex	Binary	-0.1087256
i_race_1	Binary	-0.1964767
i_race_2	Binary	0.2101294
i_educ_5	Binary	0.1706028
i_educ_4	Binary	-0.1650427
com_t	Contin.	-0.9871509
pcs_sd	Contin.	0.7536967
i_aqoc	Binary	-0.1682356

There are 8 unbalanced covariates as mentioned above. We use the metric ASD (absolute standardized difference) for evaluating balances. Absolute value of the absolute standardized difference > 0.1.

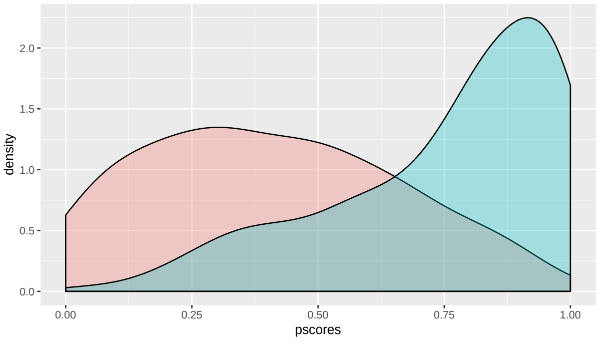
Part 2

Removing outliers





Too many probabilities on the borders. We can see clear differences in the distributions of propensity scores thus, a simple comparison of the outcomes would be confounded by differences in the background variables

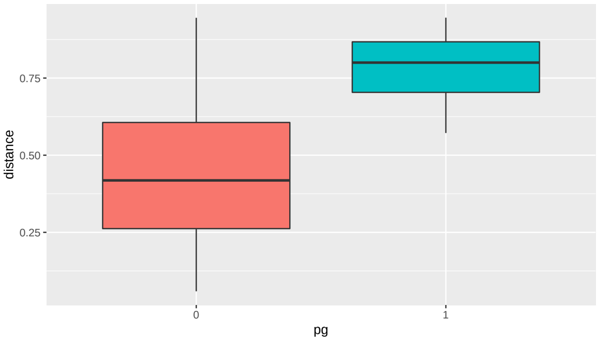


```
## [1] "48 outliers were dropped."
```

Here, as we see, 48 observations are discarded due to violation of the overlap.

One-to-one Matching and Balancing

Control units are fewer than treatment (test) units. Not all treated units are matched.



There is a very high difference between the distributions.

Most covariates are balanced now except:

Table 2: Remaining Unbalanced covariates for technique 1

	Mean Diff.
distance	-43.54492
i_age	-1349.59128
i_sex0	-55.14316
i_sex1	-55.14316
i_race1	-59.75976
i_race2	-118.03279
i_race3	-57.39645
i_race4	-71.24464
i_educ2	-84.72222
i_educ3	-30.75843
i_educ4	-42.21557
i_educ6	-36.06138
i_insu2	-91.71171
i_insu5	-10.37344
i_drug1	-68.35443
i_seve1	-38.74814
i_seve2	-20.36199
i_seve4	-129.31034
com_t	-46.21711
pcs_sd	-35.91320
mcs_sd	-109.70086

```
## [1] "21 unbalanced covariates remain"
```

Average Causal Effect

The treatment effect is at -19.58%. The confidence interval for the ATT is -32.43% to -6.75%. Since the interval does not contain zero, this is enough evidence that the treatment effect is in fact different from zero.

Logistic regression to the response variable

Table 3: Model Estimates for pg=1

	x
Estimate	-1.1584020
Std. Error	0.5829513
z value	-1.9871333
Pr(>|z|)	0.0469076

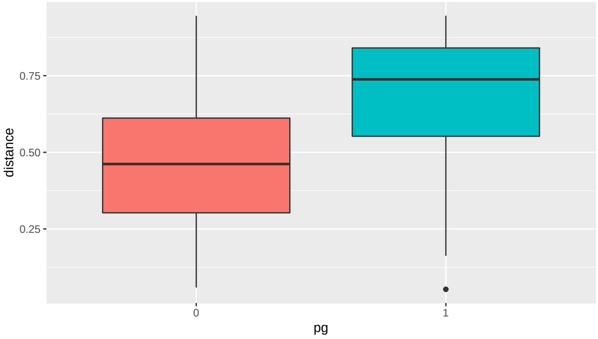
'pg' is significant in the model (has low p-value).

Causal Inference and confidence intervals

```
## Causal Odds ratio of pg=1 is 0.31 on exponential scale
## which means there is a ~69% decrease in satisfaction in people when they go to physician 2.
## The confidence interval for the ATT in exponetial scale is [0.1,0.98] which does not include 1.
## Hence, it can be reliable.
```

One-to-many matching

All records were now matched as there are enough control units.



The distributions are closer- better than last one, but there still is a large difference.

Table 4: Remaining Unbalanced covariates for technique 2

	Mean Diff.
i_age	-8405.340599
i_race1	-86.426426
i_race2	-27.213115
i_race4	-24.892704
i_educ2	-7.777778
i_drug1	-47.341772
i_seve4	-568.965517
mcs_sd	-89.776042

```
## [1] "8 unbalanced covariates remain"
```

Checking causal effect due to one-to-many matched data

The effect is negative. The effect comes out to be -15.27% which is a decrease in chances of satisfaction. Since the confidence interval is -27.29% to -3.24% (does not contain 0), this can be a significant decrease in satisfaction.

Regression model on one-to-many matched data

Table 5: Model Estimates for pg=1

	x
Estimate	-0.7670806
Std. Error	0.3883627
z value	-1.9751656
Pr(>|z|)	0.0482494

The p-value for pg=1 reveals that the covariate is significant in the model.

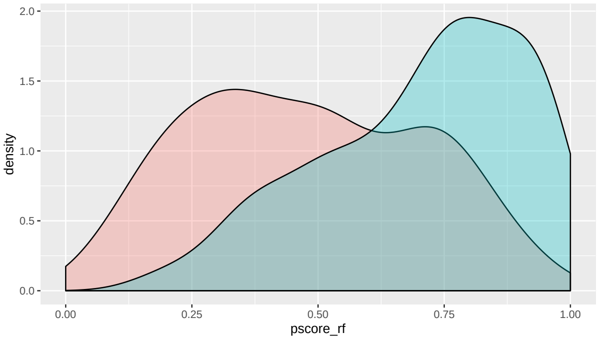
Causal Inference and confidence intervals

'pg' is significant in the model (has low p-value).

```
## Causal Odds ratio of pg=1 is 0.46 on exponential scale
## which means there is a ~54% decrease in satisfaction in people when they go to physician 2.
## The confidence interval for the ATT in exponetial scale is [0.22,0.99] which does not include 1.
## Hence, it can be reliable.
```

Part 3

Propensity Scores using Random Forest

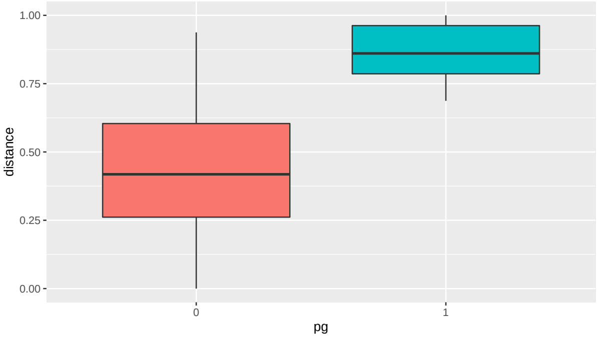


There seem to be some outliers.

Removing outliers

```
## [1] "27 outliers were dropped."
```

Control units are again fewer than treatment (test) units. Not all treated units are matched.



The difference is again magnified. It is similar to the distributions obtained from logistic regression and one-to-one matching.

```
## [1] "20 unbalanced covariates remain"
```

Table 6: Remaining Unbalanced covariates for technique 3

	Mean Diff.
distance	-47.73050
i_age	-109.09091
i_sex0	-55.17241
i_sex1	-55.17241
i_race1	-59.43888
i_race2	-59.20344
i_race3	-27.14681
i_race4	-2085.71429
i_educ2	-39.09091
i_educ3	-32.40385
i_educ4	-50.09197
i_educ6	-20.20202
i_insu2	-119.24258
i_insu3	-56.12245
i_drug1	-92.05021
i_seve1	-42.99065
i_seve2	-415.73034
i_seve4	-17.15161
com_t	-42.72570
pcs_sd	-45.18363

Checking causal effect due to one-to-one matched data

The effect shown here is also negative. The effect comes out to be -18.08% which is a decrease in chances of satisfaction. Since the confidence interval is -31.29% to -4.88% (does not contain 0), this can be a significant decrease in satisfaction.

Regression model on one-to-one matched data and propensity scores from random forest

Table 7: Model Estimates for pg=1

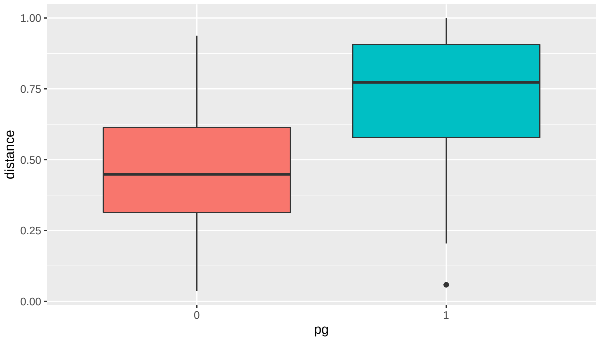
	x
Estimate	-0.9757829
Std. Error	0.5583528
z value	-1.7476097
Pr(>|z|)	0.0805316

Treatment variable 'pg' is not statistically significant anymore. neither is the propensity score.

Causal Inference and confidence intervals

```
## Causal Odds ratio of pg=1 is 0.38 on exponential scale
## which means there is a ~62% decrease in satisfaction in people when they go to physician 2.
## However, the confidence interval in exponetial scale is [0.13,1.13] which includes 1 and
## thus, the results cannot be trusted here.
```

One-to-many and Random Forest



The distributions are closer than they were in the on-on-one matched datasets. There seems to be a higher overlap in the distributions obtained from one-to-many matches from logitic regression.

Table 8: Remaining Unbalanced covariates for technique 4

	Mean Diff.
i_age	-130.588235
i_race4	-8020.000000
i_insu5	-136.551724
i_drug1	-47.615063
i_seve1	-1.765317
i_seve2	-54.157303
i_seve4	-23.062787
mcs_sd	-343.457663

```
## [1] "8 unbalanced covariates remain"
```

Checking causal effect due to one-to-many matched data and random forest

The effect is again negative. The effect comes out to be -13.99% which is a decrease in chances of satisfaction. Since the confidence interval is -25.80% to -2.17% (does not contain 0), this can be a significant decrease in satisfaction.

Regression model on one-to-many matched data and propensity scores from random forest

Table 9: Model Estimates for pg=1

	x
Estimate	-0.6698383
Std. Error	0.3790712
z value	-1.7670516
Pr(>|z|)	0.0772196

‘pg’ is not significant in the model (has high p-value).

Causal Odds ratio and Confidence Intervals

```
## Causal Odds ratio of pg=1 is 0.51 on exponential scale
## which means there is a ~49% decrease in satisfaction in people when they go to physician 2.
## However, the confidence interval in exponetial scale is [0.24,1.08] which includes 1 and
## thus, the results cannot be trusted here.
```

Table 10: Technique Comparison

Technique	Unbalanced.covariates	Causal.Odds.ratio..exp.scale.	CI.lower.bound	CI.upper.bound
Logistic & One-to-one	21	0.3139875	0.1001621	0.9842864
Logistic & One-to-many	8	0.4643667	0.2169133	0.9941136
Random Forest & one-to-one	20	0.3768972	0.1261688	1.1258841
Random Forest & one-to-many	8	0.5117913	0.2434596	1.0758677

Part 4

Let us make a table for comparing the four techniques

The method of choice here is **one-to-many matching with logistic regression** because of the following reasons:

1. After balancing, only 8 covariates were unbalanced which increases the reliability of the results.
2. As compared to the on-on-one matching, the control units were sufficient in the one-to-many matching cases and thus, we did not lose a lot of data owing to fewer control units.
3. The effect of the treatment variable (pg) ovtained from the average effects without the model and the causal odds ratios obtained from the model are in sync for the logistic regression method with one-to-many matching.
4. Moreover, the 95% confidence interval obtained for the logistic regression model does not include one in the exponential scale which increases the reliability of the results.
5. The p-value of pg=1 in logistic regression is low which reveals that it is statistically significant. Hence, the results of this model will be more relevant.