

## Causal Inference

### Problem Set 2.

The data set in [http://www.biostat.jhsph.edu/~cfrangak/biostat\\_causal/asthma.txt](http://www.biostat.jhsph.edu/~cfrangak/biostat_causal/asthma.txt) ([http://www.biostat.jhsph.edu/~cfrangak/biostat\\_causal/asthma.dta](http://www.biostat.jhsph.edu/~cfrangak/biostat_causal/asthma.dta) for STATA format) is from a study to compare the quality of services provided by two physician groups for asthma patients in California. Specifically, for patient  $i$ , let  $Y_i(z)$  be the quality of service as judged by the patient (1= satisfactory, 0 not satisfactory), if the patient is to be served by physician group  $z$ ,  $z=1, 2$ . The patients who visit the two groups can differ, and a set of covariates is measured. The variables in the data are: “pg” ( treatment assignment)– physician group (categorical); “i.age”– age (continuous); “i.sex”– sex (binary); “i.educ”– education (categorical); “i.insu”– insurance status (categorical); “i.drug”– drug coverage status (categorical); “i.seve”– severity (categorical); “com.t”– total number of comorbidity; “pcs.sd”– standard physical comorbidity scale (continuous); “mcs.sd”– standard mental comorbidity scale (continuous); “i.aqoc” (outcome)– satisfaction status of patient (binary).

For problem 1 – 3, assume the treatment (physician group) assignment is ignorable conditional on all the pre-treatment variables, i.e.,  $(Y(1), Y(2)) \perp\!\!\!\perp Z | X$ .

1. Let  $\text{pr}(Y(z) = 1) = p_z$  be the fraction of patients who would be satisfied with the service provided if all patients were to be served by physician group  $z$  ( $z=1$  or  $2$ ). The target estimand is the average causal effect  $Q = p_1 - p_2$ . To estimate it:
  - (a) Estimate the propensity score  $e$  using a logistic regression with all pre-treatment variables entering in the model as main effects.
  - (b) Check if there is any observation 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, keep them; If many, discard them and report the number of the discarded observations.
  - (c) Construct five blocks with bounds defined by the quintiles of the propensity score distribution for the whole sample. Report for each block:
    - i. For each of the pre-treatment variable, as well as for the propensity score, the difference in means for group 1 and group 2, and the t-statistic for the difference in means.
    - ii. The average causal effect and its standard error.
  - (d) Estimate the average causal effect  $Q$  using the block averages obtained above. Also report a standard error.

2. Now the target estimand is the odds ratio (OR) of satisfaction for comparing group 1 vs. group 2, i.e.,  $p_1(1 - p_2)/p_2(1 - p_1)$ . Consider the following five possible but not necessarily right ways to estimate it:

(for (a)-(c), assume the model

$$\text{logit}(\text{pr}(Y_i(z) = 1|X_i = x)) = \alpha_0 + x\beta + z\delta \quad (1)$$

is the true underlying model of  $X, Y, Z$ , and denote the corresponding  $\text{pr}(Y(z) = 1|x) = f(x, z)$ .

- (a) Regress  $Y$  on  $X$  and  $Z$  using model (1), and calculate  $\exp(\hat{\delta})$ , where  $\hat{\delta}$  is the estimate of the coefficient of  $z$  in model (1).
- (b) Regress  $Y$  on  $X$  and  $Z$  using model (1). Within each group  $z(z = 1, 2)$ , calculate  $\hat{p}_z$  as the average of  $f(x, z)$  over the  $X$  in that group. Then calculate  $\hat{p}_1(1 - \hat{p}_2)/\hat{p}_2(1 - \hat{p}_1)$ .
- (c) Regress  $Y$  on  $X$  and  $Z$  using model (1). For each observation  $i$  in the two groups, calculate two probabilities  $\hat{p}_{zi} = f(x_i, z)$ , for  $z = 1, 2$ . Take  $\hat{p}_z$  as the average of  $p_{zi}$  over all observations (not just in group  $z$ ). Then calculate  $\hat{p}_1(1 - \hat{p}_2)/\hat{p}_2(1 - \hat{p}_1)$ .
- (d) Obtain the odds ratio of satisfaction for comparing group 1 vs. 2, for each of the five propensity score blocks of problem 1, and take the average.
- (e) Within each of the five propensity score blocks, obtain the probabilities of satisfaction for both groups. Take the estimated  $\hat{p}_1$  and  $\hat{p}_2$  as the block averages, and calculate  $\hat{p}_1(1 - \hat{p}_2)/\hat{p}_2(1 - \hat{p}_1)$ .

Do all the above five procedures and compare the results. Identify the correct one(s) and explain why. If you can, also report the standard error(s) of the correct estimate(s).

3. Now let  $W$  be the gender (index as “sex” in the data), and let the target estimand be  $E(Y(1)|W = 1) - E(Y(2)|W = 1)$ . Consider the following three possible but not necessarily right ways to estimate it:

- (a) Estimate the propensity score  $e$  with all pre-treatment variables including  $W$ ; then within the subgroup of  $W = 1$ , do subclassification with the propensity score as in problem 1 and average the estimates from all the strata of  $e$ .
- (b) Same as in (a), except that here we exclude  $W$  when estimating the propensity score  $e$ .
- (c) Within the subgroup of  $W = 1$ , estimate  $e$  with all pre-treatment variables except for  $W$ ; then do subclassification with  $e$  as in problem 1 and average the estimates from all the strata of  $e$ .

Do all the above procedures and compare the results. Identify the correct one(s) and explain why. Also report the standard error(s) of the correct estimate(s).

4. Assume income ( $I$ ) is a variable that is correlated with the potential outcomes even after adjusting for all the pre-treatment variables, in the sense that,  $\text{pr}(Y(z)|X) \neq \text{pr}(Y(z)|X, I)$ , for  $z = 1, 2$ . But income is not included in the dataset. Then to estimate the three estimands in problem 1 – 3, does any of the previous procedures remain generally valid? If yes, report which one(s) and explain why. If not, give condition(s) under which those procedures remain valid.