Exercises week 5.

We consider the same tabulated data as in week 4

Suppose we have a dataset where the values of binary confounder C, a binary exposure X and a binary outcome Y are measured. Below you find for each of the combinations of values of X, C and Y the number of observations in the dataset. So there are 80 observations with X=0, C=0 and Y=0, etc.

	X	C	Y	n
1	0	0	0	80
2	0	0	1	20
3	0	1	0	20
4	0	1	1	10
5	1	0	0	80
6	1	0	1	20
7	1	1	0	80
8	1	1	1	40

a. Calculate the probability to be treated for those with C=1 and those with C=0.
   - P(X=1| C=c) for c=0,1

   What are these probabilities?

b. Suppose you want to adjust for confounding by calculating propensity score weights. Calculate the weights for X=1,0 and C=1,0

c. Estimate $E\big(Y(x)\big)$, for x=0,1 using propensity score weighting
   - Estimate the ATE (risk difference)

d. Try to derive what the weights would be if you would want to estimate the treatment effect in the treated

2. Propensity score weighting

We use the dataset "rhc_exercise.Rdata on which we performed G-computation last week. This week we will account for confounding by using propensity score methods

The treatment is right heart catheterization (RHC), a diagnostic procedure used for critically ill patients. In the data it is held in the variable called treatment and is coded 1 if RHC was administered within 24 hours from admission, 0 otherwise. We are interested in its effect on death within 30 days.

Libraries

To run this exercise in R you will need to install the following libraries:

- cobalt
- survey

a. Load the RHC data into R . Fit the propensity model with the following set of variables: transhx, age, scoma1, hrt1, bili1, wtkilo1 ,cat1 , i.e. all variables in the dataset besides surv2md1 and aps1. Examine the output.

b. Consider whether this set of variables would lead to appropriate control for confounding by examining whether the two excluded variables, surv2md1 and aps1, should actually be excluded. By definition the first is associated with mortality, while the latter is also likely to be so. If they were also associated with treatment, excluding them would lead to important treatment group imbalances.

   Consider this by doing the following:
   - Generate a new variable called ps1 that holds the predicted values of the propensity of receiving RHC derived from the model fitted in a.;
   - Generate the corresponding weights ipw1 (remember that weights depend on the exposure status of each individual);

c. Examine whether the covariates are balanced between the two treatment groups, first as they are and then after reweighing the observations according to ipw1. You can check for balance as follows. First define a new data frame `covariates` with all variables for which you would like to asses balance:

```
 vars1 <- c("transhx","age","scoma1","hrt1","bili1","wtkilo1","cat1",
"surv2md1","aps1")
       covariates <-rhc[, vars1]
```

   Then use the function bal.tab from the package cobalt, which will calculate standardized differences between the treated and untreated patients:

```
  bal.tab(covariates, treat = rhc$treatment, weights = rhc$ipw1, method
= "weighting", un=TRUE)
```

   A so-called "Love-plot" can be obtained by typing:
   love.plot(covariates, treat = rhc$treatment, weights = rhc$ipw1,method = "weighting", binary = "std", threshold = .1)

   What do you conclude?

d. Add the two remaining variables to your propensity model: surv2md1 and aps1. Generate a new variable ps2 that holds the predicted values of the propensity of receiving RHC.
   Calculate the minimum and maximum of ps2 . Also examine the distribution of ps2 , by making a histogram separately for the two treatment What can you conclude about the positivity assumption?

e. Use the ps2 values to create the corresponding weights, ipw2. Check the covariate balance across treatment groups after weighing the two treatment groups using these new weights.

f. We are ready to estimate the Average Treatment Effect, using the weights ipw2. We will use the function weighted.mean. Therefore we need to recode the outcome variable rhc$death30 as a 0/1 variable, with 1 if a person died within 30 days and 0 is the person is alive after 30 days. Check by

```
table(rhc$death30 ) and
table(as.numeric(rhc$death30 ))
```

how the variable rhc$death30 is coded.

Calculate a new binary variable

```
rhc$death30.num <- as.numeric(rhc$death30)-1
```

g. Estimate $E\big(Y(1)\big)$ by calculating the weighted mean outcome in the treatment group (which is equal to the weighted proportion of patients who died within 30 days after receiving RHC) using the function weighted.mean .
Calculate also the weighted mean outcome in the untreated group and find the estimate for the ATE.

h. To calculate the standard errors, the weights should be taken into account. This can be done using software for survey methods and to perform a linear regression with death20.num as dependent and . Therefore install and load the package survey and type

```
d.w <- svydesign(~1, weights = rhc$ipw2, data = rhc)
fit.ipw <- svyglm(death30.num ~ treatment, design = d.w)
summary(fit.ipw)
```

Write down the estimated ATE and the 95% confidence interval. Conclusion?

i. Examine the weights ipw2, using summary. Are there any very large weights? What would happen if there are very large weights?

j. Calculate new weights which are truncated at the 99th percentile of ipw2. This can be done in R by

```
ipw2.p99 <-quantile(rhc$ipw2,0.99)
rhc$ipw2.trunc <- ifelse(rhc$ipw2>ipw2.p99, ipw2.p99, rhc$ipw2)
```

Estimate the ATE using these truncated weights. What do you observe?

3. Propensity score matching

To run this exercise in R you will need to install the following libraries:

- survey

a. We use the rch dataset. The function matchit of the package MatchIt can be used to match using propensity scores by :

```
match.out1 <- matchit(treatment ~ transhx + age + scoma1 + hrt1 + bili1 +
    wtkilo1 + cat1 + surv2md1 + aps1, data=rhc,
```

```
                    method = "nearest", distance = "glm")
```

You can check what has been done by typing

```
match.out1
```

and a summary of the matching procedure by

```
summary(match.out1)
```

Which treatment effect is being targeted, the ATE or the ATT? How many individuals are matched? Consider the standardized mean differences, before and after matching. Are the baseline covariates balanced after matching?

b. Closer matching can be performed by adding the option caliper=0.1. What do you observe regarding the number of patients matched and the balance.

c. If you are happy with the results of b. (Note that there are many different matching options, but we will not explore them further here), a data.frame with the matched subsample can be obtained by :

```
match.data <- match.data(match.out2)
```

View the matched data. The variable `subclass` in the matched data contains the pair number. Sort by this variable and examine if within the pairs the propensity score values are similar.

d. The matched dataset can be analyzed as a randomized controlled data (to calculate standard errors the pairing should be taken into account). Calculate the percentage of people that died in the two treatment groups and calculate the difference. Compare the results with Exercise 2. .