Exercises week 4.

Exercise 4.1

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 observed risk on Y=1 for those with X=1 and for those with X=0. Also calculate the difference in risk. Why can you not interpret this risk as a causal effect of X on Y
b. Estimate the following quantities using standardization via outcome modelling
   - $E(C = c,\ X = x)$, for c=0,1, x=0,1
   - $P(C = 1)$ and $P(C = 0)$
   - Estimate $E\big(Y(x)\big)$, for x=0,1
   - Estimate the ATE (average treatment effect, expressed as risk difference) and the ATT (average treatment effect in the treated)

Exercise 4.2

In this exercise we will use data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT).

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 will study its effect on death within 30 days.

The data for this exercise are in the dataset "rhc_exercise.Rdata". It holds the following variables:

Variable name	Variable label
id	Identifier
treatment	Right heart catheterization (RHC) within 24 hours

death30	Death within 30 days
transhx	Transfer (>24 hr) from another hospital
age	Age at admission (years)
surv2md1	Support model which estimates at baseline the 2m surv prob
scoma1	Glasgow coma score
hrt1	Heart rate
bili1	Bilirubin (g/l)
wtkilo1	Weight (kg)
cat1	Primary disease category
aps1	Acute Physiology Score

Libraries

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

- stdReg

a. Load the RHC data into R using load(file="rhc_exercise.Rdata"). Familiarize yourself with the data. In particular examine the frequency of RHC, held in the variable called treatment, and of the outcome variable called death30, which holds the information on whether the patient died within 30 days.
b. Fit an outcome model for the relation between treatment (exposure), and the outcome death30. The outcome is binary, so you can use a logistic regression model. Calculate from this model the odds ratio of death for patient with treatment=RHC versus treatment = no RHC. Does this odds ratio have a causal interpretation?
c. We assume that the following variables have been identified as potential confounding variables: `transhx, age, surv2md1, scoma1, hrt1, bili1, wtkilo1, cat1, aps1.` Fit an outcome model for the relation between treatment (exposure), the confounders and the outcome death30 without interactions.
   Study the output of the model. What is the estimated adjusted odds ratio for the effect of being treated on death within 30 days? What do you observe if you compare the result to exercise c?
d. Fit a second logistic model adding interactions of treatment with each of the confounders.
e. To estimate $Y(1)$, the potential outcome if the patient was receiving RHC, create a dataset where all observations have `treatment ==”RHC”.` Predict in this dataset for every person the probability that Y=1 .
f. Calculate the average of the predicted probabilities. How do you interpret this average?
g. Repeat for `treatment ==”no RHC”`.
h. Estimate the average treatment effect (ATE) and the causal odds ratio using the answers at f. and g.
i. What is the difference between this odds ratio and the odds ratio calculated in c.?
j. Alternatively to perform regression standardization with the model fitted in (d), we can use the function `stdGlm` from the `stdReg` package (Sjölander, Eur J Epidemiol (2016) 31:563–574) . The function takes a fitted model as input, with the data frame that was used to fit the model, and standardizes to the confounder distribution in the data frame. It will also calculate analytic standard errors x. This is done by installing the package, and by typing

```
fit.std <- stdGlm(fit= …, data= .. , X=”..”)
```

where you have to specify after `fit=` the name of the outcome regression model , `data=` the dataset, and X= the name of the exposure variable. Type

```
summary(fit.std)
```

to obtain the estimated potential outcomes, and

```
summary(fit.std, contrast='difference', reference="no RBC")
```

to obtain the estimate of the causal risk difference with 95% CI.

k. Calculate the average treatment effect in the treated(ATT), by using the function stdGlm, with as extra option `, subsetnew=treatment=="RBC" .`