Practical 2

Part 1 or 3: Assumptions in causal inference

(a) We investigate the effect of a specific motivational training on the performance of students in elementary school. We recruit several school classes and in each class, we randomly allocate one half of the students to the motivational training, while the other half serves as the control group.

(i) It is likely that the students in the motivation group share what they learned in the motivational training with the other students. Which assumption is violated?

(ii) Name sources of (unobserved) confounding in this study, if any.

(b) We are interested in the effect of antihypertensives (drugs that lower the blood pressure) on the incidence of chronic heart disease (CHD). Our data base contains information on prescribed antihypertensives, CHD and further health-related as well as demographic information for a large number of patients over time.

(i) There are several different classes of antihypertensives that lower the blood pressure by different means. Some of them are never prescribed to patients with specific comorbidities (e.g. diuretics are contra-indicated in patients suffering from chronic obstructive pulmonary disease). Which assumption(s) may be violated? How could the research question be modified?

(ii) Name sources of (unobserved) confounding in this study, if any.

Part 3: Adjustment using standardisation and IPTW by hand

The following examples are taken from the book *Causal Inference: What If*, by Miguel Hernán and Jamie Robins.

	$Y(0)$	$Y(1)$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0
Σ	10	10

	A	C	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	1	0	0
Poseidon	1	0	0
Hera	1	0	0
Zeus	1	0	1
Artemis	0	1	1
Apollo	0	1	1
Leto	0	1	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0
Σ	13	12	10

(a) The **right**-hand table contains data for twenty Greek gods waiting for heart transplants. The variable A describes the treatment ($A = 0$ no transplant received, $A = 1$ transplant received). The outcome of interest is survival five years after treatment ($Y = 0$ alive, $Y = 1$ dead). By divine revelation we know the potential outcomes for each of the gods, shown in the left-hand table.

(i) Is there an individual causal effect for Zeus?

(ii) What is the average causal effect? (causal risk difference)

(b) The **right**-hand table contains related data from an observational study. Shown are the actual treatment A, the actual outcome Y and a risk factor C. Assume that within strata of C, the treatment A can be considered like randomised. Fill in the missing numbers in the following (the gaps are labelled for easier reference).

(i) The crude (unadjusted) difference is
$\Delta = \hat{P}(Y = 1 \mid A = 1) - \hat{P}(Y = 1 \mid A = 0) = (1)$.

(Compare this to the average causal effect from (a,ii) – association is not causation!)

(ii) Estimate the average causal effect via standardisation (see slide 46).
Subgroup with $C = 0$:
The estimated risk of dying when treated is
$\hat{P}(Y = 1 \mid A = 1, C = 0) =$(2).
The estimated risk of dying when untreated is
$\hat{P}(Y = 1 \mid A = 0, C = 0) =$(3).

Subgroup with $C = 1$:
The estimated risk of dying when treated is
$\hat{P}(Y = 1 \mid A = 1, C = 1) =$(4).
The estimated risk of dying when untreated is
$\hat{P}(Y = 1 \mid A = 0, C = 1) =$(5).

Standardisation:
The estimated standardised risk of dying for the whole population when treated is
$\hat{P}(Y(1) = 1)$
$= \hat{P}(Y = 1 \mid A = 1, C = 0) \cdot \hat{P}(C = 0) + \hat{P}(Y = 1 \mid A = 1, C = 1) \cdot \hat{P}(C = 1)$
=(6).
The estimated standardised risk of dying for the whole population when untreated is
$\hat{P}(Y(0) = 1)$
$= \hat{P}(Y = 1 \mid A = 0, C = 0) \cdot \hat{P}(C = 0) + \hat{P}(Y = 1 \mid A = 0, C = 1) \cdot \hat{P}(C = 1)$
=(7).
The estimated average causal effect is
$\hat{ACE} = \hat{P}(Y(1) = 1) - \hat{P}(Y(0) = 1) =$(8).
(Compare this to the average causal effect from (a,ii).)

(iii) Estimate the average causal effect via inverse probability of treatment weighting.
The estimated probability of treatment given $C = 0$ is
$\hat{P}(A = 1 \mid C = 0) =$(9).
The estimated probability for having $C = 0$, being treated and dying is
$\hat{P}(C = 0, A = 1, Y = 1) =$(10).
Weighting with the inverse probability of treatment given $C = 0$ yields
$\frac{\hat{P}(C=0,A=1,Y=1)}{\hat{P}(A=1|C=0)} =$(11).
The estimated probability for having $C = 0$, not being treated and dying is
$\hat{P}(C = 0, A = 0, Y = 1) =$(12).
Weighting with the inverse probability of no treatment given $C = 0$ yields
$\frac{\hat{P}(C=0,A=0,Y=1)}{\hat{P}(A=0|C=0)} =$(13).

The estimated probability of treatment given $C = 1$ is
$\hat{P}(A = 1 \mid C = 1) =$(14).
The estimated probability for having $C = 1$, being treated and dying is
$\hat{P}(C = 1, A = 1, Y = 1) =$(15).
Weighting with the inverse probability of treatment given $C = 0$ yields
$\frac{\hat{P}(C=1,A=1,Y=1)}{\hat{P}(A=1|C=1)} =$(16).

The estimated probability for having $C = 1$, not being treated and dying is $\hat{P}(C = 1, A = 0, Y = 1) =$(17).
Weighting with the inverse probability of no treatment given $C = 1$ yields $\frac{\hat{P}(C=1,A=0,Y=1)}{\hat{P}(A=0|C=1)} =$(18).

The estimated weighted risk of dying for the whole population when treated is $\hat{P}(Y(1) = 1) = \frac{\hat{P}(C=0,A=1,Y=1)}{\hat{P}(A=1|C=0)} + \frac{\hat{P}(C=1,A=1,Y=1)}{\hat{P}(A=1|C=1)} =$(19).
The estimated weighted risk of dying for the whole population when not treated is
$\hat{P}(Y(0) = 1) = \frac{\hat{P}(C=0,V=0,Y=1)}{\hat{P}(A=0|C=0)} + \frac{\hat{P}(C=1,A=0,Y=1)}{\hat{P}(A=0|C=1)} =$(20).
The estimated average causal effect is
$\hat{ACE} = \hat{P}(Y(1) = 1) - \hat{P}(Y(0) = 1) =$(21).
(Compare this to the average causal effect from (a,ii).)

Part 3: (Augmented) IPTW

(a) Consider an exposure A (binary), a sufficient set of covariates X (discrete), and an outcome Y; assume $\pi(x) = P(A = 1|X = x)$ is known.
Show that $\mathbb{E}[AY/\pi(X)] = \mathbb{E}[Y|\text{do}(A = 1)]$ or $\mathbb{E}[AY/\pi(X)] = \mathbb{E}Y(1)$. State clearly on what assumptions you rely. (Trick: use iterated conditional expectation.)

(b) In the same situation as above:

(i) show that for

$$\hat{\mu}_1 = m(X) + \frac{A}{\pi(X)}(Y - m(X))$$

we have: $\mathbb{E}\hat{\mu}_1 = \mathbb{E}[Y|\text{do}(A = 1)]$ if either $m(X) = \mathbb{E}[Y \mid A = 1, X]$ or if $\pi(x) = P(A = 1|X = x)$.

(ii) What does the corresponding $\hat{\mu}_0$ have to look like?

Part 3: Causal effect estimation with R

You can find instructions to install all required packages and data sets here:

`https://github.com/bips-hb/APTS_Causal_Inference`

Consider the 'Rotterdam breastcancer' data on 2982 women diagnosed with primary breast cancer – this is in the file `bcrot`. This was an observational study and interest lies in estimating the average causal effect of hormonal therapy (`hormon` binary yes/no) on a synthetic outcome `qol` (health related quality of life, continuous measure). The following covariates are included: `age` = age at diagnosis; `nodes` = number of positive lymph nodes (exp-transformed to `enodes`); `pr_1` = Progestorone receptors (fmol/l) (log-transformed).

(a) Compare descriptively QOL for those who do and do not take hormonal therapy.

(b) Carry out some descriptive analyses (including overlap plots using a flexible model for the propensity score) to investigate the positivity assumption. Convince yourself that it makes sense to restrict the sample to those $\geq$40 years and with at least one node. Create a dataset containing this restricted sample.

(c) Use IPTW to estimate the average causal effect in the restricted population (either 'by hand' or with `R` package `ipw`); investigate extreme weights and consider covariate balance of the re-weighted sample (`R` package `cobalt`, function `love.plot`).

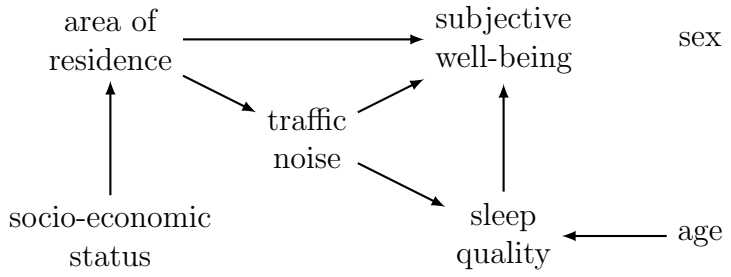
(d) Use regression standardisation to estimate the average causal effect in the restricted population (`R` package `stdReg`) choosing a flexible outcome model.

(e) Finally, use `AIPTW` to estimate the average causal effect in the restricted population (R package `AIPW` with default settings and super-learner). Investigate the covariance balance for the re-weighted sample.

(f) Compare all the above estimates including their precision (ensure that you take the behaviour of the IPT-weights into account).

Note that the covariates are from a real dataset, but the quality of life outcome in the above data is fictitious!

Part 4: Causal DAGs and regression



The figure above shows a fictional causal DAG.

(a) Which statements are correct, given the causal DAG above?

(i) In a regression of 'subjective well-being' on 'age' and 'sex', we expect both coefficients to be zero.

(ii) The coefficients in a regression of 'sleep quality' on 'traffic noise', 'subjective well-being' and 'age' have a causal interpretation.

(iii) The coefficients in a regression of 'subjective well-being' on 'traffic noise' and 'sleep quality' have a causal interpretation.

(iv) If we regress 'sleep quality' on 'age' and obtain a coefficient of about zero, we know that the DAG is wrong.

(b) Suppose we are interested in the total causal effect of 'sleep quality' on 'subjective well-being.' Which confounders should we adjust for? How could the desired effect be obtained?