Homework 8

Due November 19 at 3:00.

Searching for solutions is forbidden.

(1) Let $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$ and $A \in \{0, 1\}$. Assuming there is no unmeasured confounding. We have that, for $a \in \{0, 1\}$,

$$\mathbb{E}[Y(a)] = \int \mu(x, a)p(x)dx$$

where $\mu(x, a) = \mathbb{E}[Y|X = x, A = a] = \int yp(y|x, a)dy$. Show that

$$\mathbb{E}[Y(a)] = \mathbb{E}\left[\frac{YI(A = a)}{\pi(A|X)}\right]$$

where $\pi(a|x) = P(A = a|X = x)$. Here, $I(A = a) = 1$ if $A = a$ and $I(A = a) = 0$ otherwise.

Hint: You may take X and Y to be discrete if you wish. In that case, the integrals become sums.

Explain how would you use this formula to estimate $\mathbb{E}[Y(a)]$.

(2) Download the dataset *SAheart.csv*. The variables are

sbp	systolic blood pressure
tobacco	cumulative tobacco (kg)
ldl	low density lipoprotein cholesterol
adiposity	
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	
alcohol	current alcohol consumption
age	age at onset
chd	response, coronary heart disease

(Remove the first column which is just row numbers.)

(a) Examine the data using exploratory data analysis (i.e. plots).

(b) Do a logistic regression of `chd` on the other variables. Summarize the results.

(c) Suppose we want to estimate the causal effect of Age on chd. Assume we have measured all confounding variables. Estimate the causal effect. To do this, let Y denote chd, let A denote age and let X denote the other variables. We want to estimate the function $\psi(a) = \mathbb{E}[Y(a)]$, where $a \in \mathbb{R}$. We can restrict attention to $15 \leq a \leq 64$ since our data are in that range. Then recall that the plugin estimator is

$$\widehat{\psi}(a) = \frac{1}{n}\sum_i \widehat{\mu}(X_i, a)$$

where $\widehat{\mu}(x, a)$ is your estimate of $\mu(x, a) = \mathbb{E}[Y|X = x, A = a]$ which you have from the logistic regression. Plot the function $\widehat{\psi}(a)$.

(d) We would like to get a confidence interval for $\psi(a)$. To do this, you can use the bootstrap which works as follows: Draw a sample of size n from your original data. For example:

```
n = length(Y)
I = sample(1:n,size=n,replace=TRUE)
XX = X[I,]
YY = Y[I]
AA = A[I]
```

After you draw the bootstrap sample, estimate $\psi(a)$ again using the bootstrap sample. Repeat this process $B = 1000$ times to get $\widehat{\psi}_1(a), \ldots, \widehat{\psi}_B(a)$. For each a, let $\ell(a)$ be the 2.5 percentile of $\widehat{\psi}_1(a), \ldots, \widehat{\psi}_B(a)$ and let $u(a)$ be the 97.5 percentile of $\widehat{\psi}_1(a), \ldots, \widehat{\psi}_B(a)$. This gives a 95 percent confidence band for $\psi(a)$. Plot $\widehat{\psi}(a)$ along with $\ell(a)$ and $u(a)$.