36-402 Homework 10

Eu Jing Chua
eujingc

April 14, 2019

Question 1

We can check if the deltas are correct by adding them back to the pre-test values and seeing if they match the post-test values.

Table 1: Check for Pre-test + Delta = Post-test

	x
let	TRUE
body	TRUE
form	TRUE
numb	TRUE
relat	TRUE
clasf	TRUE

Question 2

Q2 a)

Table 2: Estimates for deltalet

	Estimate	SE
Regular watchers mean	13.220	0.810
Irregular watchers mean	2.481	0.918
Difference in means	10.739	1.224

Q2 b)

In order for this difference in means to be a sound estimate of the causal effect of switching from, there must be no other confounding sources that affect the subjects' knowledge of letters and whether they are regular watchers or not. This may not be realistic, as other variables such as age and social background could affect their knowledge of letters. We could test this by using a linear regression model of `deltalet` against `regular`, TODO

Question 3

Q3 a)

Table 3: Coefficients and SE of linear regression

	Coefficient	SE
(Intercept)	-5.3100	5.1800
factor(regular)1	8.0500	1.7100
factor(site)2	7.4900	2.1300
factor(site)3	-4.0200	1.7000
factor(site)4	-1.1900	1.8700
factor(site)5	1.4200	2.6000
factor(sex)2	1.0700	1.1600
age	0.1830	0.1150
factor(setting)2	0.2070	1.5100
factor(encour)1	0.9680	1.6300
peabody	-0.0145	0.0457
prelet	-0.5360	0.0984
prebody	0.0524	0.1360
preform	0.3870	0.2660
prenumb	0.1860	0.1300
prerelat	-0.0123	0.3080
preclasf	-0.0707	0.2270

Q3 b)

`id` should not be included in the regression as it is simply the ID number of the subject, having no relationship at all to the study besides identifying subjects.
`viewcat` should not be included too as the other covariate `regular` is a direct indicator of `viewcat`. Since `regular` is directly derived from `viewcat`, including both would be redundant and introduce problems with highly correlated covariates in linear regression.
Similarly, we exlude all the `post` variables as it is essentially the same as what we want to predict, as the `post` variables are the result of `pre` variables added with the `delta` variables. If we already knew the `post` variables, we would not be predicting anything useful or new.

Q3 c)

Someone who only took 401 might report that the average effect of making a child become a regular watcher of Sesame Street is an increase of 8.05 in score of the letter test.

Q3 d)

To infer the causal effect of becoming a regular watcher of Sesame Street on the change in score of the letter test based on the above model, we would first need to assume there are no other confounding sources between the two variables. Additionally, we also need to assume that all the additional covariates we are are controlling do not create new confounding sources by controlling for them. This is plausible but highly unlikely as including everything blindly increases the chances of creating new confounding sources.

Question 4

Q4 a) The set of variables are `setting` and `site`.

Q4 b) Using a kernel regression with cross-validated bandwidths,

Table 4: Average effect of regular watching

	x
Average treatment effect	8.6600
SE	0.0283

Question 5

Q5 a) Now just `prelet` satisfies the backdoor criterion. This is because all backdoor paths from `regular` to `deltalet` now pass through a chain where `prelet` is in the middle, so blocking this bath would block all backdoor paths. In the previous graph, there was no path from `regular` to `deltalet` through prelet, hence this was not possible.

Q5 b) The previous set of variables are no longer sufficient to satisfy the backdoor criterion, as there is still an open backdoor path from `regular` ← `U` → `prelet` → `deltalet`.

Q5 c) Using a kernel regression with cross-validated bandwidths,

Table 5: Average effect of regular watching

	x
Average treatment effect	10.700
SE	0.213

Q5 d)

In figure 1, `regular` $\perp\!\!\!\perp$ `peabody` | `setting`, `site` but not in figure 2. In both figures, `peabody` is only connected to `U`. In figure 1, blocking `setting` and `site` would block all paths from `regular` to `U`, but in figure 2 there is still an extra direct dependence of `regular` on `U`.
In figure 2, `deltalet` $\perp\!\!\!\perp$ `peabody` | `regular`, `prelet` but not in figure 1. In figure 2, `deltalet` is only connected to `regular` and `prelet`, so blocking these would block all paths to `U` and hence `peabody`. However in figure 1, `deltalet` has a dependence on `U` directly.

Question 6

Q6 a)

Table 6: Estimates of effect of regular watching on deltalet

	Estimate	SE
Naive	10.739	1.224
Linear Reg. with All	8.055	1.714
Control for setting and site	8.663	0.028
Control for prelet	10.672	0.213

Q6 b)

The naive estimate is compatible with controlling for prelet, while the linear regression with all covariates is more loosely compatible with controlling for setting and site.
The estimate from controlling for setting and site seemed the most trustworthy, as it has the smallest standard error. Its assumptions also make sense, as social background might have psychological impacts on learning and the site of watching might have an impact on the focus level.

Question 7

Q7 a)

We factor the joint probability according the original graph as

$$\Pr(Y = y, X = x', T = t, V = v) = \Pr(X = x' \mid T = t) \Pr(Y = y \mid Par(Y)) \Pr(T = t \mid Par(T)) \Pr(V = v)$$

When we set $X = x$, we make a new graph where all the edges of X are removed. Thus $\Pr(X = x'|T = t)$ becomes $\Pr(X = x')$. Since X is set and no longer random, $\Pr(X = x) = 1$.

Thus for the new graph,

$$\begin{aligned}
&\Pr(Y = y, X = x', T = t, V = v \mid do(X = x)) \\
&= \Pr(X = x') \Pr(Y = y \mid Par(Y)) \Pr(T = t \mid Par(T)) \Pr(V = v) \\
&= \begin{cases} \Pr(Y = y \mid Par(Y)) \Pr(T = t \mid Par(T)) \Pr(V = v), & \text{if } x' = x \text{ so } \Pr(X = x') = 1 \\ 0, & \text{otherwise as } \Pr(X = x') = 0 \end{cases} \\
&= \begin{cases} \frac{\Pr(Y=y,X=x',T=t,V=v)}{\Pr(X=x'|T=t)} & \text{if } x' = x \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

Q7 b)

$$\begin{aligned}
\Pr(Y = y, X = x', T = t, V = v \mid do(X = x)) &= \begin{cases} \frac{\Pr(Y=y,X=x,T=t,V=v)}{\Pr(X=x'|T=t)} & \text{if } x' = x \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{\Pr(Y=y,V=v|X=x,T=t) \Pr(X=x,T=t)}{\Pr(X=x'|T=t)} & \text{if } x' = x \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \Pr(Y = y, V = v \mid X = x, T = t) \Pr(T = t) & \text{if } x' = x \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \Pr(Y = y, X = x, T = t, V = v \mid X = x, T = t) \Pr(T = t) & \text{if } x' = x \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

Q7 c)

$$
\begin{aligned}
\Pr(Y = y \mid do(X = x)) &= \sum_{x'} \sum_{t} \sum_{v} \Pr(Y = y, X = x', T = t, V = v \mid do(X = x)) \\
&= \sum_{x'} \sum_{t} \sum_{v} \mathbb{I}_{\{x'=x\}} \Pr(Y = y, X = x, T = t, V = v \mid X = x, T = t) \Pr(T = t) \\
&= \sum_{t} \sum_{v} \Pr(Y = y, V = v \mid X = x, T = t) \Pr(T = t) \\
&= \sum_{t} \Pr(T = t) \sum_{v} \Pr(Y = y, V = v \mid X = x, T = t) \\
&= \sum_{t} \Pr(T = t) \Pr(Y = y \mid X = x, T = t) \\
&= \sum_{t} \Pr(Y = y \mid X = x, T = t) \Pr(T = t)
\end{aligned}
$$