Answers to inverse probability of treatment weighting exercise

November 21, 2018

Throughout this exercise we will be addressing the question: what is the causal effect of quitting smoking on weight change between 1971 and 1982? We will be estimating this effect using data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The dataset and codebook are available from the website for the Causal Inference book by Miguel Hernan and James Robins: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Begin by loading the dataset into SAS.

Importing data

```
libname causinf "path to data";

/* some preprocessing of the data */
data nhefs;
set causinf.nhefs;
match=1;
cens= (wt82 eq .);
run;

data nhefs_nmv;
set nhefs;
if wt82 ne .; * provisionally ignore subjects with missing values for weight in 1982;
run;
```

Q1. Our two variables of interest are the exposure *qsmk*, quit smoking between 1st questionnaire and 1982, and the outcome *wt82_71*. For the purposes of these exercises, let's assume that the only confounders we are worried about are *sex*, *race*, *age*, *education*, *smokeintensity*, *smokeyrs*, *exercise*, *active*, and *wt71*. Run some descriptive statistics that you would normally include in a table 1 (the average value or proportion of confounders in each of the exposure groups). On the last page there are two tables to help you keep track of the of balance across covariates from different models and the estimates themselves from different models. You can either fill them in by hand or create a similar table in another program.

```
proc means data= nhefs_nmv;
class qsmk;
var age wt71 smokeintensity smokeyrs education active exercise;
run;
proc freq data= nhefs_nmv;
table qsmk*(sex race education exercise active) / nopercent nocol;
run;
```

Q2. Use outcome regression to estimate the crude and adjusted effect of *qsmk* on *wt82_71*.

```
proc genmod data= nhefs_nmv;
model wt82_71= qsmk;
estimate 'Smoking cessation' intercept 1 qsmk 1;
estimate 'No smoking cessation' intercept 1 qsmk 0;
run;
quit;

proc genmod data= nhefs_nmv;
class seqn;
model wt82_71= qsmk sex race age education
```

```
smokeintensity smokeyrs
exercise active wt71;
estimate 'Smoking cessation' intercept 1 qsmk 1;
estimate 'No smoking cessation' intercept 1 qsmk 0;
repeated subject=seqn / type=ind;
run;
quit;
```

Q3. Now we will start estimating the weights. Estimate a regression where the exposure, *qsmk*, is the dependent variable in your model and the confounders are the independent variables. Use SAS to output a variable that is the probability of quitting smoking in each observation.

```
proc logistic data= nhefs_nmv descending;
ods exclude ClassLevelInfo ModelAnova Association FitStatistics GlobalTests;
class exercise active education;
model qsmk = sex race age education
smokeintensity smokeyrs
exercise active wt71;
output out=est_prob p=p_qsmk;
run;
```

Q4. Create a variable that is the probability that each observation receives the exposure, *qsmk*, that they were observed to receive. In other words, among those who quit smoking ($qsmk = 1$), the value should be the probability that they would quit smoking given confounders, $P(qsmk = 1|L = l)$. Among those who did not quit smoking ($qsmk = 0$), the value should be the probability that they would not quit smoking give confounders, $P(qsmk = 0|L = l)$ which is equivalent to $(1 - P(qsmk = 1|L = l))$. Taking the inverse of the this variable will give you the IPTW weights. Plot your weights on a histogram and check their mean and standard distribution.

```
data nhefs_w;
set est_prob;

if qsmk=1 then w= 1/p_qsmk;
else if qsmk=0 then w= 1/(1-p_qsmk);

run;

proc univariate data=nhefs_w;
id seqn;
var w;
run;

proc univariate data=nhefs_w; var w; histogram; run;

proc univariate data=nhefs_w;
var p_qsmk;
where qsmk=1;
histogram;
run;

proc univariate data=nhefs_w;
var p_qsmk;
where qsmk=0;
histogram;
run;
```

Q5. These weights can now be used to create a pseudopopulation where the confounders have a different disitribution relative to exposure. Use the same functions you used in Q1 to create your table 1 but add the

'weight' option to use the weights you created in Q4. Has the balance between those who quit smoking and those who didn't quit smoking changed?

```
proc means data= nhefs_w;
class qsmk;
weight w;
var age wt71 smokeintensity smokeyrs education active exercise;
run;

proc freq data= nhefs_w;
weight w;
table qsmk*(sex race education exercise active) / nopercent nocol;
run;

proc genmod data= nhefs_w;
class seqn;
weight w;
model wt82_71= qsmk;
estimate 'Smoking cessation' intercept 1 qsmk 1;
estimate 'No smoking cessation' intercept 1 qsmk 0;
repeated subject=seqn / type=ind;
run;
quit;
```

Q6. One of the advanatages of the propensity scores is that you can play with the model to get the balance right between the two groups without worrying about p-hacking (because you aren't looking at the outcome model). Add some interactions or non-linear effects to the model you used in Q3, recalculate the weights in Q4 and use the new weights in Q5. Have you improved the balance between the exposure groups?

```
proc logistic data= nhefs_nmv descending;
ods exclude ClassLevelInfo ModelAnova Association FitStatistics GlobalTests;
class exercise active education;
model qsmk = sex race age age*age education
smokeintensity smokeintensity*smokeintensity smokeyrs smokeyrs*smokeyrs
exercise active wt71 wt71*wt71;
output out=est_prob p=p_qsmk2;
run;

data nhefs_w2;
set est_prob;

if qsmk=1 then w2= 1/p_qsmk2;
else if qsmk=0 then w2= 1/(1-p_qsmk2);

run;

proc means data= nhefs_w2;
class qsmk;
weight w2;
var age wt71 smokeintensity smokeyrs education active exercise;
run;
proc freq data= nhefs_w2;
weight w2;
table qsmk*(sex race education exercise active) / nopercent nocol;
run;

proc univariate data=nhefs_w2;
var w2;
```

```
histogram;
run;

proc univariate data=nhefs_w2;
var p_qsmk2;
where qsmk=1;
histogram;
run;

proc univariate data=nhefs_w2;
var p_qsmk2;
where qsmk=0;
histogram;
run;
```

Q7. Once you are satisfied with the balance between your exposure groups, estimate your outcome model using *qsmk* as the only independent variable but weight your model using the weights you created in Q5. How does the result compare to what you estimated in Q2?

```
proc genmod data= nhefs_w2;
class seqn;
weight w2;
model wt82_71= qsmk;
estimate 'Smoking cessation' intercept 1 qsmk 1;
estimate 'No smoking cessation' intercept 1 qsmk 0;
repeated subject=seqn / type=ind;
run;
quit;
```

Q8. Now we will calculate stabilized weights. Although stabilized weights will not change anything in this example, they can increase the precision of our estimates with time-varying exposures. To do this we simply multiply our weights by the average of the treatment received. In other words, among those who quit smoking, multiply the weights by the proportion of people who quit smoking in the entire sample. Likewise, among those who did not quit smoking, multiply the weights by the proportion of people who did not quit smoking in the entire sampple. Rerun the analysis in Q7.

```
proc logistic data=nhefs_nmv descending;
ods exclude ClassLevelInfo ModelAnova Association FitStatistics GlobalTests;
class exercise active education;
model qsmk = sex race age age*age education
smokeintensity smokeintensity*smokeintensity smokeyrs smokeyrs*smokeyrs
exercise active wt71 wt71*wt71;
output out=est_prob_d p=pd_qsmk;
run;
proc sort; by seqn; run;

proc logistic data=nhefs_nmv descending;
ods exclude ClassLevelInfo ModelAnova Association FitStatistics GlobalTests Oddsratios;
model qsmk = ;
output out=est_prob_n (keep= seqn pn_qsmk) p=pn_qsmk;
run;
proc sort; by seqn; run;

data nhefs_sw;
merge est_prob_d est_prob_n ;
by seqn;
if qsmk=1 then sw_a= pn_qsmk / pd_qsmk;
```

```
else if qsmk=0 then sw_a= (1-pn_qsmk) / (1-pd_qsmk);
run;

proc genmod data= nhefs_sw;
class seqn;
weight sw_a;
model wt82_71= qsmk;
estimate 'Smoking cessation' intercept 1 qsmk 1;
estimate 'No smoking cessation' intercept 1 qsmk 0;
repeated subject=seqn / type=ind;
run;
quit;
```

Table 1: Fill in this table with average value of the confounderns in each of the exposure groups. The results for weight2 will depend on the model chosen.

	Original		Weights1		Weights2		Std weights	
Variable	qsmk=0	qsmk=1	qsmk=0	qsmk=1	qsmk=0	qsmk=1	qsmk=0	qsmk=1
sex	53.40	45.41	51.30	50.50	51.22	51.08	51.22	51.08
race	14.62	8.93	13.18	13.15	13.18	13.41	13.18	13.41
age	46.20	42.80	43.68	44.05	43.62	43.69	43.62	43.69
education	2.68	2.79	2.72	2.75	2.72	2.74	2.72	2.74
smokeintensity	21.19	18.60	20.57	20.65	20.49	20.20	20.49	20.20
smokeyrs	24.08	26.03	24.62	24.93	24.58	24.54	24.58	24.54
exercise	1.18	1.25	1.19	1.19	1.19	1.18	1.19	1.18
active	0.63	0.69	0.65	0.65	0.65	0.65	0.65	0.65
wt71	70.30	72.35	70.76	70.60	70.80	70.66	70.80	70.66

Table 2: Fill in this table with the estimates from each different analysis.

Weights	Estimate	Low.CI	High.CI
None	2.54	1.66	3.42
Outcome regression	3.35	2.42	4.28
Weights1	3.34	2.31	4.36
Weights2	3.44	2.41	4.47
Std weights	3.44	2.41	4.47