**EPID824**

**IPW in time-dependent confounding**

Due: 03/29/21 before the beginning of the session

NAME: Maxwell Salvatore

The aims of this homework are to:

1. Identify a variable introducing time-dependent confounding in a causal structure and explain the reason.
2. Define a causal question to estimate the average overall effect of a time-varying treatment.
3. Calculate inverse probability weights in a time-dependent confounding scenario.
4. Pinpoint the implication of applying inverse probability weights to estimate average treatment effects in a time-dependent confounding situation.
5. Implement inverse probability weighting (IPW) on a dataset to control for time-dependent confounding.
6. Compare the results of biased conventional analyses with those controlled for time-dependent confounding through IPW.

Data and code for this homework are provided in SAS, but feel free to use any statistical software of your preference.

Birth weight is an important indicator of the maternal and child health status of a population, and a predictor of adverse health outcomes of infants. Identifying modifiable causes of birth weight is therefore a relevant research endeavor. Smoking during pregnancy has been associated with lower birth weight.

For this exercise, investigators are concerned with the effect of smoking through two consecutive pregnancies on the second offspring’s birth weight. Using a state-wide birth registry covering a 20 year period, they identified all women who had two consecutive singleton pregnancies resulting each on a live birth. Their question of interest is the effect of smoking (a time-varying exposure) in both pregnancies compared to not smoking on any of the pregnancies on birth weight of the second offspring.

From prior knowledge, they make the following assumptions (for simplicity, assume direct effects):

1. Smoking during the first pregnancy causes the first offspring to be born small-for-gestational age (SGA).
2. Smoking during the first or second pregnancies causes birth weight of the second offspring.
3. Smoking during the first pregnancy causes smoking during the second pregnancy.
4. SGA of the first offspring causes birth weight of the second offspring.
5. A traumatic birth outcome of first pregnancy (e.g. SGA) causes smoking status during second pregnancy.
6. A defective placentation gene is an unmeasured common cause of SGA in first pregnancy and birth weight in second pregnancy.
7. There is no additional confounding, selection, or information bias.

The dataset hw\_09.sas7bdat has birth registry information for the first two consecutive singleton pregnancies resulting in a live birth among 1,080,616 women. The variables with their coding are labeled as follows:

# Variable Type Label

1 id Num mother id (arbitrary)

2 a0 Num smoking 1st pregnancy (0=no, 1=yes)

3 a1 Num smoking 2nd pregnancy (0=no, 1=yes)

4 z1 Num sga 1st pregnancy (0=no, 1=yes)

5 y Num birth weight 2nd pregnancy, g (continuous)

1. Draw a DAG representing the causal system per the prior knowledge assumptions.

Diagram, text

Description automatically generated

1. Which variable is introducing time-dependent confounding in the estimation of the overall average effect of smoking during pregnancy (A) on birth weight of the second offspring (Y)?

As an undesired birth outcome, a first pregnancy resulting in birth small for gestational age (SGA1) is a time-varying confounder.

1. Why?

SGA1 is a common cause of a time-varying exposure (Smoking2) and the outcome of interest (SGA2).

1. Can the overall average effect of smoking during pregnancy (A) on birth weight of second offspring (Y) be identified using standard methods (restriction, stratification, matching, or conditioning in a regression model)? If so, explain how, if not, explain why.

No because SGA1 is both a confounder and a mediator.

1. A binary (dichotomous) exposure (e.g. smoking) assessed at two time points (e.g. two consecutive pregnancies) results in four possible “treatment trajectories”: unexposed in both pregnancies (A0=0, A1=0), exposed in first pregnancy only (A0=1, A1=0), exposed in second pregnancy only (A0=0, A1=1), or exposed in both pregnancies (A0=1, A1=1). The overall average effect of a time-varying treatment is often estimated as the outcome difference between “always vs. never treatment”, i.e. being exposed at all times (in this case A0=1, A1=1) and being unexposed at all times (e.g. A0=0, A1=0). Some call this contrast the “joint” effect of exposure at all time points.
2. Write:
3. The equation of a marginal structural model (MSM) that would allow you to estimate treatment effects in the presence of time-dependent confounding [hint: Daniel et al., equation (4), page 1592].
4. The contrast (combination of coefficients) that would allow you to estimate specifically the “always vs. never treatment” effect (average treatment effect).
5. Why is this model called marginal and why is it called structural?

The model is called marginal because it is the marginal distribution (unconditional on the time-varying confounder) of the potential outcome that is modelled. The model is called structural because it is a model for potential outcomes and not a model for an observed outcome.

1. In this specific exercise, is this model parametric or nonparametric? Why?

Because A0 and A1 are both binary, the right-hand side of the equation in 5.a.i. is saturated (there are as many parameters as there are potential outcomes), there is not parametric smoothing, and thus is a nonparametric model.

1. Estimate the average effect of smoking during pregnancy (A) on birth weight of the second offspring using conventional linear regression, through fitting two separate “naïve” models. For each model: i) report the average effect point estimate and 95% CI, ii) describe the association in words, iii) state whether the estimate has a causal interpretation or not and why.
2. Model 1: ignoring Z1

*E*(*Y* | *A*0, *A*1) = *α*int + *α*0*A*0 + *α*1*A*1 + *α*01*A*0*A*1

**proc** **genmod** data=*yourlibrary*.hw\_09;

title 'CRUDE MODEL';

model y = a0 a1 a0\*a1;

estimate "always vs. never" a0 **1** a1 **1** a0\*a1 **1**;

**run**;

1. The average effect estimate for always smoking during pregnancy and birthweight of second offspring is -227.97 (95% CI: -231.20, -224.74).
2. Always smoking during pregnancy is associated with a 228 (95% CI: 225, 231) gram lower gestational birthweight of the second offspring, on average.
3. Since this is a conventional regression model that does not include potential outcomes nor adjust for confounders, it does not have a causal interpretation.
4. Model 2: conditioning on Z1

*E*(*Y* | *A*0, *A*1, *Z*1) = *β*int + *β*0*A*0 + *β*1*A*1 + *β*01*A*0*A*1 + *βzZ*1

**proc** **genmod** data=*yourlibrary*.hw\_09;

title 'Z1-ADJUSTED MODEL';

model y = a0 a1 a0\*a1 z1;

estimate "always vs. never" a0 **1** a1 **1** a0\*a1 **1**;

**run**;

1. The average effect estimate for always smoking during pregnancy and birthweight of second offspring is -198.81 (95% CI: -201.96, -195.65).
2. Conditional on low gestational birthweight of the first offspring, always smoking during pregnancy is associated with a 198 (95% CI: 196, 202) gram lower gestational birthweight of the second offspring, on average.
3. Since this is a conventional regression model, it does not have a causal interpretation.
4. IPW of MSM allows to simulate a situation in which exposure level is marginal to post-baseline confounders. This allows unbiased estimation of the exposure effect parameters through a re-weighted version of conventional regression. Perform IPW and specify a MSM to estimate the average effect of smoking during pregnancy on birth weight of second offspring.
5. Compute stabilized weights. The denominator is each person’s probability that they experience their particular exposure trajectory, conditional on their covariate history (hint: general equation under Daniel et al. section 4.2.3, page 1599; and equation under 4.2.5 for application to a scenario comparable to this exercise). The numerator is the marginal probability of each exposure category.
6. Manually calculate the stabilized weights (sw) for subjects with A0=0, Z1=0, A1=1 (hint: Daniel et al. example, page 1600). To facilitate the computation, obtain the sample sizes for each exposure trajectory in the dataset. Write out the notation involving the computation and perform it. Show calculations.

**proc** **freq** data=*yourlibrary*.hw\_09;

tables a0\*z1\*a1 / list nopercent nocum;

**run**;

1. Compute stabilized weights for all subjects using the code provided. Report the weights for each exposure/covariate trajectory. You may check that your calculation in i. is correct

**proc** **logistic** data=*yourlibrary*.hw\_09 desc noprint;

model a0=;

output out=b0 (keep=id p\_num\_a0) p=p\_num\_a0; \* P(A0=1);

**run**;

**proc** **logistic** data=*yourlibrary*.hw\_09 desc noprint;

model a1=;

output out=b1 (keep=id p\_num\_a1) p=p\_num\_a1; \* P(A1=1);

**run**;

**proc** **logistic** data=*yourlibrary*.hw\_09 desc noprint;

model a1=z1 a0 z1\*a0;

output out=b2 (keep=id ip\_0 ip\_1) predprobs=i; \* P(A1=1|Z1=z1,A0=a0);

**run**;

**proc** **sort** data=*yourlibrary*.hw\_09; by id; **run**;

**proc** **sort** data=b0 ; by id; **run**;

**proc** **sort** data=b1 ; by id; **run**;

**proc** **sort** data=b2 ; by id; **run**;

**data** hw\_09\_ipw;

merge *yourlibrary*.hw\_09 b0 b1 b2;

by id;

if a1=**0** then p\_den\_a1=ip\_0;

else if a1=**1** then p\_den\_a1=ip\_1;

\* STABILIZED WEIGHTS;

\* P(A0=0)P(A1=0) / P(A1=0|Z1=z1,A0=a0)P(A0=0);

if a0=**0** and a1=**0** then

sw=((**1**-p\_num\_a0)\*(**1**-p\_num\_a1)) / (p\_den\_a1\*(**1**-p\_num\_a0));

\* P(A0=0)P(A1=1) / P(A1=1|Z1=z1,A0=a0)P(A0=0);

if a0=**0** and a1=**1** then sw=((**1**-p\_num\_a0)\*p\_num\_a1) / (p\_den\_a1\*(**1**-p\_num\_a0));

\* P(A0=1)P(A1=1) / P(A1=1|Z1=z1,A0=a0)P(A0=1);

if a0=**1** and a1=**1** then sw=(p\_num\_a0\*p\_num\_a1) / (p\_den\_a1\*p\_num\_a0);

\* P(A0=1)P(A1=0) / P(A1=0|Z1=z1,A0=a0)P(A0=1);

if a0=**1** and a1=**0** then sw=(p\_num\_a0\*(**1**-p\_num\_a1)) / (p\_den\_a1\*p\_num\_a0);

drop ip\_0 ip\_1;

**run**;

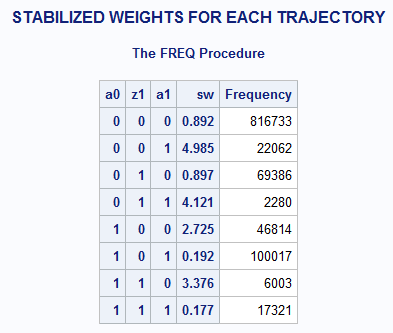
**proc** **freq** data=hw\_09\_ipw;

title "STABILIZED WEIGHTS FOR EACH TRAJECTORY";

tables a0\*z1\*a1\*sw / list nopercent nocum;

format sw **5.3**;

**run**;



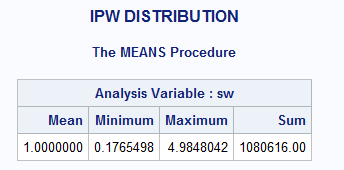
1. What is the expected value (mean) of the weights? Check the distribution of the weights and explain whether or not it is consistent with the expectation.

**proc** **means** data=hw\_09\_ipw mean min max sum;

title "IPW DISTRIBUTION";

var sw;

**run**;



The expected value of the weights is 1. The distribution of stabilized weights is consistent with expectation.

1. Check the effects of re-weighting the study population with stabilized weights.
2. What is the expected size of the pseudopopulation created by re-weighting the sample? Calculate the pseudopopulation and explain whether or not its size is consistent with the expectation.

**proc** **sort** data=hw\_09\_ipw;

by a0 z1 a1;

**run**;

**proc** **means** data=hw\_09\_ipw noprint;

by a0 z1 a1;

output out=b3(keep=a0 z1 a1 y sw \_freq\_ rename=(\_freq\_=n)) mean(sw y)=;

**run**;

**data** pseudopop;

set b3;

pseudopop=n\*sw;

**run**;

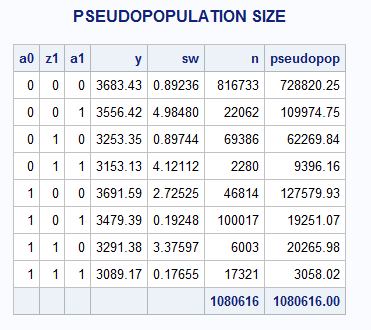
**proc** **print** noobs data=pseudopop;

title 'PSEUDOPOPULATION SIZE';

var a0 z1 a1 y sw n pseudopop;

sum n pseudopop;

**run**;



1. Examine the association between smoking during pregnancy and Z1 in the pseudopopulation. Describe the association in words, redraw the DAG to represent the causal system after re-weighting, and explain the implications of what has happened in the pseudopopulation on the estimation of the average effect of smoking during pregnancy on second offspring birthweight.

Pr[A1 = 1| Z1 = 1] = (9396.16+3058.02) / (62269.84 + 9396.16 + 20265.98 + 3058.02) = 0.1311

Pr[A1 = 1| Z1 = 0] = (109974.75 + 19251.07) / (728820.25 + 109974.75 + 127579.93 + 19251.07) = 0.1311

Pr[A1 = 1| Z1 = 1] - Pr[A1 = 1| Z1 = 0] = 0 (There is no associational difference of SGA1 on smoking at second pregnancy)

Pr[Z1 = 1| A0 = 1] = (20265.98 + 3058.02) / (127579.93 + 19251.07 + 20265.98 + 3058.02) = 0.137

Pr[Z1 = 1| A0 = 0] = (62269.84 + 9396.16) / (728820.25 + 109974.75 + 62269.84 + 9396.16) = 0.0787

Pr[Z1 = 1| A0 = 1]- Pr[Z1 = 1| A0 = 0] = 0.137-0.0787 = 0.058. (There is an associational difference of SGA1 on smoking at first pregnancy)

Z1 and A1 are independent, and we can remove the arrow from SGA1 to Smoking2. Therefore, SGA1 no longer acts as a confound and we no longer need to adjust for it. At the same time, it remains a mediator on the path from Smoking1 to SGA2.

Diagram, text

Description automatically generated

1. Estimate the average effect of smoking during pregnancy on second offspring birthweight in the pseudopopulation using a MSM. Compare this effect with the naïve estimates obtained in 6.

**proc** **genmod** data=hw\_09\_ipw;

class id;

title "MARGINAL STRUCTURAL MODEL";

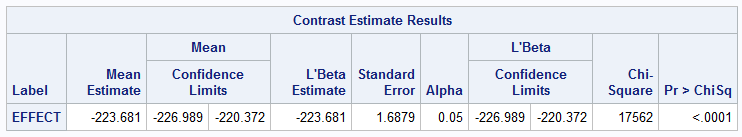
weight sw;

model y = a0 a1 a0\*a1;

repeated subject=id / type=ind;

estimate "EFFECT" a0 **1** a1 **1** a0\*a1 **1**;

**run**;



This estimate suggests a marginally smaller impact of the effect of smoking during pregnancy on second offspring birthweight than that obtained via naïve regression in 6.