## Comparison of the Four Methods

1. Draw a causal DAG corresponding to the R code of `compare.r` below. Run the simulation and store the data in `comparedat`.

```
compare.r<-function()
{
set.seed(999999999)
H<-rbinom(n=100000,size=1,prob=.4)
T<-rbinom(n=100000,size=1,prob=.5)
probA<-T*H*.6 + (1-T)*H*.3 + T*(1-H)*.3 + (1-T)*(1-H)*.1
A<-rbinom(n=100000,size=1,prob=probA)
probS<-A*.5 + (1-A)*.1
S<-rbinom(n=100000,size=1,prob=probS)
probY<-H*S*.75 + H*(1-S)*.5 + (1-H)*S*.5 + (1-H)*(1-S)*.2
Y<-rbinom(n=100000,size=1,prob=probY)
out<-data.frame(cbind(T,H,A,S,Y))
out
}
```

2. Assuming consistency, derive the true values for (1) $E(Y(1)|A = 1)$ and (2) $E(Y(0)|A = 1)$, as well as the average effect of treatment on the treated (ATT) expressed in terms of the (3) risk difference, (4) relative risk, and (5) odds ratio.

3. Use the outcome-modeling approach to standardization with a nonparametric model to estimate all five quantities and provide bootstrap confidence intervals.

4. Letting $Y_0 = H$ and $Y_1 = Y$, use the difference-in-difference approaches based on a linear, loglinear, and logistic model to estimate quantities (3), (4), and (5) and provide bootstrap confidence intervals.

5. Use the front-door method modified for the ATT to estimate all five quantities and provide bootstrap confidence intervals.

6. Suppose data on $S$ and $H$ are unavailable, and use the instrumental variables approach with the (a) linear, (b) loglinear, and (c) logistic structural nested mean models to estimate all five quantities and provide jackknife confidence intervals. Note that you will have three estimates of each of the five quantities. Also note that running the jackknife with such a large dataset takes a very long time. In case you run out of time, modify the sample size to 1,000.

7. Comment on the validity of the estimates you provided for questions 3 through 6. Make sure to note necessary assumptions.

8. Suppose compare.r generated data from a true associational DAG, but that the DAG was not causal. Suppose the additive equiconfounding assumption for the difference-in-differences approach is known to hold (see equation 18 from Module 2), with $Y_0 = H$. Comment on the validity of the estimates you provided for questions 3 through 6.

## Solutions

1. Draw a causal DAG corresponding to the `R` code of `compare.r` below. Run the simulation and store the data in `comparedat`.

```
compare.r<-function()
{
set.seed(999999999)
H<-rbinom(n=100000,size=1,prob=.4)
T<-rbinom(n=100000,size=1,prob=.5)
probA<-T*H*.6 + (1-T)*H*.3 + T*(1-H)*.3 + (1-T)*(1-H)*.1
A<-rbinom(n=100000,size=1,prob=probA)
probS<-A*.5 + (1-A)*.1
S<-rbinom(n=100000,size=1,prob=probS)
probY<-H*S*.75 + H*(1-S)*.5 + (1-H)*S*.5 + (1-H)*(1-S)*.2
Y<-rbinom(n=100000,size=1,prob=probY)
out<-data.frame(cbind(T,H,A,S,Y))
out
}
```

The DAG is shown below.

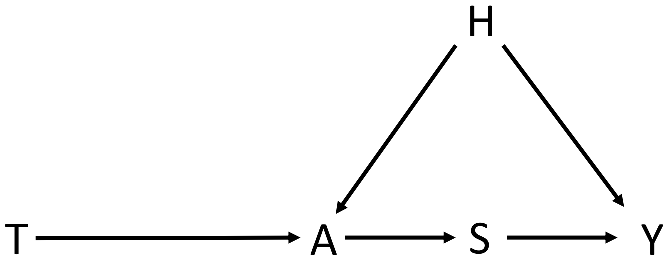

Figure 1: DAG corresponding to `compare.r`

2. Assuming consistency, derive the true values for (1) $E(Y(1)|A=1)$ and (2) $E(Y(0)|A=1)$, as well as the average effect of treatment on the treated (ATT) expressed in terms of the (3) risk difference, (4) relative risk, and (5) odds ratio.

Given consistency, we can estimate $E(Y(1)|A=1)$ via $E(Y|A=1)$. From the DAG, we see that $H$ is a sufficient confounder, so that $E(Y(0)|A=1) = E_{H|A=1}E(Y|H, A=0)$.

From the code,

$$E(Y|H,S,A) = .75HS + .5H(1-S) + .5(1-H)S + .2(1-H)(1-S),$$

so that

$$E(Y|H,A) = .75HE(S|A,H) + .5H(1 - E(S|A,H)) + .5(1 - H)E(S|A,H) + .2(1 - H)(1 - E(S|A,H)).$$

From the code,

$$E(S|A,H) = .5A + .1(1 - A) = .4A + .1,$$

so that

$$E(Y|H,A) = .75H(.4A + .1) + .5H(.9 - .4A) + .5(1 - H)(.4A + .1) + .2(1 - H)(.9 - .4A). \tag{1}$$

From the multiplication rule,

$$P(H = h|A = a) = \frac{P(A = a|H = h)P(H = h)}{P(A = a)}.$$

From the law of total probability,

$$P(A = 1|H = h) = \Sigma_t P(A = 1|H = h, T = t)P(T = t). \tag{2}$$

From the code and double expectation, we have that

$$P(H = 1) = .4$$

and

$$P(A = 1) = .4 * .5 * .6 + .5 * .4 * .3 + .3 * .5 * .6 + .1 * .6 * .5 = 0.3.$$

From (2) we have that

$$P(A = 1|H = 0) = .1 * .5 + .3 * .5 = .2$$

and

$$P(A = 1|H = 1) = .3 * .5 + .6 * .5 = .45.$$

Finally,

$$P(H = 1|A = 1) = P(A = 1|H = 1)P(H = 1)/P(A = 1) = .45 * .4/.3 = .6,$$

so that

$$P(H = 0|A = 1) = 0.4.$$

Returning to (1) and substituting in,

$$E(Y|H, A = 0) = .75H.1 + .5H.9 + .5(1 - H).1 + .2(1 - H).9.$$

Therefore,

$$E(Y(0)|A = 1) = .525E(H|A = 1) + .23(1 - E(H|A = 1)) = .525 * .6 + .23 * .4 = 0.407.$$

Furthermore,

$$E(Y|H, A = 1) = .75H.5 + .5H.5 + .5(1 - H).5 + .2(1 - H).5 = .625H + .35(1 - H),$$

so that

$$E(Y|A=1) = .625E(H|A=1) + .35(1 - E(H|A=1)) = .625 * .6 + .35 * .4 = 0.515.$$

Therefore the true values are

$$\begin{aligned} E(Y(0)|A=1) &= 0.407 \\ E(Y(1)|A=1) &= 0.515 \\ RD &= 0.108 \\ RR &= 1.265 \\ OR &= 1.547 \end{aligned}$$

3. Use the outcome-modeling approach to standardization with a nonparametric model to estimate all five quantities and provide bootstrap confidence intervals.

   We use the following code together with the bootstrap.

```
standatt.r<-function(data,ids)
{
dat<-data[ids,]
EHA<-mean(dat$H[dat$A==1])
beta<-lm(Y~A*H,data=dat)$coef
EY0A<-beta[1]+beta[3]*EHA
EY1A<-beta[1]+beta[2]+beta[3]*EHA+beta[4]*EHA
rd<-EY1A-EY0A
logrr<-log(EY1A/EY0A)
logor<-log(EY1A*(1-EY0A)/((1-EY1A)*EY0A))
c(EY0A,EY1A,rd,logrr,logor)
}
```

   The estimates and confidence intervals are presented in Table 1.

4. Letting $Y_0 = H$ and $Y_1 = Y$, use the difference-in-difference approaches based on a linear, loglinear, and logistic model to estimate quantities (3), (4), and (5) and provide bootstrap confidence intervals.

   We use the following code together with the bootstrap.

```
> mklong.r
function(dat=exam2dat)
{
longdat<-data.frame("Y"=rep(0,2*nrow(dat)),"A"=rep(dat[,"A"],each=2),
  "time"=rep(c(0,1),times=nrow(dat)))
longdat$Y[c(TRUE,FALSE)]<-dat[,"H"]  # this is Y_0
longdat$Y[c(FALSE,TRUE)]<-dat[,"Y"]  # this is Y_1
longdat
}
```

```
> didlinear.r
function(data=exam2dat,ids=c(1:nrow(exam2dat)))
{
dat<-data[ids,]
dat<-mklong.r(dat)
beta<-lm(Y~A+time+A*time,data=dat)$coef
rd<-beta[4]
EY1<-mean(dat$Y[(dat$A==1)&(dat$time==1)])
EY0<-EY1-rd
logrr<-log(EY1)-log(EY0)
logor<-log(EY1)-log(1-EY1)-log(EY0)+log(1-EY0)
c(EY0,EY1,rd,logrr,logor)
}
> didloglinear.r
function(data=exam2dat,ids=c(1:nrow(exam2dat)))
{
dat<-data[ids,]
dat<-mklong.r(dat)
beta<-glm(Y~A+time+A*time,family=poisson,data=dat)$coef
logrr<-beta[4]
EY1<-mean(dat$Y[(dat$A==1)&(dat$time==1)])
EY0<-EY1/exp(logrr)
rd<-EY1-EY0
logor<-log(EY1)-log(1-EY1)-log(EY0)+log(1-EY0)
c(EY0,EY1,rd,logrr,logor)
}
> didlogistic.r
function(data=exam2dat,ids=c(1:nrow(exam2dat)))
{
dat<-data[ids,]
dat<-mklong.r(dat)
beta<-glm(Y~A+time+A*time,family=binomial,data=dat)$coef
logor<-beta[4]
EY1<-mean(dat$Y[(dat$A==1)&(dat$time==1)])
tmp<-log(EY1/(1-EY1)) - logor
EY0<-exp(tmp)/(1+exp(tmp))
rd<-EY1-EY0
logrr<-log(EY1)-log(EY0)
c(EY0,EY1,rd,logrr,logor)
}
```

The estimates and confidence intervals are presented in Table 1, with 4a, 4b, and 4c using the linear, loglinear, and logistic DiD approaches, respectively.

5. Use the front-door method modified for the ATT to estimate all five quantities and provide bootstrap confidence intervals.

   We use the following code together with the bootstrap.

```
frontdooratt.r<-function(data=exam2dat,ids=c(1:nrow(exam2dat)))
{
dat<-data[ids,]
tmp00<-(1-mean(dat$S[dat$A==0]))*mean(dat$Y[(dat$S==0)&(dat$A==1)])

tmp01<-(mean(dat$S[dat$A==0]))*
mean(dat$Y[(dat$S==1)&(dat$A==1)])

EY0<-tmp00+ tmp01

EY1<-mean(dat$Y[dat$A==1])
RD<-EY1-EY0
logRR<-log(EY1/EY0)
logOR<-log(EY1/(1-EY1)) - log(EY0/(1-EY0))
c(EY0,EY1,RD,logRR,logOR)
}
```

The estimates and confidence intervals are presented in Table 1.

6. Suppose data on $S$ and $H$ are unavailable, and use the instrumental variables approach with the (a) linear, (b) loglinear, and (c) logistic structural nested mean models to estimate all five quantities and provide jackknife confidence intervals. Note that you will have three estimates of each of the five quantities. Also note that running the jackknife with such a large dataset takes a very long time. In case you run out of time, modify the sample size to 1,000.

   We use the following code together with the jackknife.

```
ividentity.r<-function(data)
{
dat<-data
Deta<-predict(glm(Y~A*T,data=dat),type="link")
Ystar<-Deta
Astar<- dat$A
Z<- dat$T
beta<-ivreg(formula=Ystar~Astar,instruments=~Z)$coef[2]
EY1<-mean(Deta[dat$A==1])
EY0<-mean((Deta-dat$A*beta)[dat$A==1])
RD<-EY1-EY0
logRR<-log(EY1/EY0)
logOR<-log(EY1/(1-EY1)) - log(EY0/(1-EY0))
c(EY0,EY1,RD,logRR,logOR)
}
ivlog.r<-function(data)
{
dat<-data
niter=10
A<-dat$A
Z<-dat$T
Deta<-predict(glm(Y~A*T,family=poisson,data=dat),type="link")
betat<--1
for (i in 1:niter)
{
#cat("i = ",i,"\n")
Ystar<-exp(Deta-A*betat)*(1+A*betat)
Astar<-A*exp(Deta-A*betat)
betat<-ivreg(formula=Ystar~Astar,instruments=~Z)$coef[2]
#cat("betat = ",betat,"\n")
}
beta<-betat
EY1<-mean(exp(Deta)[A==1])
EY0<-mean(exp(Deta-A*beta)[A==1])
RD<-EY1-EY0
logRR<-log(EY1/EY0)
logOR<-log(EY1/(1-EY1)) - log(EY0/(1-EY0))
c(EY0,EY1,RD,logRR,logOR)
}
ivlogit.r<-function(data)
{
dat<-data
niter<-10
A<-dat$A
Z<-dat$T
Deta<-predict(glm(Y~A*T,family=binomial,data=dat),type="link")
betat<-0
for (i in 1:niter)
{
#cat("i = ",i,"\n")
tmp<-exp(Deta-A*betat)/(1+exp(Deta-A*betat))
Ystar<-tmp*(1+A*betat*(1-tmp))
Astar<- A*tmp*(1-tmp)
betat<-ivreg(formula=Ystar~Astar,instruments=~Z)$coef[2]
#cat("betat = ",betat,"\n")
}
beta<-betat
EY1<-mean((exp(Deta)/(1+exp(Deta)))[A==1])
EY0<-mean((exp(Deta-A*beta)/(1+exp(Deta-A*beta)))[A=1])
RD<-EY1-EY0
logRR<-log(EY1/EY0)
logOR<-log(EY1/(1-EY1)) - log(EY0/(1-EY0))
c(EY0,EY1,RD,logRR,logOR)
}
```

The estimates and confidence intervals are presented in Table 1.

Table 1: Estimates and Confidence Intervals for Questions 3 through 6

| Question | $\hat{E}(Y(0)\vert A=1)$ | $\hat{E}(Y(1)\vert A=1)$ | $\hat{RD}$ | $\hat{RR}$ | $\hat{OR}$ |
|---|---|---|---|---|---|
| 3 | 0.406 (0.402, 0.411) | 0.520 (0.515, 0.526) | 0.114 (0.107, 0.121) | 1.28 (1.26, 1.30) | 1.58 (1.54, 1.63) |
| 4a | 0.612 (0.605, 0.619) | 0.520 (0.514, 0.526) | -0.092 ( -0.099, -0.084) | 0.850 (0.839, 0.862) | 0.0.688 (0.666, 0.711) |
| 4b | 0.620 (0.610, 0.630) | 0.520 (0.514, 0.526) | -0.100 ( -0.110, -0.090) | 0.839 (0.824, 0.854) | 0.664 (0.636, 0.693) |
| 4c | 0.613 (0.605, 0.620) | 0.520 (0.514, 0.526) | -0.092 (-0.101, -0.084) | 0.849 (0.837, 0.861) | 0.685 (0.663, 0.709) |
| 5 | 0.411 (0.404, 0.417) | 0.520 (0.515, 0.526) | 0.110 (0.105, 0.114) | 1.27 (1.25, 1.28) | 1.56 (1.53, 1.59) |
| 6a | 0.419 (0.395, 0.444) | 0.520 (0.514, 0.526) | 0.101 (0.075, 0.126) | 1.24 (1.17, 1.32) | 1.50 (1.35, 1.66) |
| 6b | 0.412 (0.387, 0.438) | 0.520 (0.514, 0.526) | 0.108 (0.082, 0.133) | 1.26 (1.19, 1.39) | 1.54 (1.39, 1.72) |
| 6c | 0.419 (0.394, 0.444) | 0.520 (0.514, 0.526) | 0.101 (0.075, 0.127) | 1.24 (1.17, 1.32) | 1.50 (1.35, 1.67) |

7. Comment on the validity of the estimates you provided for questions 3 through 6. Make sure to note necessary assumptions.

   The estimates all require the consistency assumption. The positivity assumption can be checked from the simulation and it holds. The estimates for question 3 are valid because $H$ is a sufficient confounder and we used a nonparametric model. Indeed, the confidence intervals all cover the true values. The estimates for question 4 are unlikely to be valid because, as we went over in the short course, typically if standardization is valid difference-in-differences is not valid. Indeed, the confidence intervals for question 4, with the exception of those for $E(Y(1)|A = 1)$ which do not rely on the DiD modeling assumptions, do not cover the true values and they all suggest that the causal effect is in the opposite direction to the truth. This example shows the danger of just using one method to analyze the data. The estimates for question 5 are valid because the structure of A,H,S,Y in the DAG is a front-door structure. Indeed, the confidence intervals all cover the true values. The estimates for question 6 require exclusion, which can be seen to hold from the simulation, but they also require that their respective structural nested mean model holds. We see that for the linear SNMM, the confidence intervals all contain the true values, suggesting that the linear SNMM is approximately true. We see that for the loglinear SNMM, the confidence intervals also contain the true values, suggesting that the loglinear SNMM is approximately true. Finally, we see that for the logistic SNMM, the confidence intervals contain the true values, so that the logistic SNMM is also approximately true.

8. Suppose compare.r generated data from a true associational DAG, but that the DAG was not causal. Suppose the additive equiconfounding assumption for the difference-in-differences approach is known to hold (see equation 18 from Module 2), with $Y_0 = H$. Comment on the validity of the estimates you provided for questions 3 through 6.

   In this case, the estimates for question 4 with the linear DiD model validly estimate the average effect of treatment on the treated. Standardization, instrumental variables, and the front door method all fail in this case – we see that their estimates are far away from the DiD estimates. The conflict between the answers to this question and the previous one can be viewed as a generalization of Lord's Paradox (see Module 2).