Babette A. Brumback

Odd Exercises Solutions Manual for Fundamentals of Causal Inference

1

Chapter 1

1. Item (8), experiment.

3. Items (6) and (7), plausibility and coherence.

5. Item (3), specificity.

7. Item (4), temporality.

2

Chapter 2

1. We have that the number of people in the What-If? Study is 165, from

```
> sum(data.frame(xtabs(~T+A+H+Y,whatifdat))$Freq)
[1] 165
```

   The left-hand side equals

$$P(A = 1, T = 1) = (27 + 3 + 9 + 13)/165 = 0.315.$$

   The right-hand side equals the product of

$$P(A = 1|T = 1) = (27 + 3 + 9 + 13)/80 = 0.650$$

   and

$$P(T = 1) = 80/165 = 0.485,$$

   noting that

```
> table(whatifdat$T)
 0  1
85 80
```

   This product is

$$0.650 * 0.485 = 0.315,$$

   which equals the left-hand side.

3. First, we show it mathematically. The textbook states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

   and the Law of Total Probability states that

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}).$$

   Second, we show it empirically. To do this, we make the following computations in R.

```
> table(whatifdat$H)
  0   1
106  59
> table(whatifdat$T)
 0  1
85 80
> table(whatifdat$H,whatifdat$T)
     0  1
  0 58 48
  1 27 32
```

The left-hand side is

$$P(T = 1|H = 1) = 32/59 = 0.542.$$

The pieces of the right-hand side are

$$P(H = 1|T = 1) = 32/80 = 0.4,$$

$$P(T = 1) = 80/165 = 0.485,$$

$$P(H = 1|T = 0) = 27/85 = 0.318,$$

and

$$P(T = 0) = 85/165 = 0.515.$$

Thus, the right-hand side equals

$$0.4 * 0.485/(0.4 * 0.485 + 0.318 * 0.515) = 0.542,$$

which equals the left-hand side.

5. First, we do the computation for the nonparametric estimator:

```
> npboot.r
function ()
{
estimator<-function(data,ids)
{
dat<-data[ids,]
npest<-mean(dat$Y[(dat$A==1)&(dat$H==0)&(dat$T==1)])
}
boot.out<-boot(data=whatifdat,statistic=estimator,R=1000)
est<-summary(boot.out)$original
SE<-summary(boot.out)$bootSE
lci<-est-1.96*SE
uci<-est+1.96*SE
list(est=est,lci=lci,uci=uci)
}
> npboot.r()
$est
```

```
[1] 0.1
$lci
[1] -0.011105
$uci
[1] 0.21111
```

Second, we do the computation for the parametric estimator:

```
> lmodexboot.r
function ()
{
estimator<-function(data,ids)
{
dat<-data[ids,]
coef<-glm(Y~A+T+H, family=binomial,data=dat)$coef
xbeta<-sum(coef[c(1:3)])
xbeta
}
boot.out<-boot(data=whatifdat,statistic=estimator,R=1000)
logitest<-summary(boot.out)$original
SE<-summary(boot.out)$bootSE
logitlci<-logitest-1.96*SE
logituci<-logitest+1.96*SE
est<-exp(logitest)/(1+exp(logitest))
lci<-exp(logitlci)/(1+exp(logitlci))
uci<-exp(logituci)/(1+exp(logituci))
list(est=est,lci=lci,uci=uci)
}
<bytecode: 0x0000000044d08a50>
> lmodexboot.r()
$est
[1] 0.090013
$lci
[1] 0.037288
$uci
[1] 0.20168
```

We find that the nonparametric estimator and 95% confidence interval is 0.1 (-0.011, 0.211), whereas for the parametric estimator we have 0.09 (0.037, 0.202). The estimates are fairly close together, but as expected, the confidence interval for the nonparametric estimator is wider than that for the parametric estimator, even including negative values, which are not possible. There are other ways to make a confidence interval for the nonparametric estimator that do not include negative values, and we explore two such ways here.

```
> lmodnpexboot.r
function ()
```

```
{
estimator<-function(data,ids)
{
dat<-data[ids,]
out<-glm(Y~A*T*H, family=binomial,data=dat)
newdata<-data.frame(A=1,T=1,H=0)
npest<-predict(out,newdata=newdata,type="link")
npest
}
boot.out<-boot(data=whatifdat,statistic=estimator,R=1000)
logitest<-summary(boot.out)$original
SE<-summary(boot.out)$bootSE
logitlci<-logitest-1.96*SE
logituci<-logitest+1.96*SE
est<-exp(logitest)/(1+exp(logitest))
lci<-exp(logitlci)/(1+exp(logitlci))
uci<-exp(logituci)/(1+exp(logituci))
list(est=est,lci=lci,uci=uci)
}
> lmodnpexboot.r()
$est
[1] 0.1
$lci
[1] 9.4805e-05
$uci
[1] 0.99238
```

We see now that the nonparametric method (rounded) returns 0.1 (0.0001,0.992), with a confidence interval that does not include negative numbers. However, due to the small sample of the $A = 1$, $T = 1$, and $H = 0$ stratum, with just 30 participants, only 3 of whom have $Y = 1$, this function has an insidious error due to the glm function not working properly for the saturated model, and this invalidates the confidence interval.

We can see the error from running

```
> nplogitboot.r
function ()
{
estimator<-function(data,ids)
{
dat<-data[ids,]
npest<-logit(mean(dat$Y[(dat$A==1)&(dat$H==0)&(dat$T==1)]))
}
boot.out<-boot(data=whatifdat,statistic=estimator,R=1000)
est<-summary(boot.out)$original
SE<-summary(boot.out)$bootSE
lci<-expit(est-1.96*SE)
```

```
uci<-expit(est+1.96*SE)
est<-expit(est)
list(est=est,lci=lci,uci=uci)
}
> nplogitboot.r()
$est
[1] 0.1

$lci
[1] NaN

$uci
[1] NaN
```

which returns NaN for the confidence limits due to the logit function not being able to take the log of zero for some bootstrap samples.

For larger samples, these two methods would both agree exactly and return confidence intervals that do not escape the unit interval. There do exist methods for confidence intervals for small samples, but we do not explore them here.

3

Chapter 3

1.
```
> # make the dataset
> gssgun<-gss[,c("owngun","conservative","gt65","white","female")]
> gssguncc<-gssgun[complete.cases(gssgun),]
> # fit the parametric logistic model
> summary(glm(owngun~conservative+white+female+gt65,family=binomial,data=gssguncc))
glm(formula = owngun ~ conservative + white + female + gt65,
    family = binomial, data = gssguncc)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.213      0.143   -8.51  < 2e-16
conservative    0.632      0.121    5.23  1.7e-07
white           0.903      0.142    6.35  2.2e-10
female         -0.570      0.114   -5.02  5.1e-07
gt65            0.193      0.137    1.40     0.16
```

   We see that conservative political views and self-reporting as white increases the chance of owning a gun, whereas being a female decreases the chance, and age group is not statistically significant.

   We modify the function `estimator` in `bootu.r` as follows:

```
estimator<-function(data,ids)
{
dat<-data[ids,]
mod<-glm(owngun~conservative,family=gaussian,data=dat)
p0<-mod$coef[1]
p1<-mod$coef[1]+mod$coef[2]
rd<-mod$coef[2]
logrr<-glm(owngun~conservative,family=poisson,data=dat)$coef[2]
owngunstar<-1-dat$owngun
fairstar<-1-dat$conservative
logrrstar<-glm(owngunstar~fairstar,family=poisson,data=dat)$coef[2]
logor<-glm(owngun~conservative,family=binomial,data=dat)$coef[2]
c(p0,p1,rd,logrr,logrrstar,logor)
}
```

   Letting Y indicate owning a gun and T indicate conservative political views, we find that the estimate and 95% confidence interval for $E(Y|T = 1)$ are $0.479(0.426, 0.524)$, whereas those for $E(Y|T = 0)$ are $0.308(0.280, 0.336)$. Therefore, individuals with conservative political

views appear more likely to own a gun. We cannot state that conservative political views causes individuals to own guns. Estimates of our four association measures with their 95% confidence intervals are presented in Table 3.1.

TABLE 3.1
Four Association Measures Relating
Conservative Political Views to Owning a Gun

Measure	Estimate	95% Confidence Interval
RD	0.167	(0.109,0.224)
RR	1.54	(1.34, 1.78)
RR*	1.32	(1.19,1.46)
OR	2.03	(1.60, 2.59)

We observe that the association is statistically significant.

Next, we modify the function `estimator` in `lmodboot.r` as follows:

```
estimator<-function(data,ids)
{
dat<-data[ids,]
coef<-glm(owngun~conservative+white+female+gt65,family=binomial,data=dat)$coef
xbeta1<-sum(coef)-coef[4]
xbeta0<-sum(coef)-coef[4]-coef[2]
p1<-exp(xbeta1)/(1+exp(xbeta1))
p0<-exp(xbeta0)/(1+exp(xbeta0))
rd<-p1-p0
logrr<-log(p1)-log(p0)
logrrstar<-log((1-p0))-log((1-p1))
logor<-log(p1/(1-p1))-log(p0/(1-p0))
c(p1,p0,rd,logrr,logor,logrrstar)
}
```

Let $H = h$ denote setting the confounders to indicate a male who is greater than 65 and self-reported as white. We find that the estimate and 95% confidence interval for $E(Y|T = 1, H = h)$ are $0.626(0.554, 0.698)$, whereas those for $E(Y|T = 0, H = h)$ are $0.471(0.403, 0.538)$. Therefore, male participants who are older than 65 and are self-reported as white appear more likely own a gun if they have conservative political views. While it is likely that conservative political views precede owning a gun for some respondents, it is also likely that they succeed owning a gun for other respondents. Therefore, temporality prevents a causal interpretation. Even if temporality were not an issue, we would need our parametric logistic model to hold and we would also need the consistency assumption, positivity, and the potential outcomes to be independent of T conditional

on H. Estimates of our four association or effect measures with their 95% confidence intervals are presented in Table 3.2.

TABLE 3.2
Four Conditional Association or Effect Measures Relating Conservative Political Views to Owning a Gun

Measure	Estimate	95% Confidence Interval
RD	0.155	(0.098,0.213)
RR	1.33	(1.19, 1.48)
RR*	1.42	(1.22,1.64)
OR	1.88	(1.48, 2.40)

Compared to the unconditional analysis, we observe that the associations appear slightly weaker for RD, RR, and OR but slightly stronger for RR*. They are all statistically significant.

3. Let Y denote zero drinks and $H = h$ denote the confounder settings.

First, we fit the logistic model to investigate the fitted coefficients.

```
> summary(glm(zerodrinks~rural+female+whitenh+
blacknh+hisp+multinh+gt65+gthsedu,family=binomial,data=brfss))
glm(formula = zerodrinks ~ rural + female + whitenh + blacknh +
    hisp + multinh + gt65 + gthsedu, family = binomial, data = brfss)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.54220    0.01793   30.23  < 2e-16
rural        0.25981    0.00983   26.44  < 2e-16
female       0.48698    0.00711   68.51  < 2e-16
whitenh     -0.64373    0.01716  -37.52  < 2e-16
blacknh     -0.14106    0.02130   -6.62  3.5e-11
hisp        -0.27710    0.02127  -13.03  < 2e-16
multinh     -0.30779    0.02964  -10.39  < 2e-16
gt65         0.55710    0.00740   75.27  < 2e-16
gthsedu     -0.77127    0.00768 -100.38  < 2e-16
```

We observe that responders living in rural counties, who are female, with less than high school education are less likely to drink. The reference category for race and ethnicity is `othernh`; respondents in all other categories are more likely to drink. The categories selected by the BRFSS do not include Asians, therefore, Asian non-hispanics are included in `othernh`.

We modify the `estimator` code in `lmodboot.r` as follows:

```
estimator<-function(data,ids)
{
dat<-data[ids,]
```

```
out<-glm(zerodrinks~rural+female+whitenh+blacknh+hisp+multinh+gt65+gthsedu,
family=binomial,data=dat)
dat0<-data.frame(rural=0,female=0,whitenh=1,blacknh=0,hisp=0,multinh=0,
gt65=0,gthsedu=1)
dat1<-data.frame(rural=1,female=0,whitenh=1,blacknh=0,hisp=0,multinh=0,
gt65=0,gthsedu=1)
p1<-predict(out,newdata=dat1,type="response")
p0<-predict(out,newdata=dat0,type="response")
rd<-p1-p0
logrr<-log(p1)-log(p0)
logrrstar<-log((1-p0))-log((1-p1))
logor<-log(p1/(1-p1))-log(p0/(1-p0))
c(p1,p0,rd,logrr,logor,logrrstar)
}
```

Furthermore, we only use 500 iterations of the bootstrap, to save some time because the dataset is so large. Results are presented in Table 3.3. To

TABLE 3.3
Rural Residence and Refraining from Drinking Alcohol

Measure	Estimate	95% Confidence Interval
$P(Y(1)\|H=h)$	0.351	(0.346, 0.356)
$P(Y(0)\|H=h)$	0.295	(0.292, 0.298)
RD	0.057	(0.052, 0.061)
RR	1.19	(1.18, 1.21)
RR*	1.087	(1.080, 1.095)
OR	1.30	(1.27, 1.32)

interpret these estimates causally, we would need to assume consistency, positivity, and we would need for our parametric logistic model to be correct and for H to represent a sufficient set of confounders.

Next part:

First we restrict the dataset using

```
> brfssdrinkers<-brfss[brfss$zerodrinks==0,]
```

Second, we fit the loglinear model to investigate the fitted coefficients.

```
> summary(glm(maxdrinks~rural+female+whitenh+blacknh+
hisp+multinh+gt65+gthsedu,family=poisson,data=brfssdrinkers))
glm(formula = maxdrinks ~ rural + female + whitenh + blacknh +
    hisp + multinh + gt65 + gthsedu, family = poisson, data = brfssdrinkers)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.68599    0.00686  245.88  < 2e-16
rural        0.02979    0.00384    7.77  8.0e-15
```

```
female      -0.46762    0.00275 -170.32  < 2e-16
whitenh     -0.02975    0.00664   -4.48  7.5e-06
blacknh     -0.16753    0.00867  -19.33  < 2e-16
hisp         0.01539    0.00805    1.91    0.056
multinh      0.07551    0.01087    6.95  3.7e-12
gt65        -0.52603    0.00325 -161.70  < 2e-16
gthsedu     -0.17052    0.00297  -57.43  < 2e-16
```

We observe that, conditional on drinking, rural residents are more likely to consume a greater number of drinks on their biggest occasion. Respondents greater than 65 years of age, females, and those with more than a high school education consume fewer maximum drinks, whereas among the racial-ethnic categories, compared to `othernh`, only those self-identifying as multiracial nonhispanic report consuming more drinks per occasion.

Then, we modify the `estimator` function in `lmodboot.r` as follows:

```
estimator<-function(data,ids)
{
dat<-data[ids,]
out<-glm(maxdrinks~rural+female+whitenh+blacknh+hisp+multinh+gt65+gthsedu,
family=poisson,data=dat)
dat0<-data.frame(rural=0,female=0,whitenh=1,blacknh=0,hisp=0,multinh=0,
gt65=0,gthsedu=1)
dat1<-data.frame(rural=1,female=0,whitenh=1,blacknh=0,hisp=0,multinh=0,
gt65=0,gthsedu=1)
p1<-predict(out,newdata=dat1,type="response")
p0<-predict(out,newdata=dat0,type="response")
rd<-p1-p0
logrr<-log(p1)-log(p0)
c(p1,p0,rd,logrr)
}
```

We do not use RR* or OR for this analysis as they do not make sense. We interpret RD as a difference in means and RR as a ratio of means. Once again, we only use 500 iterations of the bootstrap, to save some time because the dataset is so large. Results are presented in Table 3.4. To

TABLE 3.4
Rural Residence and Maximum Number of Drinks on One Occasion

Measure	Estimate	95% Confidence Interval
$E(Y(1)\|H=h)$	4.55	(4.49, 4.61)
$E(Y(0)\|H=h)$	4.42	(4.39, 4.45)
RD	0.133	(0.070, 0.197)
RR	1.03	(1.02, 1.04)

interpret these estimates causally, we would consistency, positivity, and

we would need for our parametric loglinear model to be correct and for H to represent a sufficient set of confounders. We observe that rural residents consume a statistically significantly higher maximum number of drinks on one occasion, conditional on the chosen values of the confounders, although their average maximum number is not that much higher than that for non-rural residents. The dataset is large, and therefore statistical significance is easily attained.

4

Chapter 4

1. We let M indicate aged 18-29 years, Y indicate a cost barrier, and T indicate having a disability. Results are presented below.

TABLE 4.1
Effect-measure Modification in the Brumback et al. Example

Measure	$M = 0$	$M = 1$	Modification
$\hat{E}(Y(0)\|M)$	0.040	0.225	NA
$\hat{E}(Y(1)\|M)$	0.072	0.383	NA
$\hat{RD}$	0.032	0.158	0.126
$\hat{RR}$	1.80	1.70	0.944
$\hat{RR}*$	1.03	1.26	1.21
$\hat{OR}$	1.85	2.14	1.16

We observe that RD, RR*, and OR suggest a stronger effect in the younger group whereas RR suggests a stronger effect in the older group. If we could interpret these results causally, dissolving the cost barrier in the younger group would result in more benefit per dollar than dissolving it in the older group, assuming (1) that dissolving the cost barrier costs the same per person for the younger group as it does for the older group and (2) that dissolving the cost barrier produces a benefit for a member of the younger group that is the same as it would produce for a member of the older group. We also need to assume that our results are statistically significant, which we cannot verify without the data.

3. We restrict the dataset as follows:

```
> brfsslt65<-brfss[brfss$gt65==0,]
```

We let Y be `insured`, T be `gthsedu`, and M be `whitenh`. We compute results using `boot.r` and `bootinside.r` and present them below.

TABLE 4.2
Effect-measure Modification in the BRFSS Study

Measure	$M = 0$	$M = 1$	Modification
$\hat{E}(Y(0)\|M)$ (95% CI)	0.697 (0.691, 0.703)	0.845 (0.842, 0.848)	0.148 (0.141, 0.154)
$\hat{E}(Y(1)\|M)$ (95% CI)	0.881 (0.878, 0.885)	0.933 (0.932, 0.935)	0.052 (0.048, 0.056)
$\hat{RD}$ (95% CI)	0.184 (0.178, 0.191)	0.088 (0.085, 0.092)	-0.096 (-0.103, -0.088)
$\hat{RR}$ (95% CI)	1.26 (1.25, 1.28)	1.10 (1.10, 1.11)	0.874 (0.865, 0.882)
$\hat{RR}*$ (95% CI)	2.55 (2.46, 2.64)	2.32 (2.25, 2.39)	0.910 (0.870, 0.951)
$\hat{OR}$ (95% CI)	3.23 (3.09, 3.37)	2.56 (2.48, 2.65)	0.79 (0.754, 0.838)

We observe that obtaining more than a high school education is associated with an increased chance of having health insurance irrespective of whether a respondent self-identifies as white non-Hispanic or not. If a causal interpretation were possible, we would observe that all four measures suggest that the effect of obtaining more than a high school education on having health insurance is more pronounced for those who do not self-identify as white non-Hispanic than it is for those who do. If we were not interested in estimates for RD and OR, we could have determined that all four measures agree for this example solely by computing RR and RR*.

5. We make the dataset

```
> head(gssch4)
  attend gthsedu female id
1      1       1      0  1
2      0       0      1  2
3      0       1      0  3
4      1       1      1  4
5      1       1      0  5
6      0       1      1  6
```

Then, we do the analysis, again using the `geeglm` function in the `geepack` package to easily obtain the P-values.

```
> summary(geeglm(attend~gthsedu*female,data=gssch4,id=id))
Call:
geeglm(formula = attend ~ gthsedu * female, data = gssch4, id = id)
 Coefficients:
              Estimate Std.err   Wald Pr(>|W|)
(Intercept)     0.2590  0.0170 232.13  < 2e-16
gthsedu         0.0908  0.0297   9.34   0.0022
female          0.1148  0.0244  22.16  2.5e-06
gthsedu:female -0.0762  0.0406   3.53   0.0602
```

If we were to additionally assume independence between the four potential outcomes and the two causes, and monotonicity of the causal types, we could use the interaction term in the above analysis to investigate synergy. That term, which is also the difference of risk differences, is estimated at -0.076, but it is not statistically significantly different from zero (P=0.06). Therefore, it is plausible that it is positive, which would be sufficient for synergy assuming monotonicity of the causal types, and it is plausible that it is zero or negative, in which case we would not be sure. We conclude that we cannot be sure whether there is a causal synergy of `gthsedu` and `female` or not.

5

Chapter 5

1. We let W_1 indicate the healthy diet, W_3 indicate the exercise program, W_2 indicate weight loss, A indicate increased nutrient intake, and Y indicate increased strength. Given the causal DAG of Figure **??**, we could measure the effect of nutrient intake on increased strength using any comparison of $P(Y = 1|A = 1)$ with $P(Y = 1|A = 0)$ (e.g. RD, RR, RR*, or OR). If the DAG is true, we would observe no effect. A reviewer might suppose the missing arrows from W_1 and W_2 to Y and from W_2 and W_3 to A should be present. Therefore, the reviewer might speculate that our null result would be overturned if we adjusted for confounders (assuming faithfulness did not hold). Another critique the reviwer might have is that the temporality between A and Y could be in either direction. It is possible that for some participants, A is causing Y, whereas for others, Y is causing A, and that these effects cancel each other out in the population.

3. The causal DAG is shown in Figure 5.1.

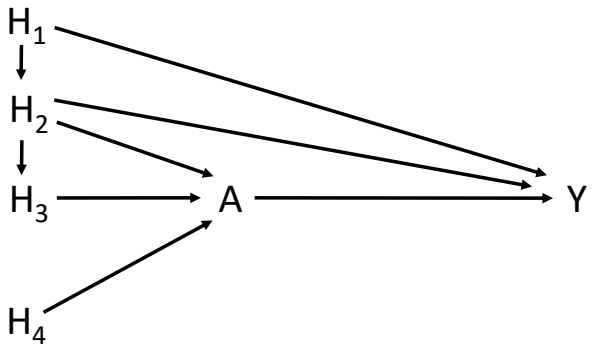

FIGURE 5.1: Causal DAG for Exercise 3

The true confounders are H_1 and H_2. The smallest sufficient set of true confounders is H_2.

The causal DAG redrawn with potential outcomes behind the scenes is shown in Figure 5.2.

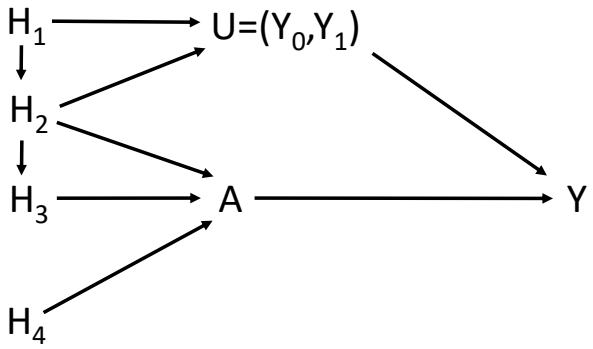

FIGURE 5.2: Causal DAG with Potential Outcomes Behind the Scenes for Exercise 3

We see that U blocks all backdoor paths from A to Y. The potential outcomes are not true confounders in this example because there is not a directed path from U to A.

5. The causal DAG is shown in Figure 5.3.

FIGURE 5.3: Causal DAG for Exercise 5

From `probY`, the risk difference for the entire population is $P(Y = 1|T = 1) - P(Y = 1|T = 0) = 0.01$. We have that

$$P(Y = 1|T, O = 1) = \frac{P(O = 1|Y = 1, T)P(Y = 1|T)}{P(O = 1|T)},$$

that $O \amalg T|Y$, and that

$$P(O = 1|T) = 0.01 * T + 0.1(1 - 0.01 * T) = 0.1 + 0.009 * T,$$

so that

$$P(Y = 1|T = 1, O = 1) = 1 * 0.01/.109 = 0.0917$$

and

$$P(Y = 1|T = 0, O = 1) = 1 * 0/0.1 = 0.$$

Therefore, the risk difference in the $O = 1$ group is 0.0917. We observe very much bias.

Empirically, we find

```
> ex5.dat<-simex5.r()
> xtabs(~Y+T+O,data=ex5.dat)
, , O = 0

   T
Y      0    1
  0 4522 4458
  1    0    0

, , O = 1

   T
Y      0    1
  0  488  469
  1    0   63

> 63/(63+469)
[1] 0.11842
> xtabs(~Y+T,data=ex5.dat)
   T
Y      0    1
  0 5010 4927
  1    0   63
> 63/(63+4927)
[1] 0.012625
```

So we see that the risk difference for the population is estimated as 0.0126 whereas for the $O = 1$ group it is estimated at 0.1184. We observe very much bias.

For the odds ratios, symmetry implies that

$$\frac{P(Y = 1|T = 1, O = 1)P(Y = 0|T = 0, O = 1)}{P(Y = 0|T = 1, O = 1)P(Y = 1|T = 0, O = 1)}$$

equals

$$\frac{P(T = 1|Y = 1, O = 1)P(T = 0|Y = 0, O = 1)}{P(T = 0|Y = 1, O = 1)P(T = 1|Y = 0, O = 1)}.$$

Because the causal DAG implies $T \amalg O|Y$, this latter odds ratio equals

$$\frac{P(T=1|Y=1)P(T=0|Y=0)}{P(T=0|Y=1)P(T=1|Y=0)},$$

which, by symmetry, equals

$$\frac{P(Y=1|T=1)P(Y=0|T=0)}{P(Y=0|T=1)P(Y=1|T=0)}.$$

For rare $Y = 1$, the odds ratio approximates the relative risk because $P(Y = 0|T = 0)$ and $P(Y = 0|T = 1)$ are approximately equal to one. These results are useful because it means that we can use the case-control study design to estimate relative risks with rare outcomes, which are harder to study using standard designs.

For our example, the odds ratio and relative risk are infinite due to $P(Y = 1|T = 0) = 0$ (we define these measures as the limit as the probability goes to zero from the right.)

6

Chapter 6

1. We can estimate with the `R` code

```
> fluout.r
function(data=brfsslt65,ids=c(1:nrow(brfsslt65)))
{
dat<-data[ids,]
lmod<-glm(flushot~insured+rural+female+whitenh+blacknh+hisp+multinh+gthsedu,
family=binomial,data=dat)
dat0<-dat1<-dat
dat0$insured<-0
dat1$insured<-1
EYhat0<-predict(lmod,newdata=dat0,type="response")
EYhat1<-predict(lmod,newdata=dat1,type="response")
EY0<-mean(EYhat0)
EY1<-mean(EYhat1)
rd<-EY1-EY0
rr<-log(EY1/EY0)
or<-log(EY1*(1-EY0)/(EY0*(1-EY1)))
c(EY0,EY1,rd,rr,or)
}
```

   and the bootstrap.

   The results are presented in Table 6.1, and we observe that the proportion who would have received a flu shot had everyone had health insurance would have been 0.440 (0.438, 0.442) versus 0.215 (0.210, 0.220) had no one had health insurance. We see that the risk difference, risk ratio, and odds ratio all indicate a strong effect of health insurance. If we had measured enough confounders and our outcome model is correct, we could conclude that having health insurance increases the chance of getting a flu shot.

3. We can estimate with the `R` code

```
> fludr.r
function(data=brfsslt65,ids=c(1:nrow(brfsslt65)))
{
dat<-data[ids,]
```

TABLE 6.1
Outcome-model Standardization
Measuring the Effect of Having Health
Insurance on Getting a Flu Shot

Measure	Estimate	95% CI
$\hat{E}(Y(0))$	0.215	(0.210, 0.220)
$\hat{E}(Y(1))$	0.440	(0.438, 0.442)
$\hat{RD}$	0.225	(0.219, 0.230)
$\hat{RR}$	2.04	(2.00, 2.09)
$\hat{OR}$	2.86	(2.78, 2.95)

```
e<-fitted(glm(insured~rural+female+whitenh+blacknh+
hisp+multinh+gthsedu,family=binomial,data=dat))
lmod<-glm(flushot~insured+rural+female+whitenh+blacknh+hisp+multinh+gthsedu,
family=binomial,data=dat)
dat0<-dat1<-dat
dat0$insured<-0
dat1$insured<-1
EYhat0<-predict(lmod,newdata=dat0,type="response")
EYhat1<-predict(lmod,newdata=dat1,type="response")
EY0<-mean(dat$flushot*(1-dat$insured)/(1-e)+EYhat0*(e-dat$insured)/(1-e))
EY1<-mean(dat$flushot*(dat$insured/e) - EYhat1*(dat$insured-e)/e)
rd<-EY1-EY0
rr<-log(EY1/EY0)
or<-log(EY1*(1-EY0)/(EY0*(1-EY1)))
c(EY0,EY1,rd,rr,or)
}
```

and the bootstrap.

The results are presented in Table 6.2, and we observe that the proportion who would have received a flu shot had everyone had health insurance would have been 0.440 (0.437, 0.442) versus 0.202 (0.196, 0.208) had no one had health insurance. We see that the risk difference, risk ratio, and odds ratio all indicate a strong effect of health insurance, and nearly agree with those from the exposure modeling approach. This might be taken to suggest that if either model is correct, it would be more likely to be the exposure model. If we had measured enough confounders and either our outcome model or our exposure model is correct, we could conclude that having health insurance increases the chance of getting a flu shot.

5. We can estimate with the `R` code

```
> twopartexp.r
```

TABLE 6.2
Doubly Robust Standardization
Measuring the Effect of Having Health
Insurance on Getting a Flu Shot

Measure	Estimate	95% CI
$\hat{E}(Y(0))$	0.202	(0.196, 0.208)
$\hat{E}(Y(1))$	0.440	(0.437, 0.442)
$\hat{RD}$	0.238	(0.231, 0.244)
$\hat{RR}$	2.18	(2.12, 2.24)
$\hat{OR}$	3.10	(2.99, 3.22)

```
function(data=brfss,ids=c(1:nrow(brfss)))
{
dat<-data[ids,]
dat$A<-dat$rural
e<-fitted(glm(rural~gt65+female+whitenh+blacknh+
hisp+multinh+gthsedu,family=binomial,data=dat))
dat$W<-(1/e)*dat$A + (1/(1-e))*(1-dat$A)
beta<-glm(maxdrinks~A,data=dat,weights=W)$coef
EY0<-beta[1]
EY1<-beta[1]+beta[2]
rd<-EY1-EY0
rr<-log(EY1/EY0)
c(EY0,EY1,rd,rr)
}
```

and the bootstrap.

The results are presented in Table 6.3, and we see that the results are very similar to those computed using the two-part outcome model. The rate difference is estimated at -0.133 (-0.159, -0.106), and the rate ratio is estimated at 0.921 (0.905, 0.937) for the effect of living in a rural county on maximium number of alcoholic drinks consumed on any occasion. If we have measured enough confounders and our exposure model is correct, we could conclude that living in a rural county decreases the maximum number of alcoholic drinks consumed on any occasion compared to living in an urban county.

7. We can estimate with the `R` code

```
> twopartatt.r
function(data=brfss,ids=c(1:nrow(brfss)))
{
dat<-data[ids,]
```

TABLE 6.3
Exposure-model Standardization Measuring the Effect of Living in a Rural County on Maximum Drinks Consumed on Any Occasion

Measure	Estimate	95% CI
$\hat{E}(Y(0))$	1.67	(1.66, 1.68)
$\hat{E}(Y(1))$	1.54	(1.51, 1.56)
$\hat{RD}$	-0.133	(-0.159, -0.106)
$\hat{RR}$	0.921	(0.905, 0.937)

```
dat$A<-dat$rural
e<-fitted(glm(rural~gt65+female+whitenh+blacknh+
hisp+multinh+gthsedu,family=binomial,data=dat))
e0<-mean(dat$A)
dat$W<-(1-dat$A)*e/(e0*(1-e))
EY0<-mean(dat$maxdrinks*dat$W)
EY1<-mean(dat$maxdrinks[dat$A==1])
rd<-EY1-EY0
rr<-log(EY1/EY0)
c(EY0,EY1,rd,rr)
}
```

and the bootstrap.

The results are presented in Table 6.4, and we see that they are fairly similar to those for exercise 5, although the estimated average potential outcomes among the rural residents are slightly lower than those among the population as a whole, reflecting the influence of fewer drinkers in the rural subpopulation. The rate difference is estimated at -0.127 (-0.152, -0.101), and the rate ratio is estimated at 0.921 (0.906, 0.937) for the effect of living in a rural county on maximium number of alcoholic drinks consumed on any occasion, specific to the subpopulation living in a rural county. If we have measured enough confounders and our exposure model is correct, we could conclude that living in a rural county decreases the maximum number of alcoholic drinks consumed on any occasion compared to what we would observe if those same residents lived in an urban county.

TABLE 6.4
Exposure-model Standardization for the ATT Measuring the Effect of Living in a Rural County on Maximum Drinks Consumed on Any Occasion

Measure	Estimate	95% CI
$\hat{E}(Y(0))$	1.61	(1.60, 1.62)
$\hat{E}(Y(1))$	1.48	(1.46, 1.51)
$\hat{RD}$	-0.127	(-0.152, -0.101)
$\hat{RR}$	0.921	(0.906, 0.937)

7

Chapter 7

1. We used the R programs

```
> mklongsb.r
function(dat=sepsisb)
{
longdat<-NULL
for (i in 1:nrow(dat))
{
Zubrod<-dat[i,"Zubrodbase"]
A<-dat[i,"shock"]
time<-0
longdat<-rbind(longdat,c(Zubrod,A,time))
Zubrod<-dat[i,"Zubrod1yr"]
A<-dat[i,"shock"]
time<-1
longdat<-rbind(longdat,c(Zubrod,A,time))
}
dimnames(longdat)[[2]]<-c("Zubrod","A","time")
data.frame(longdat)
}
> did.r
function(data=sepsisb,ids=c(1:nrow(sepsisb)))
{
dat<-data[ids,]
dat<-mklongsb.r(dat)
beta<-lm(Zubrod~A+time+A*time,data=dat)$coef
rd<-beta[4]
beta<-glm(Zubrod~A+time+A*time,data=dat,family=poisson)$coef
logrr<-beta[4]
beta<-glm(Zubrod~A+time+A*time,data=dat,family=binomial)$coef
logor<-beta[4]
c(rd,logrr,logor)
}
> standatt.r
function(data=sepsisb,ids=c(1:nrow(sepsisb)))
{
dat<-data[ids,]
dat$A<-dat$shock
dat$H<-dat$Zubrodbase
dat$Y<-dat$Zubrod1yr
```

```
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)/(EY0A*(1-EY1A)))
c(EY0A,EY1A,rd,logrr,logor)
}
```

together with the bootstrap for all of the estimation.

Results are presented in Table 7.1. We observe that the point estimates all suggest that septic shock increases the risk of a poor Zubrod score in one year, with the standardized ATT suggesting slightly stronger effects than the DiD methods. However, the DiD results are not statistically significant, whereas the standardized ATT results are. We should tell our collaborators that there looks like there may be an effect but that we cannot be certain due to sampling variability. We need a larger study.

TABLE 7.1
Estimates and 95% Confidence Intervals for Exercise 1

Method	RD (95%CI)	RR (95% CI)	OR (95% CI)
DiD Linear	0.129 (-0.028, 0.286)	NA	NA
DiD Loglinear	NA	1.5 (0.839, 2.68)	NA
DiD Logistic	NA	NA	1.82 (0.817, 4.04)
Standardized ATT	0.166 (0.034, 0.298)	1.75 (1.17, 2.63)	2.23 (1.21, 4.09)

3. Assumptions A1 and A2 holding are equivalent to

$$a - b = c - d$$

and

$$a/b = c/d.$$

We have that from the second equality that $a = bc/d$, so that from the first equality $bc/d - b = c - d$, which implies that $b(c/d - 1) = c - d$, which implies that either $c = d$ (and also $a = b$) or that $b = (c - d)/(c/d - d/d) = d$, and thus that $a = c$.

In summary, either $a = b$ and $c = d$, or $b = d$ and $a = c$. In practice, these constraints are not likely to hold.

5. Assumptions A1 and A3 holding are equivalent to

$$a - b = c - d$$

and

$$\frac{a(1-b)}{b(1-a)} = \frac{c(1-d)}{d(1-c)}.$$

These equalities imply that

$$\frac{ad}{bc} = \frac{(1-d)(1-a)}{(1-b)(1-c)} = \frac{1-d-a+ad}{1-b-c+bc},$$

which, by A1, implies that

$$\frac{ad}{bc} = \frac{1-b-c+ad}{1-b-c+bc}. \tag{7.1}$$

It is easy to show that

$$\frac{x+y}{x+z} = \frac{y}{z}$$

implies that $z = y$ or $x = 0$. This together with (7.1) implies that either $ad = bc$ or $b + c = 1$, which implies either that $a/b = b/c$ (i.e. that A2 holds) or that $a+d = b+c = 1$. If A1, A2, and A3 hold, then either $a = b$ and $c = d$, or $b = d$ and $a = c$. If instead $a + d = b + c = 1$, there is no simpler relationship between a,b,c, and d.

In summary, either $a = b$ and $c = d$, or $b = d$ and $a = c$, or $a+d = b+c = 1$.

8

Chapter 8

1. No it is not valid. Although $E(Y(a)) = E(Y|A = a)$, it is not true that $E(Y|S = s, A) = E(Y|S = s)$, and therefore

$$E(Y|A = a) = \Sigma_s E(Y|S = s, A = a)P(S = s|A = a),$$

   which will not generally equal (8.1).

3. We stored the dataset in `dat83`, and we estimated using `est83.r` together with the bootstrap.

```
> est83.r
function(data=dat83,ids=c(1:nrow(dat83)))
{
dat<-data[ids,]
tmp00<-(1-mean(dat$S[dat$A==0]))*
( mean(dat$Y[(dat$S==0)&(dat$A==0)])*(1-mean(dat$A)) +
  mean(dat$Y[(dat$S==0)&(dat$A==1)])*mean(dat$A) )

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

EY0<-tmp00+ tmp01

tmp10<-(1-mean(dat$S[dat$A==1]) )*
( mean(dat$Y[dat$S==0 & dat$A==0])*(1-mean(dat$A)) +
  mean(dat$Y[dat$S==0 & dat$A==1])* mean(dat$A) )

tmp11<-mean(dat$S[dat$A==1]) *
( mean(dat$Y[dat$S==1 & dat$A==0])*(1-mean(dat$A)) +
  mean(dat$Y[dat$S==1 & dat$A==1])*mean(dat$A) )

EY1<-tmp10 + tmp11
frontdoor<-EY1-EY0

beta<-lm(Y~A*X,data=dat)$coef
EX<-mean(dat$X)
standardization<-beta[2]+beta[4]*EX
c(frontdoor,standardization)
}
```

We found that the front-door method gave an estimate of 0.077 (0.070, 0.084) and that standardization gave 0.079 (0.060, 0.099). From `sim8ex3.r` we know that both estimators are unbiased. It is of interest that the confidence interval from the front-door method is quite a bit shorter than that from standardization. We speculate that the front-door method may be more efficient in general.

5. We can redraw the causal DAG as in Figure 8.1 to include the potential outcomes and note that all of the assumptions about the potential outcomes remain unchanged by the addition of W. Therefore, the theory still holds.

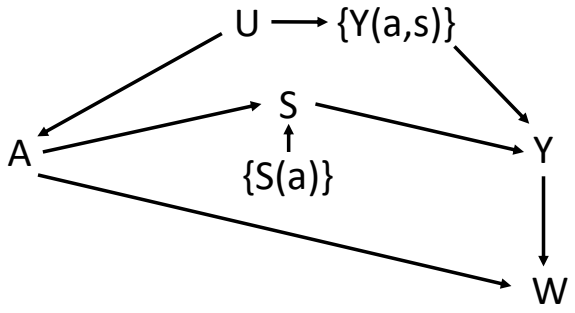

FIGURE 8.1: Causal DAG for Exercise 5 Solutions

We estimate the effect of AD_0 on A in the Double What-If study using the front-door approach with $A = AD_0$, $S = VL_0$, and $Y = A$, noting from our argument above that we do not need to worry about $W = VL_1$. We use the `frontdoor.r` code with the bootstrap, and find that $\hat{E}(Y(0)) = 0.248(0.215, 0.281)$, $\hat{E}(Y(1)) = 0.244(0.212, 0.276)$, and $RD = -0.0038(-0.029, 0.022)$. We can tell from the DAG that AD_0 and A are independent, and thus that the true effect is zero, and our results are consistent with that.

9

Chapter 9

1. The assumption imposed by the linear SNMM is that

$$E(Y|A=1,T) = E(Y(0)|A=1,T) + \beta. \tag{9.1}$$

The constraint is that β does not depend on whether T equals one or zero. When the placebo group cannot access the treatment, $T = 1$ whenever $A = 1$. Therefore, (9.1) reduces to

$$E(Y|A=1,T=1) = E(Y(0)|A=1,T=1) + \beta,$$

which imposes no constraint at all, because $E(Y|A = 1,T)$ and $E(Y(0)|A = 1,T = 1)$ are constants.

Similarly, the assumptions imposed by the loglinear and logistic SNMMs reduce to

$$\log E(Y|A=1,T=1) = \log E(Y(0)|A=1,T=1) + \beta$$

and

$$\text{logit}E(Y|A=1,T=1) = \text{logit}E(Y(0)|A=1,T=1) + \beta,$$

which are both free of constraints.

3. We used `est93.r` together with the bootstrap to analyze the data.

```
> est93.r
function(data=sim9ex3dat,ids=c(1:nrow(sim9ex3dat)))
{
dat<-data[ids,]
EYA1<-mean(dat$Y[dat$A==1])
tmp0<-mean(dat$Y[(dat$S==0)&(dat$A==1)])*(1-mean(dat$S[dat$A==0]))
tmp1<-mean(dat$Y[(dat$S==1)&(dat$A==1)])*mean(dat$S[dat$A==0])
EY0A1<-tmp0+tmp1
frontdoorATT<-EYA1-EY0A1
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]
```

```
ivATT<-mean(Deta[dat$A==1])-mean((Deta-dat$A*beta)[dat$A==1])
dat$H<-dat$U
EHA<-mean(dat$H[dat$A==1])
beta<-lm(Y~A*H,dat=dat)$coef
standATT<-beta[2]+beta[4]*EHA
c(frontdoorATT,ivATT,standATT)
}
```

TABLE 9.1
Estimates and 95% Confidence Intervals for the ATT of Exercise 3 Using Three Different Methods

Method	RD (95%CI)
Front-door	0.531 (0.496, 0.566)
Linear SNMM	0.811 (0.735, 0.887)
Standardization	0.518 (0.499, 0.538)

We observe that the estimates from the front-door approach the outcome-modeling standardization approach agree, but that the one from the linear SNMM is quite different. From `sim9ex3.r`, we know that the front-door approach and that standardization taking U as the sufficient confounder and using a nonparametric model are both valid. Therefore, we know that the linear SNMM is invalid for this example.

5. The assumption imposed by the loglinear SNMM is that

$$\log E(Y|A=1,T) = \beta + \log E(Y(0)|A=1,T).$$

Unlike in the logistic SNMM case for the previous example, the analytic formula for $E(Y(0)|A=1,T)$ is intractable. That is why, in the function `sim9ex5.r`, it is calculated using Monte Carlo methods (i.e. by simulation) based on a sample size of 1,000,000. We see that

```
logEYcA1T0<-log(mean(Y0[(A==1)&(T==0)]))+ beta
logEYcA1T1<-log(mean(Y0[(A==1)&(T==1)]))+ beta
```

which is later followed by

```
Y1T0<-rpois(n=nsim,lambda=exp(logEYcA1T0))
Y1T1<-rpois(n=nsim,lambda=exp(logEYcA1T1))
Y<-Y0*(1-A) + Y1T0*A*(1-T) + Y1T1*A*T
```

which shows us that the loglinear SNMM is satisfied. The code also shows

us that randomization and exclusion hold. The true value of β is 4.0, and it represents a log rate ratio:

$$\beta = \log E(Y|A=1,T) - \log E(Y(0)|A=1,T),$$

comparing $E(Y|A=1,T)$ to $E(Y(0)|A=1,T)$. We are more interested in comparing $E(Y|A=1)$ with $E(Y(0)|A=1)$. For computation, we use the code `est95.r` together with the jackknife.

```
> est95.r
function (data=sim9ex5dat)
{
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)
{
Ystar<-exp(Deta-A*betat)*(1+A*betat)
Astar<-A*exp(Deta-A*betat)
betat<-ivreg(formula=Ystar~Astar,instruments=~Z)$coef[2]
}
beta<-betat
EY1<-mean(exp(Deta)[A==1])
EY0<-mean(exp(Deta-A*beta)[A==1])
RD<-EY1-EY0
logRR<-log(EY1/EY0)
c(beta,EY0,EY1,RD,logRR)
}
```

The estimates are presented in Table 9.2.

TABLE 9.2
Estimation of ATT Using a Loglinear SNMM for Exercise 5

Measure	Estimate	95% CI
$\hat{\beta}$	4.02	(3.89, 4.15)
$\hat{E}(Y(0)\|A=1)$	4.39	(3.82, 4.96)
$\hat{E}(Y\|A=1)$	246	(244, 247)
$\hat{RD}$	241	(240, 243)
$\hat{RR}$	55.9	(49.1, 63.6)

We observe that the confidence interval for β, which is (3.89, 4.15), covers

the true value, 4.0, and it is fairly narrow. The other estimates are also quite precise, and there is a statistically signficant effect of A on Y in those with $A = 1$. We observe that with the same sample size (5,000) as in the logistic SNMM, the estimates for this example with the loglinear SNMM are much more precise.

10

Chapter 10

1. We estimated the propensity score and checked for overlap using the function

```
> plotprop.r
function (data=brfsslt65,ids=c(1:nrow(brfsslt65)))
{
pdf("H:\\Fundamentals of Causal Inference\\BRFSS\\BRFSS Chapter 10\\einsured.pdf")
dat<-data[ids,]
e<-fitted(glm(insured~rural+female+whitenh+blacknh+
hisp+multinh+gthsedu,family=binomial,data=dat))
a<-range(density(e[dat$insured==1])$y)
b<-range(density(e[dat$insured==0])$y)
a<-range(a,b)
plot(c(0,1),a,type="n",xlab="propensity score", ylab="density")
lines(density(e[dat$insured==1],bw=.05),lty=1)
lines(density(e[dat$insured==0],bw=.05),lty=2)
legend("topleft",c("insured=0","insured=1"),lty=c(2,1))
dev.off()
}
```

   The results are graphed in Figure 10.1. We see that the overlap is strong.

3. We use the function

```
> estand.r
function (data=brfsslt65,ids=1:nrow(brfsslt65))
{
dat<-data[ids,]
emod<-glm(insured~rural+female+whitenh+blacknh+
hisp+multinh+gthsedu,family=binomial,data=dat)
e<-fitted(emod)
lmod<-glm(flushot~insured+e,family=binomial,data=dat)
dat0<-dat1<-dat
dat0$insured<-0
dat1$insured<-1
dat0$e<-dat1$e<-e
EYhat0<-predict(lmod,newdata=dat0,type="response")
EYhat1<-predict(lmod,newdata=dat1,type="response")
EY0<-mean(EYhat0)
```

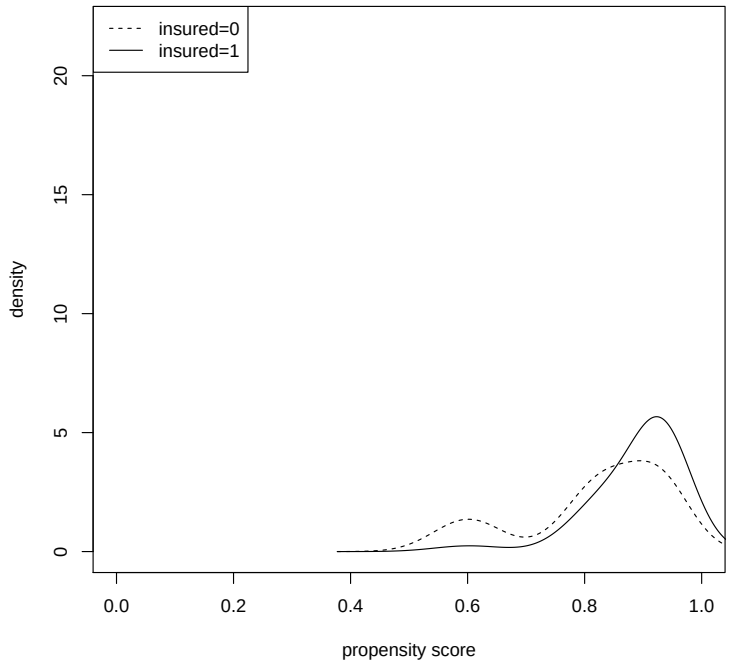

FIGURE 10.1: Checking for Overlap with the Flu Shot BRFSS Example

```
EY1<-mean(EYhat1)
rd<-EY1-EY0
logrr<-log(EY1/EY0)
c(EY0,EY1,rd,logrr)
}
```

together with the bootstrap, to produce the results presented in Table 10.1. We observe that they are very similar to those obtained by ordinary outcome-modeling standardization in Exercise 1 of Chapter 6.

TABLE 10.1
Outcome-model Standardization Using the Propensity Score of the Effect of Having Health Insurance on Receiving a Flu Shot

Measure	Estimate	95% CI
$\hat{E}(Y(0))$	0.214	(0.209, 0.219)
$\hat{E}(Y(1))$	0.440	(0.438, 0.442)
$\hat{RD}$	0.226	(0.220, 0.231)
$\hat{RR}$	2.05	(2.00, 2.11)

5. We implemented this with the R code

```
> match.r
function ()
{
SEe<-sqrt(var(e))
Match(Y=brfsslt65$flushot,Tr=brfsslt65$insured,X=e,
estimand="ATE",caliper=0.25, replace=T,ties=F)
}
> match.out<-match.r()
> matchsummary.r
function ()
{
summary.Match(match.out)
}
> matchbalance.r
function ()
{
MatchBalance(flushot~insured+rural+female+whitenh+blacknh+
hisp+multinh+gthsedu,data=brfsslt65,match.out=match.out)
}
```

which returned

```
> matchsummary.r()

Estimate...  0.24072
SE.........  0.0013656
T-stat.....  176.27
p.val......  < 2.22e-16

Original number of observations..............  215435
Original number of treated obs...............  189720
Matched number of observations...............  215435
Matched number of observations  (unweighted).  215435

Caliper (SDs)........................................   0.25
Number of obs dropped by 'exact' or 'caliper'  0
```

and

```
> matchbalance.r()

***** (V1) insured *****
                     Before Matching           After Matching
mean treatment........    0.94398                        1
mean control..........    0.83583                        0
std mean diff.........     47.029                      Inf

mean raw eQQ diff.....    0.10815                        1
med  raw eQQ diff.....          0                        1
max  raw eQQ diff.....          1                        1

mean eCDF diff........   0.054075                      0.5
med  eCDF diff........   0.054075                      0.5
max  eCDF diff........    0.10815                        1

var ratio (Tr/Co).....    0.38539                      NaN
T-test p-value........ < 2.22e-16             < 2.22e-16


***** (V2) rural *****
                     Before Matching           After Matching
mean treatment........    0.13201                  0.14418
mean control..........    0.15279                  0.14418
std mean diff.........    -6.1398                        0

mean raw eQQ diff.....   0.020783                        0
med  raw eQQ diff.....          0                        0
max  raw eQQ diff.....          1                        0

mean eCDF diff........   0.010392                        0
med  eCDF diff........   0.010392                        0
```

```
max  eCDF diff........   0.020783                        0

var ratio (Tr/Co).....    0.88517                        1
T-test p-value........ < 2.22e-16                        1


***** (V3) female *****
                        Before Matching           After Matching
mean treatment........    0.58306                     0.5306
mean control..........    0.49309                    0.53083
std mean diff.........     18.246                   -0.046505

mean raw eQQ diff.....   0.089968                  0.00023209
med  raw eQQ diff.....          0                           0
max  raw eQQ diff.....          1                           1

mean eCDF diff........   0.044981                  0.00011604
med  eCDF diff........   0.044981                  0.00011604
max  eCDF diff........   0.089962                  0.00023209

var ratio (Tr/Co).....     0.9726                     1.0001
T-test p-value........ < 2.22e-16                     0.63244


***** (V4) whitenh *****
                        Before Matching           After Matching
mean treatment........       0.77                     0.73838
mean control..........    0.71642                     0.73815
std mean diff.........     12.734                    0.052805

mean raw eQQ diff.....   0.053589                  0.00023209
med  raw eQQ diff.....          0                           0
max  raw eQQ diff.....          1                           1

mean eCDF diff........   0.026795                  0.00011604
med  eCDF diff........   0.026795                  0.00011604
max  eCDF diff........   0.053589                  0.00023209

var ratio (Tr/Co).....     0.8717                     0.99943
T-test p-value........ < 2.22e-16                     0.63244


***** (V5) blacknh *****
                        Before Matching           After Matching
mean treatment........   0.072064                    0.080943
mean control..........   0.087224                    0.080943
std mean diff.........    -5.8624                           0

mean raw eQQ diff.....   0.015159                           0
```

```
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       0

mean eCDF diff........    0.00758                       0
med  eCDF diff........    0.00758                       0
max  eCDF diff........    0.01516                       0

var ratio (Tr/Co).....    0.83992                       1
T-test p-value........ < 2.22e-16                       1


***** (V6) hisp *****
                       Before Matching           After Matching
mean treatment........   0.081923                  0.10001
mean control..........     0.1128                  0.10001
std mean diff.........    -11.258                        0

mean raw eQQ diff.....   0.030867                       0
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       0

mean eCDF diff........   0.015437                       0
med  eCDF diff........   0.015437                       0
max  eCDF diff........   0.030875                       0

var ratio (Tr/Co).....    0.75156                       1
T-test p-value........ < 2.22e-16                       1


***** (V7) multinh *****
                       Before Matching           After Matching
mean treatment........   0.022341                 0.024759
mean control..........    0.02647                 0.024759
std mean diff.........    -2.7939                        0

mean raw eQQ diff.....  0.0041231                       0
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       0

mean eCDF diff........  0.0020646                       0
med  eCDF diff........  0.0020646                       0
max  eCDF diff........  0.0041292                       0

var ratio (Tr/Co).....    0.84759                       1
T-test p-value........ 7.1667e-10                       1


***** (V8) gthsedu *****
                       Before Matching           After Matching
```

```
mean treatment........    0.74543                  0.67936
mean control..........    0.63262                  0.67936
std mean diff.........     25.897                        0

mean raw eQQ diff.....    0.11281                       0
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       0

mean eCDF diff........   0.056406                       0
med  eCDF diff........   0.056406                       0
max  eCDF diff........    0.11281                       0

var ratio (Tr/Co).....     0.8165                       1
T-test p-value........ < 2.22e-16                       1
```

We see that the average treatment effect assessed as a risk difference is estimated at 0.241 with a P-value less than 2.22e-16. We observe that all confounders are very balanced after matching. The estimated ATE is similar to, but slightly higher than, that using outcome-modeling standardization with the propensity score and also to that using the average of the effects within quartiles after merging the higher two.

11

Chapter 11

1. First we made the dataset using `mkepilave.r`

```
> mkepilave.r
function ()
{
epilave<-NULL
dat<-epil
for (i in dat$subject)
{
dati<-dat$y[dat$subject==i]
ave<-mean(dati)
trt<-dat$trt[dat$subject==i][1]
trt<-as.numeric(trt)-1
base<-dat$base[dat$subject==i][1]
lbase<-dat$lbase[dat$subject==i][1]
epilave<-rbind(epilave,c(base,lbase,trt,ave))
}
dimnames(epilave)[[2]]<-c("base","lbase","trt","ave")
data.frame(epilave)
}
```

   Then we analyzed the data using

```
> precisionepilave.r
function(data=epilave,ids=1:nrow(epilave))
{
dat<-data[ids,]
RD1<-summary(lm(ave~trt,data=dat))$coef[2]
RD2<-summary(lm(ave~trt+base,data=dat))$coef[2]
c(RD1,RD2)
}
```

   and the bootstrap. Not including `base`, we estimated -0.621 (-3.39, 2.15), and including it, we estimated -0.912 (-2.51, 0.685). Including `base` reduced the sampling variability and the length of the confidence interval.

3. We constructed and analyzed the data using

```
> mktoe.r
```

```
function ()
{
toe<-NULL
dat<-toenail
ids<-unique(dat$ID)
for (id in 1:length(ids))
{
i<-ids[id]
dati<-dat[dat$ID==i,]
y0<-dati$outcome[dati$visit==1]
yvek<-as.numeric(dati$outcome[-1])
y<-mean(yvek)
trt<-dati$treatment[1]
toe<-rbind(toe,c(y0,trt,y))
}
dimnames(toe)[[2]]<-c("y0","trt","y")
data.frame(toe)
}
> precisiontoe.r
function(data=toe,ids=1:nrow(toe))
{
dat<-data[ids,]
RD1<-summary(lm(y~trt,data=dat))$coef[2]
RD2<-summary(lm(y~trt+y0,data=dat))$coef[2]
c(RD1,RD2)
}
```

together with the bootstrap.

Not including `y0` led to an estimate and confidence interval of -0.031 (-0.098, 0.035), and including it led to -0.035 (-0.081, 0.012). Including `y0` did reduced the sampling variability and hence the length of the confidence interval.

5. To do this, we wrote the function `manyprecisionsim.r`:

```
> manyprecisionsim.r
function()
{
set.seed(999)
Nsim<-1000
leng1<-leng2<-cover1<-cover2<-NULL
precisionsim.r<-function ()
{
nsim<-90
V<-rnorm(nsim)
T<-rbinom(n=nsim,size=1,prob=0.5)
EY<-.5*T+ T*V
Y<-rnorm(n=nsim,mean=EY)
dat<-cbind(V,T,Y)
```

```
dat<-data.frame(dat)
dat
}
bootprecision.r<-function(data=dat)
{
out<-boot(data=dat,statistic=precisionsimdat.r,R=1000)
est<-summary(out)$original
SE<-summary(out)$bootSE
lci<-est-1.96*SE
uci<-est+1.96*SE
list(est=est,SE=SE,lci=lci,uci=uci)
}
for (i in 1:Nsim)
{
dat<-precisionsim.r()
out<-bootprecision.r(dat)
leng1tmp<-out$uci[1]-out$lci[1]
leng2tmp<-out$uci[2]-out$lci[2]
cover1tmp<-cover2tmp<-0
if ((out$uci[1] > 0.5)&(out$lci[1] < 0.5)) cover1tmp<-1
if ((out$uci[2] > 0.5)&(out$lci[2] < 0.5)) cover2tmp<-1
leng1<-c(leng1,leng1tmp)
leng2<-c(leng2,leng2tmp)
cover1<-c(cover1,cover1tmp)
cover2<-c(cover2,cover2tmp)
}
avelength1<-mean(leng1)
avelength2<-mean(leng2)
cover1<-mean(cover1)
cover2<-mean(cover2)
list(avelength1=avelength1,avelength2=avelength2,cover1=cover1,cover2=cover2)
}
```

We found that the average length when we do not include V is 1.007, and when including V it is 0.933. Therefore, we observe a reduction. The coverage of the two methods is about the same: not including V it is 0.942 (0.928, 0.956) and including V it is 0.94 (0.925, 0.955). Both confidence intervals include 95%. To compute the confidence intervals of the coverage, we used `R` as follows:

```
> sqrt(0.942*(1-0.942)/1000)
[1] 0.0073916
> 0.942-1.96*0.0073916
[1] 0.92751
> 0.942+1.96*0.0073916
[1] 0.95649
> sqrt(0.94*(1-0.94)/1000)
[1] 0.00751
> 0.94-1.96*0.00751
```

```
[1] 0.92528
> 0.94+1.96*0.00751
[1] 0.95472
```

12

Chapter 12

1. We compute

```
> summary(glm(owngun~conservative+white+gt65+female,data=gssguncc))
Call:
glm(formula = owngun ~ conservative + white + gt65 + female,
    data = gssguncc)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.2422     0.0280    8.65  < 2e-16
conservative   0.1434     0.0267    5.37  9.0e-08
white          0.1807     0.0277    6.52  9.8e-11
gt65           0.0424     0.0301    1.41     0.16
female        -0.1243     0.0244   -5.10  3.8e-07
> summary(glm(owngun~white+gt65+female,data=gssguncc))
Call:
glm(formula = owngun ~ white + gt65 + female, data = gssguncc)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2709     0.0278    9.76  < 2e-16
white         0.1967     0.0278    7.07  2.5e-12
gt65          0.0554     0.0303    1.83    0.068
female       -0.1247     0.0246   -5.07  4.5e-07
> summary(glm(conservative~white+gt65+female,data=gssguncc))
Call:
glm(formula = conservative ~ white + gt65 + female, data = gssguncc)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20058    0.02699    7.43  1.8e-13
white        0.11110    0.02707    4.10  4.3e-05
gt65         0.09108    0.02947    3.09    0.002
female      -0.00245    0.02391   -0.10    0.918
```

Using the difference method, we estimate 0.1967-0.1807 = 0.0160. Using the product method, we estimate 0.1111*0.1434 = 0.0159. Therefore, the methods agree. We used the bootstrap to estimate a confidence interval for the difference method programmed with

```
> meddiffgun.r
function(dat=gssguncc,ids=1:nrow(gssguncc))
{
```

```
data<-dat[ids,]
d1<-glm(owngun~white+gt65+female,data=data)$coef[2]
d2<-glm(owngun~white+conservative+gt65+female,data=data)$coef[2]
d1-d2
}
```

We estimate the NIE at 0.0159 (0.0060, 0.0258); it is small but statistically signficant.

3. We compute

```
> summary(glm(flushot~whitenh+rural+female+gthsedu,data=brfsslt65))
Call:
glm(formula = flushot ~ whitenh + rural + female + gthsedu, data = brfsslt65)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26054    0.00267    97.6   <2e-16
whitenh      0.05614    0.00242    23.2   <2e-16
rural       -0.03844    0.00301   -12.8   <2e-16
female       0.08074    0.00211    38.3   <2e-16
gthsedu      0.11042    0.00228    48.4   <2e-16
> summary(glm(flushot~whitenh+insured+rural+female+gthsedu,data=brfsslt65))
Call:
glm(formula = flushot ~ whitenh + insured + rural + female +
    gthsedu, data = brfsslt65)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10441    0.00356    29.3   <2e-16
whitenh      0.03682    0.00242    15.2   <2e-16
insured      0.21449    0.00328    65.3   <2e-16
rural       -0.03549    0.00298   -11.9   <2e-16
female       0.07653    0.00209    36.6   <2e-16
gthsedu      0.08586    0.00229    37.5   <2e-16
> summary(glm(insured~whitenh+rural+female+gthsedu,data=brfsslt65))
Call:
glm(formula = insured ~ whitenh + rural + female + gthsedu, data = brfsslt65)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72792    0.00173  419.74  < 2e-16
whitenh      0.09007    0.00157   57.29  < 2e-16
rural       -0.01378    0.00195   -7.05  1.8e-12
female       0.01963    0.00137   14.34  < 2e-16
gthsedu      0.11447    0.00148   77.28  < 2e-16
```

Using the difference method, the point estimate for the NIE is 0.05614-0.03682=0.01932, whereas using the product method, the point estimate is 0.09007*0.21449 = 0.019319. The methods agree. Using the bootstrap with the difference method computed with

```
> meddiff.r
```

```
function (data=brfsslt65,ids=1:nrow(brfsslt65))
{
dat<-data[ids,]
d1<-glm(flushot~whitenh+rural+female+gthsedu,data=dat)$coef[2]
d2<-glm(flushot~whitenh+insured+rural+female+gthsedu,data=dat)$coef[2]
d1-d2
}
```

we estimate 0.0193 (0.0185, 0.0202), which means that the NIE is statistically significant.

5.

13

Chapter 13

1. We modified the code in the book to use `scogdat` instead of `cogdat`.

   Results of applying marginal structural models with IPTW estimation to the modified data are presented in Table 13.1.

TABLE 13.1
IPTW of MSM parameters for the Sensitivity Analysis of the Hypothetical Cancer Clinical Trial

Parameter	Estimate	95% CI
β_0	0.261	(0.225, 0.297)
β_1	-0.008	(-0.061, 0.045)
β_2	0.639	(0.574, 0.703)
β_3	-0.491	(-0.641, -0.340)
$\beta_1 + \beta_3$	-0.499	(-0.639, -0.359)
$\beta_2 + \beta_3$	0.148	(0.013, 0.282)
$\beta_1 + \beta_2 + \beta_3$	0.140	(0.007, 0.272)

   We observe that the estimates of β_1 is not statistically significantly different from zero. Therefore, $A_1 = 1$ on its own does not appear to influence survival at two years versus administration of neither treatment. However, as β_2 is statistically significant, we see that administering $A_2 = 1$ on its own does increase survival at two years versus administration of neither treatment. This has changed with the sensitivity analysis. The statistical significance of the estimate of $\beta_1 + \beta_3$ suggests that administering $A_1 = 1$ followed by $A_2 = 1$ results in decreased survival relative to administering $A_1 = 0$ followed by $A_2 = 1$. The statistical significance of $\beta_2 + \beta_3$ suggests that administering $A_1 = 1$ followed by $A_2 = 1$ results in increased survival relative to administering $A_1 = 1$ and $A_2 = 0$. To compare joint administration to administration of neither treatment, we need to estimate the contrast $\beta_1 + \beta_2 + \beta_3$, which we see is just statistically significant and equal to 0.140 (0.007, 0.272). Thus, administration of both treatments is better than administeration of neither.

3. We modified the code in the textbook to use `scogdat`. We found

```
>  A1opt.r(A2opt.r(mkcogtab.r()))
  A2 H2 A1 Freq    prop A2opt propA2opt A1opt propA1opt
1  0  0  0  410 0.29268     1   1.00000     0   0.89947
2  1  0  0   30 1.00000     1   1.00000     0   0.89947
3  0  1  0  160 0.18750     1   0.66667     0   0.89947
4  1  1  0   30 0.66667     1   0.66667     0   0.89947
5  0  0  1  280 0.10714     1   0.50000     0   0.89947
6  1  0  1   20 0.50000     1   0.50000     0   0.89947
7  0  1  1  190 0.42105     0   0.42105     0   0.89947
8  1  1  1   70 0.28571     0   0.42105     0   0.89947
```

We see that the optimal A_1 is estimated as 0. Putting this together with the results of `A2opt`, we observe that the optimal dynamic treatment regime is to set $A_1 = 0$ and then to set $A_2 = 1$ regardless of H_2. The marginal survival probability following implementation of the optimal dynamic treatment regime is estimated at 0.899. Conditional on H_2, this survival probability increases to 1.0 if $H_2 = 0$ and decreases to 0.667 if $H_2 = 1$.

After modifying the rest of the code to run the bootstrap, we found

```
> bootsoptimal.out
$pA1opt
[1] 0
$lclpropA1opt
[1] 0.8468
$uclpropA1opt
[1] 0.95214
$lclpropA2optH20
[1] 1
$uclpropA2optH20
[1] 1
$lclpropA2optH21
[1] 0.49621
$uclpropA2optH21
[1] 0.83712
```

We observe that a_1^{opt} equals zero for 100% of the bootstrap samples. We now can attach confidence intervals to our estimated survival probabilities. The marginal survival probability following implementation of the optimal dynamic treatment regime is estimated at 0.8999 (0.847, 0.952). Conditional on H_2, this survival probability increases to 1.0 (1.0, 1.0) if $H_2 = 0$ and decreases to 0.667 (0.496, 0.837) if $H_2 = 1$.

5. We modified the code in the book to use `simdat`. The parameter estimates and their 95% confidence intervals are presented in Table 13.2.

TABLE 13.2
SNMM Estimation for the Simulated Apple Cider Vinegar Trial

Contrast	Estimate	95% CI
β_{20}	0.436	(0.328, 0.544)
$\beta_{20} + \beta_{22}$	0.449	(0.231, 0.668)
$\beta_{20} + \beta_{21}$	0.404	(0.234, 0.574)
$\beta_{20} + \beta_{21} + \beta_{22} + \beta_{23}$	0.386	(0.251, 0.522)
β_1	0.076	(-0.014, 0.165)

From estimation of β_1, we see that the effect of A_1 is on weightloss at 7 weeks is not statistically signficant when A_2 will not be administered afterwards. From estimation of the contrasts of β_2, we see that the effect of A_2 is to increase weightloss at 7 weeks when $A_1 = 0$ followed by no weightloss at 4 weeks ($\hat{\beta}_{20} = 0.436$), to increase weightloss at 7 weeks when $A_1 = 0$ followed by weightloss at 4 weeks ($\hat{\beta}_{20} + \hat{\beta}_{22} = 0.449$), to increase weightloss at 7 weeks when $A_1 = 1$ followed by no weightloss at 4 weeks ($\hat{\beta}_{20} + \hat{\beta}_{21} = 0.404$), and to increase weightloss at 7 weeks when $A_1 = 1$ followed by weightloss at weeks ($\hat{\beta}_{20} + \hat{\beta}_{21} + \hat{\beta}_{22} + \hat{\beta}_{23} = 0.386$). Therefore, the decision about whether to treat with $A_2 = 0$ or $A_2 = 1$ does not need to make use of information on A_1 and H_2.