Fundamentals of Causal Inference with R: Module 3 Front-Door and Instrumental Variables Methods

Babette Brumback

Professor Emerita
University of Florida, Biostatistics Department

Front-Door Method: Motivation

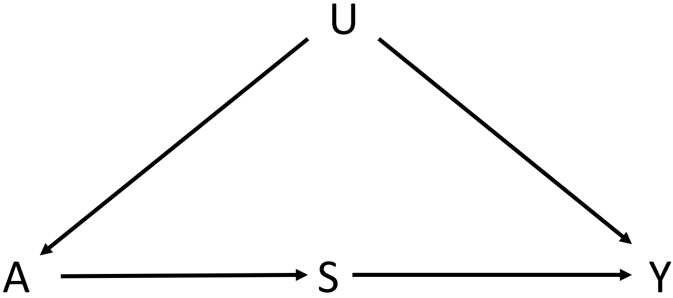
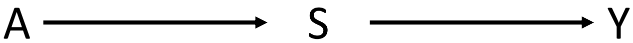


Figure 1: Front-door Causal DAG

Motivation

- The front-door method of Pearl (1995) provides a way to adjust for confounding by unmeasured variables, represented by U, of the effect of A on Y in Figure 1.
- Like the backdoor method, the front-door method is quite clever, but to date, and unlike the backdoor method, it is not in popular use.
- Pearl and MacKenzie (2018) present just one published example of its use in comparing analyses of an observational study with randomized clinical trial (RCT) results in which the front-door method produces results that agree with those from the RCT but the backdoor method does not.
- However, in that example, as in so many others, it is difficult to justify the independence assumptions encoded in Figure 1; A indicates signing up for a job-training program offered by the Department of Labor, S indicates showing up for that program, and Y represents earnings over the subsequent 18 months.
- U indicates motivation, which is an unmeasured confounder.
- However, as there is likely an arrow from U to S, the front-door method would not be expected to work.
- Because it appears to work, either that arrow must be weak, or there must be differences between the RCT population and the observational study population that counterbalance it.

Motivation



A

S

Y

Figure 2: S is a Surrogate Marker

Motivation

- To motivate the front-door method, it is helpful to introduce the concept of a *surrogate marker.*
- In Figure 2, S is a surrogate marker for the effect of A on Y.
- Note that there may be other causes of Y that are independent of A and S, or causes of S that are independent of A; these need not be depicted in the DAG.
- If there are no such other causes, then S is a *perfect surrogate marker*, and testing the effect of A on Y is effectively identical to testing the effect of A on S.
- Surrogate markers are particularly useful when Y is a long-term clinical outcome and S occurs early on.
- For example, suppose we have a randomized clinical trial in a population of patients with high cholesterol, where A indicates initiation of treatment with a statin, S indicates a healthy lipid profile, and Y indicates subsequent myocardial infarction or stroke.
- In this example, S is not a perfect surrogate, because there are other causes of heart attacks and strokes besides unhealthy lipid levels, and possibly there are also other causes of healthy lipid levels besides statins.
- However, it is plausible that statins prevent heart attacks and strokes solely by way of improving lipid levels.
- Initial trials could therefore focus solely on the surrogate marker, avoiding the lengthy trial duration a focus on clinical outcomes of real interest would require.

Motivation

- When S is a surrogate marker and the causal DAG of Figure 2 holds, we have that

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

  because Y is independent of A given S.
- Additionally, as there are no confounders for the effect of A on Y, we have that

$$E(Y(a)) = E(Y|A=a) = \Sigma_s E(Y|S=s)P(S=s|A=a). \qquad (1)$$

- Improved cholesterol can be thought of as a *specific* mechanism of the effect of statins; that is, it is the dominant effect of statins on the outcome.
- For treatments with more than one effect on the outcome, S could be a partial surrogate marker, as in Figure 3.

Motivation

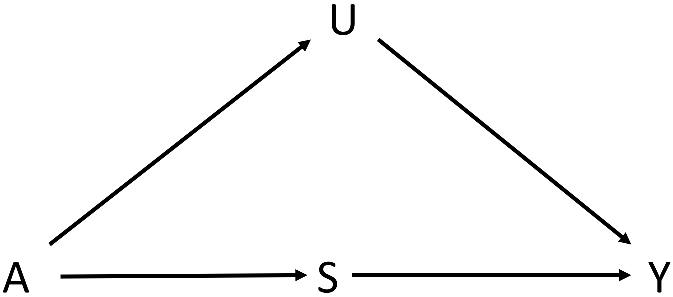


Figure 3: S is a Partial Surrogate Marker

Motivation

- The surrogate marker structure of Figure 2 plausibly arises in many RCTs of new therapeutic agents with specific mechanisms.
- Suppose such a trial confirmed a benefit of such a treatment in terms of improving the surrogate outcome, and that the treatment gained approval on this basis.
- A natural next step is to quantify the effect of the treatment on the clinical outcome of real interest.
- Although the RCT may be too small for this endeavor, a post-marketing observational study examining treatment use, surrogate outcome, and clinical outcome could conceivably be characterized by the front-door causal DAG in Figure 1.
- One needs to argue both that A does not affect Y except through S (i.e. that the mechanism of the therapeutic agent is specific), and that U does not affect S.
- In an observational study of statin use, restricted to a population with high cholesterol, U could represent family history of heart attack or stroke.
- This would plausibly affect statin use as well as subsequent heart attack or stroke.
- To use the front-door method, we would need to argue that it does not affect cholesterol reduction except via statin use.
- We would also need to rule out dietary restrictions as a successful alternative, which is perhaps somewhat plausible.
- Furthermore, we would need to consider other possible variables in U and also rule those out as direct causes of S.

Motivation

- The surrogate marker structure could instead pertain to unintended effects of a treatment.
- Second generation anti-psychotics (SGAs), which we code as A, are helpful treatments for mental illnesses including schizophrenia and bipolar disorder.
- However, they are well known to cause substantial weight gain S, with a risk of subsequent diabetes Y.
- Nevertheless, there is evidence that schizophrenia and diabetes were associated long before the introduction of SGAs.
- Given electronic health record data on A, S, and Y from a mental health clinic, one might attempt to apply the front-door method to adjust for possible unmeasured confounders U, such as severity of schizophrenia.
- It is plausible that severity of schizophrenia would not cause substantial weight gain in the time-frame of the study, which was just one year in the study that documented diabetes as an effect.
- Again, one would need to rule out other possible variables in U – affecting both SGA use and subsequent diabetes – as direct causes of substantial weight gain.

Motivation

- In much of science, we have much more information on the total effect of A on Y than we have on its mechanism.
- This could account, in large part, for the currently limited application of the front-door method.
- When we know the total effect but are interested in whether S serves as a mediator, i.e. in whether S is part of the mechanism, we would conduct a *mediation analysis*.
- Mediation analyses are the subject of Chapter 12.

Theory and Method

- When the front-door causal DAG holds, the front-door theorm of Pearl (1995) states that we can modify equation (1) to

  $$E(Y(a)) = \Sigma_s P(S = s|A = a)\Sigma_{a'} E(Y|S = s, A = a')P(A = a'), \qquad (2)$$

  where we have replaced

  $$E(Y|S = s)$$

  with

  $$\Sigma_{a'} E(Y|S = s, A = a')P(A = a'),$$

  using backdoor standardization.
- The front-door method estimates $E(Y(a))$ via

  $$\hat{E}(Y(a)) = \Sigma_s \hat{P}(S = s|A = a)\Sigma_{a'} \hat{E}(Y|S = s, A = a')\hat{P}(A = a').$$

Theory and Method

- One can use a potential outcomes framework to prove the front-door theorem.
- We let $S(a)$ denote the potential outcome corresponding to S when A is set equal to a.
- We let $Y(a, s)$ denote the potential outcome corresponding to Y when A is set equal to a and then S is set equal to s.
- We let $Y(a)$ denote the potential outcome corresponding to Y when A is set equal to a.
- We then make the following six assumptions about these potential outcomes, where a and s represent any possible values of A and S.

Theory and Method

1. We assume $Y(A) = Y$, $S(A) = S$, and $Y(A, S) = Y$. These are consistency assumptions.
2. $Y(a, S(a)) = Y(a)$, and $Y(a, S(a)) = Y(a, s)$ when $S(a) = s$. These are similar in spirit to consistency assumptions. The first states that had S been set equal to the value it would have taken were A set equal to a, we would have observed the potential outcome $Y(a)$.
3. $Y(a, s) = Y(\cdot, s)$, that is, it does not depend on a. We let $Y(\cdot, s)$ represent the potential outcome corresponding to Y when S is set equal to s. This assumption corresponds to no directed path from A to Y except through S. In particular, it rules out the partial surrogate DAG in Figure 3.
4. We assume that $S(a) \amalg A$, that is, there are no confounders for the effect of A on S.
5. We assume that $Y(\cdot, s) \amalg S|A$, that is, A is a sufficient confounder for the effect of S on Y.
6. We assume $Y(a, s) \amalg S(a)$. This is the key front-door assumption. It follows from the front-door causal DAG of Figure 1, where there is no arrow from U to S.

Theory and Method

- It is perhaps easier to grasp from the version of the front-door causal DAG depicted in Figure 4, which includes the potential outcomes.

Theory and Methods

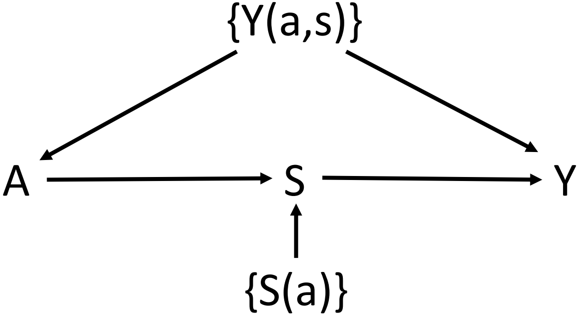


Figure 4: Front-Door Causal DAG Including Potential Outcomes

Theory and Methods

The proof of the front-door theorem at (2) is as follows.

$$E(Y(a)) = E(Y(a, S(a)))$$

by assumption 2.

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

by the law of total expectation.

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

by assumption 2.

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

by the key front-door assumption 6.

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

by assumption 3.

$$E(Y(\cdot, s)) = \Sigma_{a'} E(Y|S = s, A = a')P(A = a')$$

by assumption 5 and backdoor standardization.

$$P(S(a) = s) = P(S = s|A = a)$$

by assumption 4 and consistency Then substitute the last two relations into the third to last relation to finish proving the front-door theorem.

Simulated Example

- It is helpful to use a simulated binary dataset to better understand the theory and the method.
- One can think in terms of the SGA example, in which A indicates administration of SGA, S indicates substantial weight gain, and Y indicates diabetes at one year.
- We simulate according to the causal DAG in Figure 4.
- We let $P(Y(\cdot, 0) = 1) = 0.05$ and $P(Y(\cdot, 1) = 1) = 0.2$.
- We let $P(S(0) = 1) = 0.05$ and $P(S(1) = 1) = 0.9$.
- We let $P(A = 1|Y(\cdot, 0), Y(\cdot, 1)) = (1 - Y(\cdot, 1)) * 0.1 + Y(\cdot, 1) * 0.8$.
- Simulating these variables determines S and Y.
- Using `R`, the data are simulated with `sim1.r`.

Simulated Example

```
> sim1.r
function ()
{
set.seed(555)
nsim=10000
Ydot0<-rbinom(n=nsim,size=1,prob=0.05)
Ydot1<-rbinom(n=nsim,size=1,prob=0.2)
probA=(1-Ydot1)*.1 + Ydot1*.8
A<-rbinom(n=nsim,size=1,prob=probA)
S0<-rbinom(n=nsim,size=1,prob=0.05)
S1<-rbinom(n=nsim,size=1,prob=0.9)
S<-(1-A)*S0 + A*S1
Y<-(1-S)*Ydot0 + S*Ydot1
dat<-cbind(A,S,Y,Ydot0,Ydot1,S0,S1)
data.frame(dat)
}
```

Simulated Example

- Using our theory, we can calculate

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

  so that

$$E(Y(0)) = 0.05 * 0.95 + 0.2 * 0.05 = 0.0575$$

  and

$$E(Y(1)) = 0.05 * 0.1 + 0.2 * 0.9 = 0.185.$$

- We could instead try using equation (2) to calculate $E(Y(a))$, but this involves very many tedious applications of the multiplication rule and the law of total probability.
- Instead, we use `frontdoor.r` to estimate it, as follows.

Simulated Example

```
> frontdoor.r
function ()
{
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
list(EY0=EY0,EY1=EY1)
}
> frontdoor.r()
$EY0
[1] 0.055847
$EY1
[1] 0.18537
```

Simulated Example

- We observe that the estimates are very close to the true values, as expected given the large sample size of 10,000.
- We can compare these estimates to the unadjusted estimates of $E(Y|A=0)$ and $E(Y|A=1)$, which are

```
> mean(dat$Y[dat$A==0])
[1] 0.048892
> mean(dat$Y[dat$A==1])
[1] 0.61367
```

Simulated Example

- We see that the estimate of $E(Y|A=1)$, in particular, is far away from $E(Y(1))$.
- There is substantial confounding, which the front-door method eliminated.
- In summary, the front-door method represents a clever way to adjust for unmeasured confounding.
- However, the assumptions of the causal DAG in Figure 1 must hold.
- So far, the front-door method has not really found its way into applications.
- However, it is possible that with increasing awareness, it may prove useful for post-marketing follow-up of RCTs relying on surrogate marker outcomes, as in the statin example, or for safety analyses, as in the SGA example.

Introduction: Instrumental Variables Methods

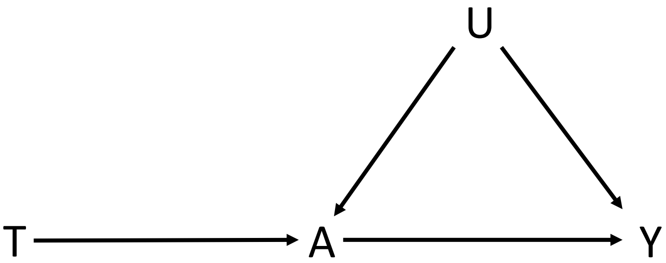


Figure 5: T is an Instrumental Variable for the Effect of A on Y

Introduction

- When the instrumental variables DAG of Figure 5 holds, we have yet another way to adjust for unmeasured confounding of the effect of A on Y.
- The variable T is as an *instrumental variable* if it (a) affects A, and (b) the only directed path from T to Y is through A, and (c) any backdoor paths from T to A or from T to Y are blocked by measured confounders in H.
- Assumption (b) is called the *exclusion assumption*, because it excludes the arrow directly from T to Y.
- When an instrumental variable exists, we can use it to adjust for unmeasured confounding using either *principal stratification* or *structural nested mean models*, under certain additional assumptions.
- We assume T is randomized so that there is no need for H, which we have therefore not included in the DAG.
- We also assume binary A and T.
- The methods can be modified to include H or to include more general A and T.

Introduction

- For example, in the What-If Study, we can use naltrexone T as an instrumental variable for the effect of reduced drinking A on unsuppressed viral load Y, provided that naltrexone does not cause unsuppressed viral load except via reduced drinking.
- In the Double What-If Study, we can be certain that naltrexone T is an instrumental variable for the effect of reduced drinking A on unsuppressed viral load VL_1, because the data were generated such that T only causes VL_1 through A.

Introduction

- We enlist the following notation.
- Let $Y(t, a)$ be the potential outcome for Y assuming we set $T = t$ and then $A = a$.
- We assume consistency.
- Due to the exclusion assumption, $Y(t, a) = Y(a)$.
- We let $A(t)$ be the potential outcome for A assuming we set $T = t$.

Introduction

- Suppose participants randomized to treatment with T either comply with the assigned treatment regime or not.
- Let the treatment actually taken be recorded in A.
- In the What-If and Double What-If Studies, we suppose that the entire purpose of treatment with naltrexone is to reduce drinking.
- Therefore we let reduced drinking equate to treatment actually taken.
- We note that A is a post-randomization event.
- When A does not equal T, two historical methods for assessing the effect of treatment on Y are called the *as-treated* analysis and the *per-protocol* analysis.
- In the as-treated analysis, we compare $E(Y|A=1)$ with $E(Y|A=0)$, and we lose the benefit of randomization.
- In the per-protocol analysis, we let $Z=1$ if $A=T$ and use ordinary stratification on $Z=1$ to compare $E(Y|T=1, Z=1)$ with $E(Y|T=0, Z=1)$.
- While this might look at first like a comparison of randomized groups, the event $Z=1$ is a post-randomization event; therefore, Z, unlike a baseline covariate, cannot be expected to be balanced across the two treatment groups.

Introduction

- For example, suppose sicker patients assigned to $T = 1$ comply, and have $A = 1$ and $Z = 1$, whereas some healthy patients avoid it due to side effects, and have $A = 0$ and $Z = 0$, while all patients assigned $T = 0$ comply with the assignment, and have $A = 0$ and $Z = 1$.
- In this case, while the as-treated and the per-protocol analyses are different, they both compare sicker patients on $T = 1$ with a blend of sicker and healthy patients on $T = 0$.
- Treatment will not appear as effective as it really is.

Introduction

- For these reasons, many studies rely on the *intent-to-treat effect* (ITT), $E(Y|T=1) - E(Y|T=0)$, which measures the effect of the intention to treat with $A=1$ versus $A=0$, rather than the effect of treatment actually received.
- Due to randomization and consistency, this equates to the causal effect

$$\text{ITT} = E(Y(1, A(1)) - E(Y(0, A(0)).$$

Complier Average Causal Effect and Principal Stratification

- Principal stratification allows us to assess the causal effect of treatment actually received, under additional assumptions.
- Principal stratification classifies participants according to the potential occurence of a post-randomization event.
- It can be a useful alternative to ordinary stratification.
- With principal stratification, we define four principal strata of participants according to their potential outcomes $A(t)$.
- If $A(0) = A(1) = 0$, then the participant is a *never taker*, and will not take the treatment regardless of randomized assignment.
- If $A(0) = A(1) = 1$, then the participant is an *always taker*, and will always take the treatment regardless of randomized assignment.
- If $A(0) = 0$ and $A(1) = 1$, then the participant is a *complier*, and will comply with whichever treatment was assigned.
- Finally, if $A(0) = 1$ and $A(1) = 0$, then the participant is a *defier*, and will refuse to comply with whichever treatment was assigned.

Complier Average Causal Effect and Principal Stratification

- Let $C = 1$ indicate a complier, i.e. that $A(t) = t$.
- Note that because (a) T is randomized, (b) C is a pre-randomization variable, and (c) $Y(t, a) = Y(a)$, it is reasonable to assume

  $$Y(a) \perp\!\!\!\perp T | C, \tag{3}$$

  that is, that T is independent of the potential outcomes $Y(a)$ within strata defined by C.
- If instead, $Y(t, a) \neq Y(a)$, then $Y(a)$ could lie on the path from T to A, in which case (3) would not hold.
- We also have that

  $$T \perp\!\!\!\perp C$$

  because C is a pre-randomization variable.

Complier Average Causal Effect and Principal Stratification

- The *complier average causal effect* (CACE) is defined as the average effect of treatment in the compliers, that is, as

$$\begin{aligned} \text{CACE} &= E(Y(1)|C=1) - E(Y(0)|C=1) \\ &= E(Y|T=1, C=1) - E(Y|T=0, C=1), \end{aligned} \tag{4}$$

  where the relation follows from consistency and assumption (3).
- The right-hand side of (4) shows the CACE as a stratified treatment effect, but $C = 1$ defines a principal stratum instead of an ordinary stratum.
- We cannot observe C for all participants.
- When $A(t) \neq t|T = t$ we know that the participant is a *non-complier*, but when $A(t) = t|T = t$ we do not know whether $A(1-t) = 1-t$ or not, and thus we cannot conclude that the participant is a complier.
- Furthermore, if we are willing to assume that there no defiers, we can observe the proportion of always-takers as $E(A = 1|T = 0)$, and similarly we can observe the proportion of never-takers as $E(A = 0|T = 1)$, but we cannot observe the proportion of compliers $P(C = 1)$.
- Ordinary stratification assesses the treatment effect within a stratum defined by an observed variable, such as Z, whereas principal stratification assesses the treatment effect within a a stratum defined by an unobserved variable, such as C.

Complier Average Causal Effect and Principal Stratification

- To estimate the CACE, we go ahead and assume that there are no defiers, i.e. participants with $A(t) = 1 - t$.
- This is also referred to as the *monotonicity assumption*, because it is equivalent to $A(t) \geq A(t-1)$, e.g. that $A(t)$ is a monotonic function of t.
- Assuming exclusion and no defiers implies that

$$E(Y|T=1, C=0) = E(Y|T=0, C=0), \tag{5}$$

because no defiers means that $C = 0$ includes only never-takers and always takers, and exclusion ensures that randomization to T equal 1 or 0 cannot affect the outcome of the never-takers and the always-takers, because T does not affect A in those subgroups.

Complier Average Causal Effect and Principal Stratification

- Therefore, we have that

  $$E(Y|T=1) = E(Y|T=1, C=1)P(C=1) + E(Y|T=1, C=0)(1-P(C=1)),$$

  and

  $$E(Y|T=0) = E(Y|T=0, C=1)P(C=1) + E(Y|T=0, C=0)(1-P(C=1)).$$

- Therefore by (5),

  $$E(Y|T=1) - E(Y|T=0) =$$
  $$E(Y|T=1, C=1)P(C=1) - E(Y|T=0, C=1)P(C=1),$$

  so that the CACE equals

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

Complier Average Causal Effect and Principal Stratification

- We note that no defiers implies

  $$P(C=1)+P(A=0|T=1)+P(A=1|T=0)=1,$$

  where the latter two quantities are the proportions of never takers and always takers, respectively.

- Thus,

  $$\begin{aligned} P(C=1) &= 1-P(A=0|T=1)-P(A=1|T=0) \\ &= P(A=1|T=1)-P(A=1|T=0). \end{aligned}$$

- Therefore, the CACE equals

  $$\text{CACE}=\frac{E(Y|T=1)-E(Y|T=0)}{P(A=1|T=1)-P(A=1|T=0)}, \tag{6}$$

  an expression phrased entirely in terms of observed variables.

Complier Average Causal Effect and Principal Stratification

- Therefore we can estimate it via

$$\hat{\text{CACE}} = \frac{\hat{E}(Y|T=1) - \hat{E}(Y|T=0)}{\hat{P}(A=1|T=1) - \hat{P}(A=1|T=0)}. \tag{7}$$

- The sampling distribution of $\hat{\text{CACE}}$ can be estimated using the bootstrap.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- A drawback of the CACE is that it applies only to the compliers, a subgroup of the population that we cannot even identify.
- An alternative method targets the average effect of treatment on the treated (ATT), which we have already studied.

$$ATT = E(Y(1) - Y(0)|A = 1) = E(Y - Y(0)|A = 1).$$

- To estimate the ATT using the instrumental variable, T, we introduce structural nested mean models.
- We start with the linear structural nested mean model (SNMM):

$$E(Y - Y(0)|A, T) = A\beta. \tag{8}$$

- We observe that $Y - Y(0)$ is assumed to be mean independent of T given A.
- This assumption for the ATT replaces the assumption of no defiers for the CACE.
- It implies that in the subset with $A = 1$, any effect modifiers of $Y - Y(0)$ are balanced across the $T = 0$ and $T = 1$ groups. This is a strong assumption.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- We also need the non-causal linear model

$$E(Y|A, T) = D\eta, \tag{9}$$

where D is a function of A and T.

- We have that A and T are binary, and we can use the saturated model with $D = (1, A, T, A * T)$, so that this step does not add any additional assumptions.
- Combining (8) and (9), we have that

$$D\eta - A\beta = E(Y(0)|A, T),$$

and thus

$$E_{A|T}(D\eta - A\beta) = E(Y(0)|T) = E(Y(0)), \tag{10}$$

where the last equality is similar to (3), and results from T being randomized and from the exclusion assumption $Y(t, a) = Y(a)$, which entails that $Y(a)$ cannot lie on the path from T to A.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- Letting $E(Y(0)) = \alpha$, multiplying both sides by $(1, T)^T$, and then taking expectations of both sides of (10) results in the *instrumental variables estimating equation*

$$E\left[(1, T)^T(D\eta - A\beta - \alpha)\right] = 0. \tag{11}$$

- To solve (11), we first estimate η by $\hat{\eta}$ using regression with the linear model (9), and then we solve

$$\Sigma_{i=1}^{n}(1, T_i)^T(D_i\hat{\eta} - A_i\beta - \alpha) = 0 \tag{12}$$

for β and α using instrumental variables regression.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- The ivreg function in the AER library in R solves the estimating equation

$$\Sigma_{i=1}^{n}(1, T_i)^T(Y_i^* - A_i^*\beta - \alpha) = 0 \tag{13}$$

  for β and α.
- Therefore, we can solve (12) by letting $Y^* = D\hat{\eta}$ and $A^* = A$.
- We provide examples using the software shortly, but it is not hard to show that for binary T and A, the β which solves (13) is

$$\hat{\text{ATT}} = \hat{\beta} = \frac{\hat{E}(Y|T=1) - \hat{E}(Y|T=0)}{\hat{P}(A=1|T=1) - \hat{P}(A=1|T=0)}, \tag{14}$$

  which happens to be identical to $\hat{\text{CACE}}$ at (7).
- The interpretation, however, is quite different.
- Note that for the linear SNMM, $E(Y - Y(0)|A=1, T) = \beta$ does not depend on T, and that is why β equals the ATT, $E(Y - Y(0)|A=1)$.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- The linear structural nested mean model can be generalized to

$$h(E(Y|A,T)) - A\beta = h(E(Y(0)|A,T)), \tag{15}$$

  where $h(\cdot)$ is either $\log(\cdot)$ (for a loglinear structural nested mean model) or logit$(\cdot)$ (for a logistic structural nested mean model), and it is paired with the non-causal model

$$h(E(Y|A,T)) = D\eta, \tag{16}$$

  where D is as before in (9).
- From (15) and (16) it follows that

$$h^{-1}(D\eta - A\beta) = E(Y(0)|A,T),$$

  so that

$$E_{A|T}\left(h^{-1}(D\eta - A\beta)\right) = E(Y(0)|T) = E(Y(0)).$$

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- Letting $E(Y(0)) = \alpha$, multiplying both sides by $(1, T)^T$, and taking expectations of both sides yields

$$E\left[(1, T)^T \left(h^{-1}(D\eta - A\beta) - \alpha\right)\right] = 0. \tag{17}$$

- To solve (17), we first solve the loglinear or logistic estimating equation associated with (16) to estimate η with $\hat{\eta}$, and then we solve (17) by iteratively linearizing the equation about current estimates β_t and α_t using a Taylor series approximation and solving an instrumental variables estimating equation of the form (13).

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- Specifically, when $h(\cdot) = \log(\cdot)$,

  $$Y^* = \exp(D\hat{\eta} - A\beta_t)(1 + A\beta_t)$$

  and

  $$A^* = A\exp(D\hat{\eta} - A\beta_t).$$

- When $h(\cdot) = \text{logit}(\cdot)$,

  $$Y^* = \text{expit}(D\hat{\eta} - A\beta_t)\,(1 + A\beta_t(1 - \text{expit}(D\hat{\eta} - A\beta_t)))$$

  and

  $$A^* = A\text{expit}(D\hat{\eta} - A\beta_t)(1 - \text{expit}(D\hat{\eta} - A\beta_t)).$$

- Shortly, we will present `R` code that implements these estimators.
- We note that the general formulation applies to the linear structural nested mean model as well, with $h(\cdot)$ as the identity function.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- Although the parameter β can be interpreted conditionally on $A=1$ and T as a risk difference, log relative risk, or log odds ratio comparing the causal effect of $A=1$ versus $A=0$, we are interested in the ATT expressed as a risk difference, log relative risk, or log odds ratio, which is conditional on $A=1$ but not on T.
- We have already shown that for the linear SNMM, β is the ATT expressed as a risk difference.
- For the loglinear SNMM, β is the log of the ATT expressed as a relative risk.
- To see this, let $h(\cdot)$ be the $\log(\cdot)$ function in equation (15) and let $A=1$. We have that
$$\frac{E(Y|A=1,T)}{E(Y(0)|A=1,T)}=\exp(\beta).$$
- Using the double expectation theorem,
$$\frac{E(Y|A=1)}{E(Y(0)|A=1)}=\frac{\Sigma_t E(Y|A=1,T=t)P(T=t|A=1)}{\Sigma_t E(Y(0)|A=1,T=t)P(T=t|A=1)},$$
which equals
$$\frac{\Sigma_t \exp(\beta)E(Y(0)|A=1,T=t)P(T=t|A=1)}{\Sigma_t E(Y(0)|A=1,T=t)P(T=t|A=1)}=\exp(\beta).$$

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- For all of the SNMMs, we can estimate $E(Y|A=1)$ simply as $\hat{E}(Y|A=1)$ since it is a function of observed data.
- For $E(Y(0)|A=1)$ we have

  $$E(Y(0)|A=1) = E_{T|A=1}\left(h^{-1}(D\eta - A\beta)\right),$$

  which we can estimate as the mean of $h^{-1}(D\hat{\eta} - A\hat{\beta})$ in the participants with $A=1$.
- We will also show how to implement this with `R` code.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- The risk difference, log relative risk, and log odds ratio measures of the ATT are readily obtained from the estimates of $E(Y|A=1)$ and $E(Y(0)|A=1)$.
- To estimate sampling distributions, we would typically turn to the bootstrap.
- However, the SNMM estimating equations are tricky.
- For the linear SNMM, the solution could be outside the range of $[-1, 1]$, which is non-sensical for risk differences, and for the loglinear and logistic SNMMs, sometimes the estimating equation has no solution.
- This occurs if the model does not fit the data very well. Even if the model fits the sampled data well, it may not fit data that are resampled using the bootstrap.
- Therefore, the bootstrap is problematic, particularly for the loglinear and logistic SNMMs.
- To circumvent this, we turn to the *jackknife*, which is a simpler alternative to the bootstrap.

Average Effect of Treatment on the Treated and Structural Nested Mean Models

- The jackknife resamples the data $i = 1, \ldots, n$ times, where the i^{th} resample includes the original sample except for participant i; then the resulting estimates are combined to obtain a good estimator of the standard error.
- By its very nature, the jackknife produces resampled datasets that are very close to the original data.
- Therefore, it would be highly unlikely to find a resampled dataset leading to failure of the loglinear and logistic SNMM estimating equations, provided that solutions exist for the original data.
- The `resample` package in `R` contains the `jackknife` function, which we will show how to use.

Examples

- First, we consider the What-If? Study, in which we are interested in the effect of reduced drinking A on unsuppressed viral load Y.
- The `R` code we will present in `iv.r` and `bootiv.r` analyzes the data.
- A challenge with these data is that the proportion with $A = 1$ is almost the same in the $T = 0$ and $T = 1$ groups (62.35% versus 65%).
- When this occurs, T is called a *weak instrument* for the effect of A on Y.
- Furthermore, for some bootstrap samples, the proportions are identical and the denominator of the $C\hat{A}CE$ and $A\hat{T}T$ equals zero, so that the estimator does not exist.
- For that reason, we use only 100 bootstrap samples instead of 1000, so that we could rerun the program until we obtained a group of 100 bootstrap samples with valid estimates.
- Typically, we would not conduct an analysis with a weak instrument, but we include it here for illustration of the difficulties.

Examples

```
> xtabs(~T+A,whatifdat)
   A
T    0  1
  0 32 53
  1 28 52
> prop.table(xtabs(~T+A,whatifdat),1)
   A
T          0       1
  0 0.37647 0.62353
  1 0.35000 0.65000
> iv.r
function(data=whatifdat,ids=c(1:nrow(whatifdat)))
{
dat<-data[ids,]
ITT<-mean(dat$Y[dat$T==1])-mean(dat$Y[dat$T==0])
denom<-mean(dat$A[dat$T==1])-mean(dat$A[dat$T==0])
IV<-ITT/denom
c(ITT,IV)
}
```

Examples

```
> bootiv.r
function ()
{
out<-boot(data=whatifdat,statistic=iv.r,R=100)
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)
}
> bootiv.r()
$est
[1] 0.0073529 0.2777778
$SE
[1]  0.071978 13.289620
$lci
[1]  -0.13372 -25.76988
$uci
[1]  0.14843 26.32543
```

Examples

- We see that the ITT is positive, at 0.0074, but with a very wide confidence interval (-0.134, 0.148).
- The CACE and ATT are estimated at 0.278, which suggests that reduced drinking increases viral load, in contradiction to our results using standardization or difference-in-differences.
- The weak instrument causes the discrepancy, and it also results in the very wide confidence interval of (-25.77, 26.32), which includes nonsensical risk differences.

Examples

- The Double What-If Study provides a better basis for comparison of the CACE and ATT with the methods of standardization and difference-in-differences.
- Furthermore, since we have the `R` code for the simulated data, we can compare to the true ATT risk difference, log relative risk, and log odds ratio, drawing on the previous calculations.
- We can therefore determine whether our estimated ATT risk difference, log relative risk, and log odds ratio using the linear SNMM, loglinear SNMM, or logistic SNMM have confidence intervals that contain the truth.
- Due to the way we simulated the data, we do not know the true CACE.
- If we assume no defiers, we know that the true CACE equals the true ATT risk difference.
- However, we cannot tell from the simulation whether there are defiers or not.

Examples

- We can determine from the `R` code for our simulation in `doublewhatifsim.r` that the linear SNMM holds but that the loglinear SNMM and the logistic SNMM do not hold.
- We can derive that

$$\begin{gathered}
E(Y(0)|A=1,T=0)=0.655\\
E(Y(0)|A=1,T=1)=0.548\\
E(Y|A=1,T=0)=0.295\\
E(Y|A=1,T=1)=0.188,
\end{gathered}$$

- and we can use these numbers to compute

$$\begin{gathered}
E(Y-Y(0)|A=1,T=0)=-0.36\\
E(Y-Y(0)|A=1,T=1)=-0.36,
\end{gathered}$$

which means that the linear SNMM holds.

Examples

- However, as

$$\log(E(Y|A=1,T=0)) - \log(E(Y(0)|A=1,T=0)) = -0.798$$
$$\log(E(Y|A=1,T=1)) - \log(E(Y(0)|A=1,T=1)) = -1.069,$$

  we observe that the loglinear SNMM does not hold.

- Similarly, as

$$\text{logit}(E(Y|A=1,T=0)) - \text{logit}(E(Y(0)|A=1,T=0)) = -1.512$$
$$\text{logit}(E(Y|A=1,T=1)) - \text{logit}(E(Y(0)|A=1,T=1)) = -1.655,$$

  the logistic SNMM does not hold.

Examples

- We compute the estimators for the linear, loglinear, and logistic SNMMs using the R functions `ividentity.r`, `ivlog.r`, `ivlogit.r`, and then we compute the jackknife confidence intervals using `jackiv.r`.

Examples

```
> ividentity.r
function(data)
{
dat<-data
Deta<-predict(glm(VL1~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)
}
```

Examples

```
> ivlog.r
function(data)
{
dat<-data
niter=10
A<-dat$A
Z<-dat$T
Deta<-predict(glm(VL1~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)
}
```

Examples

```
> ivlogit.r
function(data)
{
dat<-data
niter<-10
A<-dat$A
Z<-dat$T
Deta<-predict(glm(VL1~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)
#return(ivreg(formula=Ystar~Astar,instruments=~Z))
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)
}
```

Examples

```
> jackiv.r
function ()
{
out<-jackknife(data=doublewhatifdat,statistic=ividentity.r)
est<-out$stats$Observed
lci<-est-1.96*out$stats$SE
uci<-est+1.96*out$stats$SE
list(est=est,lci=lci,uci=uci)
}
```

Examples

- The results are presented in Table 1.
- We see that all of the 95% confidence intervals for the SNMM methods contain the truth, despite the loglinear and logistic SNMMs not holding.
- The estimates from the loglinear and logistic SNMMs are indeed further from the truth than are those from the linear SNMM, as we would expect.
- However, it is reassuring that despite the model misspecification of the loglinear and logistic SNMM, the confidence intervals contain the true value.
- In practice, we have no way of knowing which, if any, of the SNMMs are correctly specified.
- We recommend trying all three, in hopes that the results qualitatively agree.

Examples

Table 1: Instrumental Variables Analysis of the Double What-If? Study

Measure	Truth	Method	Estimate (95% CI)
$E(Y(1) - Y(0) \mid C = 1)$	?	$\widehat{CACE}$	-0.409 (-0.564, -0.254)
$E(Y - Y(0) \mid A = 1)$	-0.360	Linear SNMM	-0.409 (-0.564, -0.254)
$E(Y - Y(0) \mid A = 1)$	-0.360	Loglinear SNMM	-0.483 (-0.746, -0.219)
$E(Y - Y(0) \mid A = 1)$	-0.360	Logistic SNMM	-0.413 (-0.560, -0.265)
$\log(E(Y \mid A = 1)) - \log(E(Y(0) \mid A = 1))$	-1.033	Linear SNMM	-1.020 (-1.340, -0.700)
$\log(E(Y \mid A = 1)) - \log(E(Y(0) \mid A = 1))$	-1.033	Loglinear SNMM	-1.129 (-1.564, -0.693)
$\log(E(Y \mid A = 1)) - \log(E(Y(0) \mid A = 1))$	-1.033	Logistic SNMM	-1.025 (-1.339, -0.712)
$\text{logit}(E(Y \mid A = 1)) - \text{logit}(E(Y(0) \mid A = 1))$	-1.630	Linear SNMM	-1.780 (-2.478, -1.081)
$\text{logit}(E(Y \mid A = 1)) - \text{logit}(E(Y(0) \mid A = 1))$	-1.630	Loglinear SNMM	-2.116 (-3.409, -0.822)
$\text{logit}(E(Y \mid A = 1)) - \text{logit}(E(Y(0) \mid A = 1))$	-1.630	Logistic SNMM	-1.794 (-2.464, -1.124)