Causal inference in the social, health and business sciences: A very brief practical primer in R and Stan

Theiss Bendixen

2023-05-24

Contents

1 Introduction **2**

2 The fundamental problem of causal inference: Observed, potential and missing outcomes **2**
2.1 Some notation . . . 2
2.2 DAGs and experimental randomization . . . 3

3 Blocking backdoors and the elementary ingredients of DAGs **4**
3.1 The "Fork" . . . 4
3.2 The "Pipe" . . . 5
3.3 The "Collider" . . . 5
3.4 Take-homes . . . 6
3.4.1 "The Table 2 fallacy" . . . 6
3.4.2 100,000 regressions do not make for a causal estimate . . . 6
3.4.3 Average people vs. people on average: Marginal and conditional effects . . . 7

4 G-methods and marginal effects **8**
4.1 Inverse Probability Weighting . . . 9
4.2 Standardization . . . 11
4.3 Longitudinal analysis and time-varying treatments . . . 12
4.4 It's assumptions all the way down . . . 18

5 Missing data **19**
5.1 External validity . . . 20
5.1.1 Generalizability, Transportability, Poststratification . . . 21
5.2 Internal validity . . . 28
5.2.1 Missing (completely) at random . . . 28
5.2.2 Missing not at random . . . 28

6 To do: **29**

7 Appendix: Velux IPTW **30**

References **33**

1 Introduction

Causal inference is about what works, when, for whom and under what circumstances. When an event occurs – a medical treatment, social intervention, marketing campaign, etc. – we want to know what happened as a result, if anything. But – and here's the crucial insight – we also need to know what would have happened *in the absence of* the event, all else being equal. That is, the causal effect of an event is the difference between what actually happened and the counter-factual.

This working paper is a very brief practical primer on causal inference in the social, health, and business sciences. We'll use minimal formal notation and instead emphasize practical computational implementation using the `R` scripting language. It assumes a basic understanding of Bayesian regression and generalized linear modeling in `R`. Most importantly, however, it assumes an interest in the question of how we can ever know whether something caused something else and to what extent.

Without further ado, let's dive in.

2 The fundamental problem of causal inference: Observed, potential and missing outcomes

We said just above that the causal effect of an event is the difference between what actually happened and the counter-factual. But here immediately the Fundamental Problem of causal inference comes crashing down at our feet: We can never observe the counter-factual!

We can never re-play history or anything of the sort. For instance, we can never give a medical treatment to a patient and also *not* give the treatment to said patient. The counter-factual outcome is often missing. Causal inference is, then, fundamentally a missing data problem.

So, the best we can do is to *infer* the counter-factual by estimation, approximation or informed (hopefully) guesswork and contrast that with the observed state. More precisely, the best we can do is to aim at an **average treatment effect** or variants thereof. An average treatment effect of, say, a medical treatment is the difference in outcomes under receiving the treatment vs. not receiving the treatment not within an individual, like the individual treatment effect, but instead *between* individuals. Further, these individuals have to be similar enough in relevant characteristics such that, whatever the difference in outcome between the treatment and no treatment groups is, in fact, explained by the treatment alone, and not by some other events or characteristics.

This, in a nutshell, is the fundamental problem of causal inference. It's simple. Deceptively so, because the real world is a mess. So we have assumptions to make. But in return we get something quite extraordinary: A rigorous, tried-and-tested workflow for thinking about cause-and-effect-relationships.

The fundamental problem of causal inference, then, is also a *promise*: A promise that, if we're careful, honest, and transparent, we might actually get causal answers to the questions that matter most in life. Or, alternatively, we'll be told that there exist no valid answers to the questions we posed. Either is a very valuable yield.

2.1 Some notation

To make the discussion more succinct, we'll use some simple formal notation and terminology. It's useful to learn these for another reason, too: perhaps you'll want to navigate the causal inference literature on your own, when you're done here.

Throughout, Y is our outcome on which our cause has or hasn't an effect, and X is the cause, our predictor of interest – say, receiving a particular medical treatment $X = 1$ or not $X = 0$. That is, in its simplest form, an average treatment effect is simply the difference in expected outcomes between receiving the treatment $Y^{X=1}$ or not $Y^{X=0}$. This quantity is often written as $\mathrm{E}(Y^{X=1}) - \mathrm{E}(Y^{X=0})$, where $\mathrm{E}(.)$ is the expectation operator, i.e. it computes the expected value.

2.2 DAGs and experimental randomization

One useful tool for thinking through a causal modeling problem is **DAG**s, *directed acyclic graphs*. A DAG graphs the assumed[1] causal relationships among variables and can thereby guide subsequent analytic strategies for recovering the causal effects of interest (or can reveal that no such effects can be recovered under any circumstances, given the DAG). *Directed* means that the relationships, denoted by arrows, are causal. *Acyclic* means that variables can't cause themselves dynamically. And *graph* means... well, that should be obvious.

Figure 1 is one example of a DAG. It shows a simple but extremely common scenario in which both our cause X and outcome Y are influenced by a third variable Z. In other words, Z is a **confounder**. We'll discuss confounding in more depth in sections below. For now, suffice it to say that our estimate of the cause of X on Y will be biased if we do not take into account Z. How do we know that Z confounds the relationship between X and Y and thereby induces bias? Simple, we can read it off the DAG.

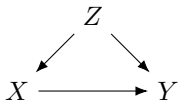


Figure 1: **Directed acyclic graph (DAG) of confounding**.

To see this, note that Z has a causal effect on both X and Y. This is technically called a **backdoor path**. A backdoor path is any set of arrows that both points into the outcome Y and predictor of interest X. In a basic sense, causal inference is about identifying such backdoor paths and then "blocking'' them, using study design or statistical adjustment. This simple insight makes analyzing DAGs a sort of party game or puzzle. It can actually *be fun* to draw out more or less complicated DAGs and then identify how and whether a causal effect can be recovered under this particular graph, through backdoor paths and the likes. There's more to graph analysis than looking for and blocking backdoor paths, as we'll see below. But it's a good start.

So how do we block backdoor paths? I said just above that we can, among other approaches, use particular study designs. Randomization is the best known example of such a study design. What randomization does, from the perspective of DAGs, is that it deletes any arrows that go into X, meaning that there can longer be any backdoor paths, by definition. Figure 2 illustrates this. It's similar to figure 1, except that there's no arrow from Z to X. Under this model, an estimate of the effect of X on Y is no longer biased and we don't need to measure Z[2].

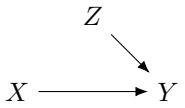


Figure 2: **DAG of randomized** X.

Randomization allows us to delete arrows going into X from all other variables because randomization is, well, random. So, whether an individual receives a randomized medical treatment will not be dependent on anything else than the randomization mechanism. In the next section, we'll show in code and simple simulated data how this works.

In sum, randomization is an efficient strategy for causal estimation, because it makes X independent of all other variables in the graph. This is the reason why randomized controlled studies are often referred to as the ''gold-standard" for causal inference. But, unfortunately, our work does not end here. Randomization is often imperfect (e.g., participants might not perfectly adhere to the randomized treatment) or not feasible for practical or ethical reasons (e.g., could we assign people to smoke in a randomized fashion such that we could estimate the causal effect of smoking on, say, cancer?).

[1] Based on prior studies, theory, commonsense, or other sources of inference.

[2] Although, on a technical aside, knowing Z potentially increases the precision of our estimate of X.

This does not mean that all hope is lost. Causal inference in observational or pseudo-randomized contexts is possible. But it does force us to make some assumptions and add some steps to our workflow. For instance, through relatively simple statistical techniques, we can block backdoor paths that couldn't be deleted by randomization. Next section introduces the promises and pitfalls of this approach, using regression modeling.

3 Blocking backdoors and the elementary ingredients of DAGs

So, how do we identify and block backdoor paths? It's actually pretty simple, at least in principle. Figure 3 shows three distinct patterns that we need to be able to recognize in a DAG. We could think of these patterns as the elementary *ingredients* of causal graphs, since any given DAG, no matter its complexity, is constructed by combining one or more of these.

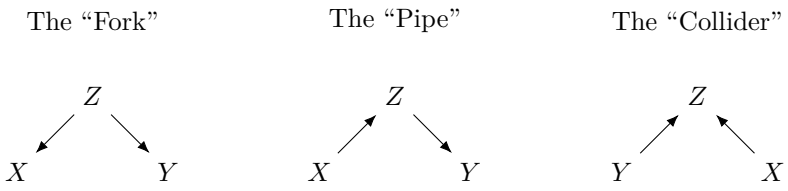


Figure 3: **Elementary ingredients of causal graphs**.

3.1 The "Fork"

To see this, consider first the "fork" on the left. You'll see that it's similar to Figure 1, except that we've left out the arrow $X \rightarrow Y$. A fork implies that X and Y are only associated because of a common causal influence of Z. In other words, the relationshiop between X and Y is *confounded*.

Given this DAG, then, we should not expect a *direct* causal influence of X on Y; they are only associated through the confounding backdoor path $X \leftarrow Z \rightarrow Y$. By statistically adjusting for Z, we'd block the backdoor path, making X and Y independent from each other. Adjusting for Z then achieves the same thing as randomization, in that it allows us to delete the arrow going into X from Z, as we saw above.

Let's verify this in a simple simulation.

```
set.seed(123)

n <- 1e4

bZ <- 0.5

Z <- rnorm(n, 0, 1)
X <- Z*bZ + rnorm(n, 0,1)
Y <- Z*bZ + rnorm(n, 0,1)

forkdat <- data.frame(Y=Y, X=X, Z=Z)

# bX is non-zero; confounded
glm(Y ~ X, data = forkdat)

# bX is zero; de-confounded
glm(Y ~ X + Z, data = forkdat)
```

3.2 The "Pipe"

Consider next the "pipe". As with the fork, the pipe implies that X and Y are only associated through Z. This time around, though, Z is not a confounder in a traditional sense, since it's causally influenced by X. In other words, Z is a *mediator* on the path between X and Y. This somewhat complicates our adjustment strategy.

If we're interested in estimating the **total effect** of X on Y, we would not adjust for Z, since adjusting for Z would block the part of the causal path from X to Y that runs through Z. However, say we hypothesized a **direct effect** of X on Y in addition to the mediating path through Z. If we were interested in estimating only that direct effect, we would have to adjust for Z, since Z would otherwise confound that path.

Again, we can verify this logic in a simple simulation.

```
set.seed(123)

n <- 1e4

bX <- 0.5
bZ <- 0.5

X <- rnorm(n, 0,1)
Z <- X*bX + rnorm(n, 0, 1)
Y <- Z*bZ + rnorm(n, 0,1)

pipedat <- data.frame(Y=Y, X=X, Z=Z)

# bX is 0.5*0.5; total effect
glm(Y ~ X, data = pipedat)

# bX is zero; direct effect
glm(Y ~ X + Z, data = pipedat)
```

3.3 The "Collider"

With both the fork and the pipe, then, X and Y are associated, unless we adjust for Z. With our final ingredient, the "collider", it's different. Actually, it's the opposite. In a collider, X and Y are associated *only if* we adjust for Z. That is, adjusting for Z opens rather than closes a backdoor path. This is sometimes referred to as *collider bias*.

If you're like the rest of us, this is not very intuitive on a first thought, so let's check it in a simple simulation.

```
set.seed(123)

n <- 1e4

bX <- 0.5
bY <- 0.5

X <- rnorm(n, 0,1)
Y <- rnorm(n, 0,1)
Z <- X*bX + Y*bY + rnorm(n, 0, 1)

colldat <- data.frame(Y=Y, X=X, Z=Z)

# bX is 0; X and Y non-associated
glm(Y ~ X, data = colldat)
```

```r
# bX is non-zero; X and Y associated
glm(Y ~ X + Z, data = colldat)
```

3.4 Take-homes

There are many pedagogical resources that discuss in more depth the promises and pitfall of statistical adjustment for causal inference (e.g., Achen 2005; Cinelli, Forney, and Pearl 2020; Lübke et al. 2020; Rohrer 2018; Westreich and Greenland 2013; Wysocki, Lawson, and Rhemtulla 2022). But, for now, I focus on three critical take-home points:

3.4.1 "The Table 2 fallacy"

First, we have shown how multiple regression can be an effective alternative to randomization for blocking backdoor paths. This is a wonderful promise, since it allows us to study causal relationships that would not be feasible or ethical to study in a RCT context. It also allows us to decompose effects of interest, which is particularly relevant in mediation scenarios [e.g., total vs direct effect; for a technical but accessible introduction, see Wang and Arah (2015)], as well estimating an unbiased causal effect even in the presence of unmeasured confounding, using a formula known as the *front-door criterion* (**to be discussed**).

However, when using statistical adjustment, we must tread with care. Just as statistical adjustment can block backdoor paths and thereby de-confound a true causal effect, statistical adjustment can induce bias, if we adjust for the wrong variables. We saw this in the pipe and collider examples: Adjusting for a mediator *blocks* a true causal effect, while adjusting for a collider *opens up* a non-causal path. A similar scenario, which we haven't discussed as of yet, is when a so-called *post-treatment* variable is adjusted for. Post-treatment bias can likewise mask a true causal effect (**to be discussed**).

All of this entails that we cannot in general treat all regression coefficients as "causes" of the outcome. A statistical adjustment set (e.g., whether or not to include Z in a regression) is built for a particular purpose and a particular set of causal assumptions (e.g., estimating the effect of X on Y, adjusting for assumed confounder Z). This in turn means that we cannot expect that the variables used for statistical adjustment also represent a particular causal effect. For all we know, parts of the adjustment set might be a collider or a mediator on some other path in the causal graph, rendering a causal interpretation invalid.

This fundamental insight stands in contrast to common practice in many empirical sciences, where a table (often Table 2 of an empirical report) with all estimated regression coefficients is often presented – implicitly or explicitly – as direct causal effects. This is a very unfortunate habit as it potentially misleads the reader to interpret coefficients of the control variables causally. Accordingly, to raise awareness about the potential pitfalls of this malpractice, it has been given a name, the "Table 2 Fallacy" (Westreich and Greenland 2013).

3.4.2 100,000 regressions do not make for a causal estimate

Another critical take-home is that causes are not in the data. For instance, it's not possible from the data alone to distinguish between a pipe and a fork, because they have similar statistical implications: In both cases, X and Y are associated, unless we condition on Z. But, for any given theory, whether Z is thought of as a mediator or a confounder is often very different.

Therefore it follows that a DAG cannot be built on the basis of any number of regressions alone. Regressions are prediction engines, so regression coefficients only have a valid causal interpretation to the extent that identification assumptions hold. Instead, regressions can be used to *test* the implications of a particular DAG or set of DAGs (for discussion of this point in the wild, see Bendixen 2023; Purzycki, Bendixen, and Lightner 2022). All this means that, at least in the context of an observational study where randomization was not feasible, variable selection – which variables to include in your adjustment set – cannot be automated; it strictly requires prior knowledge and explicit causal assumptions (Hernan and Robins 2020; Westreich 2019).

3.4.3 Average people vs. people on average: Marginal and conditional effects

The final take-home pertains to the difference between **marginal** and **conditional** estimates. Here's what I mean.

In all but the simple cases, regression analysis yields conditional effect estimates. That is, the regression coefficient of interest is not the average effect for the whole sample but rather for a subset, conditional on the other variables in the model. To see this, consider the following simulation.

Say we randomize some marketing campaign X but suspect that age A might be modifying the effect of this intervention. The DAG would then be similar to figure 2; since we randomized X, A cannot be a confounder, but we're interested in modeling A anyway, because it could interact with the intervention in interesting ways (e.g., younger people might be more or less willing to be persuaded by the "once-in-a-lifetime offer!!" that we're launching). For the sake of interpretation, we're modeling age in a centered and standardized manner, with a mean of 0 and a standard deviation of 1.

```
# inverse logit function
logistic <- function(x) exp(x)/(1+exp(x))

set.seed(123)

n <- 1e3

bX <- 1
bA <- 1
bXA <- 0.5

A <- rnorm(n, 0, 1)
X <- rbinom(n, 1, prob = 0.5)
Y <- rbinom(n, 1, prob = logistic(-0.5 + X*bX + A*bA + X*A*bXA))

model <- glm(Y ~ X * A,
             family = "binomial",
             data = data.frame(Y=Y, X=X, A=A))
```

Now, check the regression coefficient for X in the `model` output. It's around 1. But consider now what this means. Very briefly put, in multiple regression, a coefficient represent the association of a given predictor to the outcome, when all other predictors are held at 0. The regression coefficient for X represents therefore the effect of the intervention, *when age is held at 0*, that is at the average age. This is the **conditional** estimate: the intervention effect for an average-aged individual.

However, often we wouldn't be interested in this quantity, the effect of an intervention at a very particular level of a covariate (a particular year of age, in this case). Perhaps we're interested in the effect of the intervention across a range of covariate values, but the regression coefficient for X is still not directly getting us there. For instance, we could get predictions for the intervention's effect for up to ± 2 standard deviations of the mean age like so:

```
# X=0
ndX0 <- data.frame(X=rep(0,5),
                   A=c(-2,-1,0,1,2))

estX0 <- predict(model, newdata = ndX0)

# X=1
ndX1 <- data.frame(X=rep(1,5),
                   A=c(-2,-1,0,1,2))
```

```r
estX1 <- predict(model, newdata = ndX1)

# Differences
data.frame(A=c(-2,-1,0,1,2), effect=estX1 - estX0) |>
  mutate(effect = round(effect,1)) |>
  head()
```

```
##    A effect
## 1 -2    0.0
## 2 -1    0.4
## 3  0    0.9
## 4  1    1.3
## 5  2    1.7
```

The resulting table is the predicted intervention `effect` across different age groups `A`, in log-odds. Note that when $A = 0$, that's our regression coefficient for X.

This is often a good start. But imagine we want to roll out the marketing campaign at scale in a population. In that scenario, as is often the case in applied cases, we're interested in the effect of the intervention in the sample *as a whole*, averaging over the distribution of age. This is the **marginal** estimate: the intervention effect for individuals on average.

All of this is to say: there is an important distinction between making inferences for the *average* person (conditional) vs. people *on average* (marginal).

So there we have it. Our estimand is the marginal effect, but in moderately complex cases, where we're adjusting for covariates with potentially non-linear relationships, regression only yields conditional estimates. At the same time, adjustment for covariates is often critical for de-confounding, particularly in observational settings. So, what gives? Enter g-methods.

4 G-methods and marginal effects

As we saw above, multiple regression is a powerful tool for blocking backdoor paths. This allows for causal interpretation of the focal parameter(s) (but, importantly, *not* the control variables). However, regression has drawbacks that makes it insufficient for many common causal estimands, in particular when there are many covariates with complex relationships. Often, we're interested in marginal, not conditional, treatment effects – that is, treatment effects in a population as a whole, not just subsets thereof. In more complicated data structures, such as longitudinal studies with dynamical relationships between treatment and outcome, regression alone will also fall short. In all these cases, we need a few additional tools in the trunk.

One such family of tools is known as **g-methods** – *g* for *general* or *generalized* (Robins 1986). Here, we'll focus on one g-method in particular, variously referred to as **g-computation** (e.g. Snowden, Rose, and Mortimer 2011; Ahern, Hubbard, and Galea 2009), **g-formula** and **standardization** (e.g. Vansteelandt and Keiding 2011; Hernan and Robins 2020, ch. 13), while we'll also very briefly introduce another method, known as **inverse probability of treatment weighting** (IPTW). Standardization and IPTW rely on the same identification assumptions and will yield similar if not identical results (indeed, in many simpler cases, they are mathematically identical), but they arrive at their results at quite different analytic routes.

We focus primarily on standardization for a few reasons. First, since we want to perform our more real-world data analysis in a Bayesian framework, standardization is an obvious choice since the IPT weights are not obviously compatible with Bayes theorem (for some discussion see Robins, Hernán, and Wasserman 2015; Saarela et al. 2015).

Second, compared to IPTW, standardization and its practical implementation seem to be often overlooked in popular text books and primers on causal inference and econometrics in the social and health sciences (e.g., Morgan and Winship 2015; Pearl, Glymour, and Jewell 2016; Westreich 2019; Angrist and Pischke 2009; McElreath 2020), although the latter stresses contrasts. It is discussed in (Hernan and Robins 2020, ch. 13)

but there's no explicit, reproducible practical/programming application, though there's a g-methods package. IPTW also does not handle interaction terms between a predictor and the outcome.

However, given its popularity, it's useful to at least be aware of the nuts and bolts of IPTW, too. So let's briefly present the two methods in turn.

4.1 Inverse Probability Weighting

Recall that the main aim of a covariate-adjusted analysis is to obtain conditional exchangeability: When we account for imbalances in covariates between treatment and control group, we say the two groups are exchangeable conditionally on the covariates. This means that the two groups are comparable such that if we detect a difference between the groups after the intervention, we can interpret that difference as a causal effect of the intervention. Another way to think of obtaining exchangeability is as de-confounding.

The IPTW method obtains conditional exchangeability by, in effect, creating a "pseudo-population" in which covariates are balanced between treatment and control. As the name hints at, this pseudo-population is created by estimating the probability of receiving the treatment conditional on covariates. This probability is then inverted and used as weights in a regression predicting the actual outcome of interest. In code, the procedure looks like this. First, we simulate a simple, confounded data structure: a binary treatment X with a coefficient of βx, binary confounder Z and a continuous outcome Y.

```
set.seed(123)

n <- 2e3

bX <- 0.5

Z <- rbinom(n, 1, 0.3)
X <- rbinom(n, 1, 0.4 + Z*0.5)
Y <- rnorm(n, 10 + X*bX + Z*0.5)

gdat <- data.frame(Y=Y, X=X, Z=Z)
```

Next, we calculate IPT weights by, first, fitting a model that regresses X on Z. Then, for each row in the dataset, we get predictions from this model for the probability of receiving treatment. Then we compute the IPT weights and includes those weights in a regression that predicts the outcome by X. This latter step, the outcome regression, is known as a **marginal structural model** (MSM). It's a *marginal* model, because the coefficient of X is an average (or marginal, since the estimated weights *marginalizes over* the distribution of the covariate) treatment effect in the population; and *structural*, because X has a valid causal interpretation (*structural* is another word for *causal*).

```
# inverse logit function
logistic <- function(x) exp(x)/(1+exp(x))

# receiving treatment conditional on Z
treat.mod <- glm(X ~ Z, data = gdat, family = "binomial")

# probability of treatment
gdat$pd <- predict(treat.mod) |> logistic()

# compute inverse weights
gdat$w <- with(gdat, ifelse(X==1, 1/pd, 1/(1-pd)))

# MSM of outcome
glm(Y ~ X, data = gdat, weights = w)
```

The MSM recovers the simulated coefficient for X ($\beta x = 0.5$), even though the model does not adjust for Z

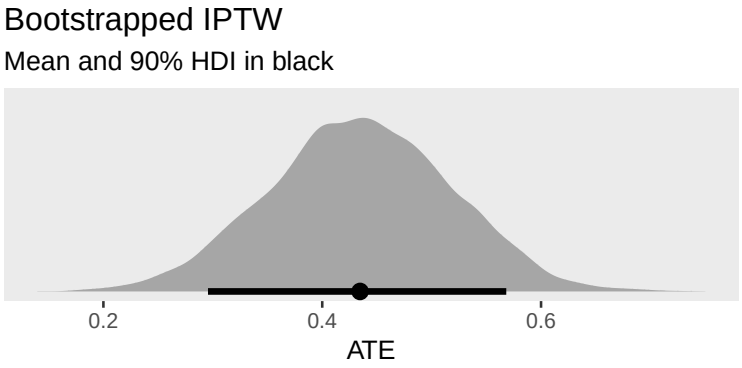
in a traditional sense. This is because we adjusted for Z using weights. Note that this simple example could also, of course, be obtained using simple multiple regression that adjusted for X and Z. This, however, will not be so, when we tackle more complex longitudinal data in the next section.

There are at least two ways to improve on our simple IPTW workflow above. First, above we calculated unstabilized weights, but in many cases (and always with continuous exposures), so-called **stabilized weights** are preferred (and in all cases, it's good to know about both). In brief, the stabilized weights are different from the unstabilized in that the numerator is the unconditional inverse probability of treatment and they guard against extreme weights and, in turn, variance-inflation (for more detail, see Chesnaye et al. 2022). The stabilized weights can be calculated with an intercept-only regression on treatment.

```
# intercept-only model on treatment
treat.mod.sw <- glm(X ~ 1, data = gdat, family = "binomial")

# probability of treatment
gdat$pn <- predict(treat.mod.sw) |> logistic()

# compute stabilized weights
gdat$sw <- with(gdat, ifelse(X==1, pn/pd, pn/(1-pd)))

# MSM of outcome with stabilized weight
glm(Y ~ X, data = gdat, weights = sw)
```

The other thing we can do to improve the workflow is to obtain a measure of uncertainty around our marginal estimate. However, since [**placeholder**], we can't rely on the standard error from our MSM. Instead, we must resort to **bootstrapping**, which involves running the same algorithmic routine R times. This will result in a distribution of estimates with length R. Here's one way to implement bootstrapping with the stabilized weights.

```
# IPTW function
iptw_fun <- function(formula, data, indices) {
  d <- data[indices,]
  fit <- glm(formula, family="gaussian", weights=sw, data=d)
  return(coef(fit))
}

iptw.result <- boot(
                # set data
                data=gdat,
                # set function
                statistic=iptw_fun, # set function
                # set R
                R=1e4,
                # specify formula
                formula=Y ~ X)

# bootstrapped point estimate for bX
iptw.point <- iptw.result$t0[2]

# bootstrapped interval for bX
iptw.interval <- boot.ci(iptw.result,
                         type = "norm",
                         index = 2)$normal
```



That's it for IPTW for now. They are popular and reasonably straightforward to implement – although practical issues remain (e.g., Austin and Stuart 2015). For instance, what do to with extreme weights? Some suggest that extreme weights need to be truncated or trimmed at some threshold. This might work, although to me it seems somewhat unprincipled (e.g., what is an extreme weight anyway?).

In any case, as mentioned, since there's no very clear way of obtaining uncertainty in the weights and then propagate that uncertainty to the MSM in a formal Bayesian framework, we leave IPTW here.

4.2 Standardization

Standardization is our main g-method here. It's arguably even more straightforward to implement than IPTW, as it requires us to only fit a single regression of the outcome and then do some post-fitting simulation.

The conceptual steps in standardization are as follows:

- Fit a theoretically informed model of the outcome including the treatment, covariates and possibly non-linear relationships and functional forms (e.g., interaction terms, quadratic terms, etc.)
- Duplicate the original data, set $X = 1$ for all individuals, and then obtain predictions of the outcome using the fitted model holding covariate(s) Z as observed z, $E[Y^{Z=z,X=1}]$.
- Duplicate the original data, set $X = 0$ for all individuals, and then obtain predictions of the outcome using the fitted model holding covariate(s) Z as observed z, $E[Y^{Z=z,X=0}]$.
- For each draw of the posterior distribution, compute contrast corresponding to the estimand of interest, $E[Y^{X=1}] - E[Y^{X=0}]$.

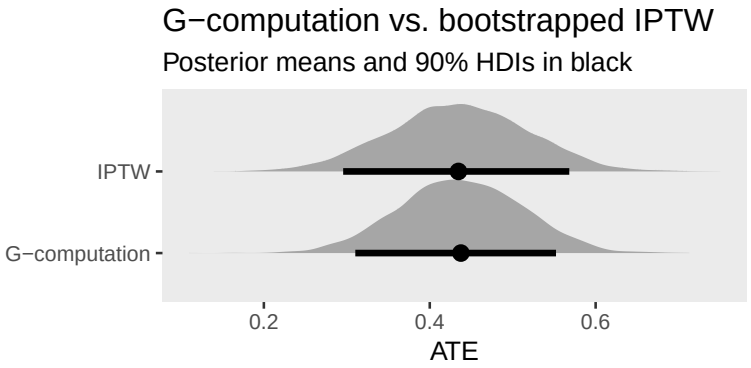
In code, using Bayesian estimation with default priors via `Stan` and `brms`, the steps are these.

```
# outcome model
stand.mod <- brm(Y ~ X + Z,
                 data = gdat,
                 family = "gaussian",
                 cores = 4)
```

```
# set X=1, Z=z
X1 <- transform(gdat, X=1)

# set X=0, Z=z
X0 <- transform(gdat, X=0)

# E[Y{Z=z, X=1}]
EX1 <- add_epred_draws(stand.mod, newdata = X1)

# E[Y{Z=z, X=0}]
EX0 <- add_epred_draws(stand.mod, newdata = X0)
```

```r
# E[Y{X=1} - Y{X=0}]
ate <- data.frame(X1 = EX1$.epred,
                  X0 = EX0$.epred,
                  draw = EX0$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(X1 - X0))
```

The distribution of contrasts calculated in that final step are marginal estimates in that the two expectations $E[Y^{Z=z,X=1}]$ and $E[Y^{Z=z,X=0}]$ marginalize or average over the covariate(s) Z. We summarize the contrast by its posterior mean and 90% HPDI and plot the distribution against the bootstrapped IPTW, using the `tidybayes` package. Results are almost identical, although g-computation looks slightly more precise in this particular case.



When performing standardization in a frequentist setting, we'd have to resort to bootstrapping in order to obtain valid uncertainty around the contrast (see e.g. Snowden, Rose, and Mortimer 2011), as with IPTW. However, in a Bayesian setting, we get uncertainty in one go, since the predicted values are valid posterior distributions of expectations.

One more thing: There's a fancy `R` package called `marginaleffects` that does all of this under the hood. Here's how:

```r
# marginaleffects solution
library(marginaleffects)

avg_comparisons(stand.mod, variables = "X")
```

It's certainly more compact but, for pedagogical purposes, we won't pursue this solution further here; black-boxing the computational steps might be convenient but it does very little to enhance understanding. It's useful to know about, however, if nothing else for double checking our results.

Okay, now that we have introduced g-methods in a very simple data context, we're ready to tackle more complex data structures, where regression modeling alone falls short.

4.3 Longitudinal analysis and time-varying treatments

For illustrating a longitudinal data context with time-varying treatment, we'll simulate from the DAG in figure 4. Suppose that our treatment X is randomized but administered according to some time-varying covariate Z. We're interested in the *joint effects* of the treatment X_0 and X_1 on our outcome Y (that is, the effects of each treatment on the outcome not through the subsequent treatment), denoted by the dashed edges.

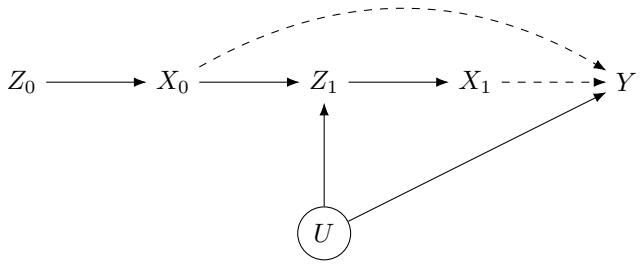


Figure 4: DAG of longitudinal treatment-confounder feedback.

However, some unobserved variable U affects both Z_1 and Y. This confounds the relationship between the treatment and the outcome via the backdoor path $Y \leftarrow U \rightarrow Z_1 \rightarrow X_1$. This in turn means that in order to estimate the hypothesized causal path $X_1 \rightarrow Y$, we'd want to adjust for Z_1.

But here's the issue: Z_1 is also a collider on the path $Y \leftarrow U \rightarrow Z_1 \leftarrow X_0$, implying that if we condition on Z_1, we open a non-causal path between X_0 and Y. All in all, under this DAG, we're not able to estimate the joint effects of X_0 and X_1 on Y using a single regression model, since the two treatment indicators imply different adjustment sets. G-methods to the rescue.

First, we simulate some data, under the model presented in figure 4. To keep things simple, we assume that the treatment, in fact, does not have any direct impact on Y (i.e., we could delete the dashed edges), such that the true coefficients of X_0 and X_1 should be (roughly) zero.

```
set.seed(1)

n <- 2e4

U <- rnorm(n, 0, 1)
Z_0 <- rbinom(n, 1, 1/(1+exp(0.5)))
X_0 <- rbinom(n, 1, 1/(1+exp(0.5 + Z_0*0.5)))
Z_1 <- rbinom(n, 1, 1/(1+exp(0.5 + X_0*0.5+U*0.5)))
X_1 <- rbinom(n, 1, 1/(1+exp(0.5 + Z_1*0.5)))

Y <- rnorm(n, 10 + U*2)

lgdat <- data.frame(Y=Y, X_0=X_0, X_1=X_1, Z_0=Z_0, Z_1=Z_1, U=U)
```

Then, we verify the havoc that Z_1 can wreak, if we haphazardly estimated the effect of X_0 on while adjusting for Z_1. The coefficient is solidly non-zero.

```
mx0 <- glm(Y ~ X_0 + Z_1, data = lgdat)
summary(mx0)
confint(mx0)
```

And similarly, we check that *not* adjusting for Z_1 can bias our estimate of X_1.

```
mx1 <- glm(Y ~ X_1, data = lgdat)
summary(mx1)
confint(mx1)
```

Then, we apply standardization to the data. The logic is the same as discussed above, only now we have two time points.

So, we first fit a model for the treatment at time point 0 and then compute the contrast in expectations between receiving and not receiving the treatment, $E[Y^{X_0=1,Z_0=z_0}] - E[Y^{X_0=0,Z_0=z_0}]$. Next, we fit a model

for receiving treatment at time point 1 and similarly compute the contrast in expectations between receiving and not receiving the treatment, $E[Y^{X_1=1,X_0=x_0,Z_1=z_1}] - E[Y^{X_1=0,X_0=x_0,Z_1=z_1}]$.

```
# Outcome model for X_0
yX0model <- brm(Y ~ X_0 + Z_0,
                data = lgdat,
                cores = 4)
```

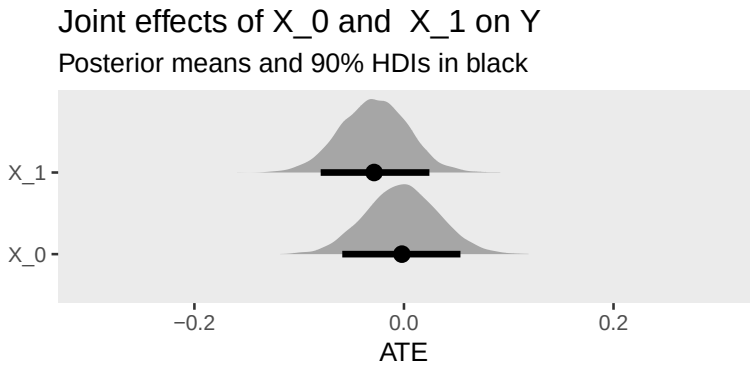
```
# E[Y{X_0=1, Z_0=z_0}]
EX01 <- add_epred_draws(yX0model, newdata = transform(lgdat, X_0=1))

# E[Y{X_0=0, Z_0=z_0}]
EX00 <- add_epred_draws(yX0model, newdata = transform(lgdat, X_0=0))

# E[Y{X=1} - Y{X=0}]
ateX_0 <- data.frame(X_01 = EX01$.epred,
                     X_00 = EX00$.epred,
                     draw = EX00$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(X_01 - X_00))
```

```
# Outcome model for X_1
yX1model <- brm(Y ~ X_0 + X_1 + Z_1,
                data = lgdat, cores = 4)
```

```
# E[Y{X_0=1, Z_0=z_0}]
EX11 <- add_epred_draws(yX1model, newdata = transform(lgdat, X_1=1))

# E[Y{X_0=0, Z_0=z_0}]
EX10 <- add_epred_draws(yX1model, newdata = transform(lgdat, X_1=0))

# E[Y{X=1} - Y{X=0}]
ateX_1 <- data.frame(X_11 = EX11$.epred,
                     X_10 = EX10$.epred,
                     draw = EX10$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(X_11 - X_10))
```

By breaking the procedure down into separate models – one for each time point and treatment – we by-pass the problem that a single regression runs into, namely that Z_1 is both a collider and a confounder. The resulting contrasts, stored in `ateX_0` and `ateX_1` and plotted below, are valid causal estimates of the joint effects of each treatment on the outcome (to the extent that the identification assumptions hold (we discuss these assumptions in the next section).

Recall that we simulated the joint effects as non-existing, so the contrasts should be around zero. They are not *exactly* centered on 0 due to sampling variability; we're only showing a single run of simulation and analysis here.



Now for the finale. The DAG in Figure 4 comes from Hernan and Robins (2020), and it's useful in that it illustrates how regression modeling alone breaks if our estimands require different adjustment sets – for instance, if a variable is both a confounder that we want to adjust for and a collider that we want to keep unadjusted.

But let's take standardization for a spin in a real-world data analysis example. VanderWeele, Jackson, and Li (2016) sets the stage for our example: religion and (mental) health.

The DAG could look something like figure 5.

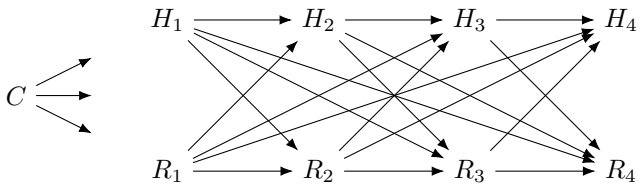


Figure 5: DAG of longitudinal exposure-outcome feedback.

C is a set of baseline covariates comprising standard demographics, including age, gender, income and education; the indefinite arrows indicate confounding on all observed variables). Our exposure R_{1-4} represent level of religiosity ("How important is religion to you?", measured on a 1 ("not at all important") to 4 ("very important") scale.) measured at four time points, and H_{1-4} are subjective general health ("How would you describe your current health?", measured on a 1 ("very bad") to 5 ("very good") scale.), a set of time-varying covariates. Note the complex causal relationships across time between R and H.

To provide some background, this dataset comes from a panel study on a pseudo-representative sample of Danes during the COVID-19 pandemic (for more detail, see Mauritsen, Bendixen, and Christensen 2022). As investigators, we are interested in assessing whether religiosity was a protective factor in terms of subjective general well-being in the context of the stress and uncertainty of a pandemic. If that was so, we'd expect a positive effect of religiosity on subjective health.

Suppose, then, that our estimand of interest is the contrast in subjective well-being at the final measurement time (that is, $Y = H_4$) between being "maximally religious" $Y^{R_1=R_2=R_3=4}$ and being "minimally religious" $Y^{R_1=R_2=R_3=1}$. That is, we're interested in the *joint effects* of being religious on health at all time points up to the final measurement, meaning partitioning the respective effects of religiosity level at each time point on the outcome, *except those running through the subsequent measurements of religiosity.*

Note that we could've made the causal diagram even more involved by hypothesizing that our two time-varying variables also caused future versions of themselves beyond the immediately subsequent time point (e.g., adding paths $R_1 \rightarrow R_3$, $R_1 \rightarrow R_4$, $H_1 \rightarrow H_3$, $H_1 \rightarrow H_4$, etc.).

However, even as it stands, this analysis is not straightforward. Consider for instance that some of the effect of R_1 on the outcome works through H_2 (and H_3) and that if we adjust for H_2 (or H_3), we'll block some of

the effect of R_1. But note too that H_2 (through H_3) confounds the relationship between R_3 and the outcome, such that if we do not adjust for H_2 (or H_3), the estimate for the effect of R_3 will be biased.

All this implies that H_2 is *both* a mediator that we'd want to keep unadjusted *and* a confounder that we'd want to adjust for. A single regression cannot handle this situation. But, as we've already seen, g-methods can.

First, we load the data and packages and set the number of iterations to use in sampling. For now, we use only the complete cases (in all variables but for R_4, which we're not actually using at this moment); below, we'll discuss in more detail what a complete cases analysis implies and ways of handling missing covariate values when using standardisation.

```
veluxData <- read.csv("velux_data.csv")

veluxData <- with(veluxData,
          data.frame(
            id,
            R1 = religion_1,
            R2 = religion_2,
            R3 = religion_3,
            H1 = health_1,
            H2 = health_2,
            H3 = health_3,
            H4 = health_4,
            G = gender,
            A = age.c,
            E = education,
            I = household_income)
)

d <- veluxData[complete.cases(veluxData[, !names(veluxData) %in% c("R4")]),]

iter <- 2000
```

Then, we fit the model for R_1 on the outcome H_4. We're interested in estimating all paths from $R_1 \rightarrow H_4$ except those through subsequent religiosity measurements, while controlling for the demographic confounders. This means that, in addition to the baseline covariates and our main exposure R_1, we only need to include R_2 in the model. Adjusting for R_2 blocks the paths from R_1 to the outcome through subsequent measures of religion. Had we assumed the path $R_1 \rightarrow R_3$, we'd have to include R_3, too, to block the path $R_1 \rightarrow R_3 \rightarrow H_4$.

```
## Outcome model for religion_t1
r1mod <- brm(H4 ~ G + A + E + I + R1 + R2,
             data = d,
             cores = 4,
             iter = iter,
             control = list(adapt_delta = 0.99))

# Min. religious: E[Y{R_1=1, C=c, R_2=r_2}]
ER10 <- add_epred_draws(r1mod, newdata = transform(d, R1=1))

# Max. religious: E[Y{R_1=4, C=c, R_2=r_2}]
ER11 <- add_epred_draws(r1mod, newdata = transform(d, R1=4))

# Contrast: E[Y{R_1=4}] - E[Y{R_1=1}]
ateR1 <- data.frame(ER10 = ER10$.epred,
                    ER11 = ER11$.epred,
```

```
                    draw = ER11$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(ER11 - ER10))

# compare with marginaleffects
median_qi(ateR1$ate)
marginaleffects::avg_comparisons(r1mod,
                                 variables = list(R1 = "minmax"))
```

Next, we fit a model for R_2. In addition to the main exposure and the demographic covariates, we need to adjust for both R_1 (because it's a confounder on the path $R_2 \leftarrow R_1 \rightarrow Z_4$), R_3 (to block the path $R_2 \rightarrow R_3 \rightarrow Z_4$) and Z_1 (because it's a confounder on the path $R_2 \leftarrow Z_1 \rightarrow Z_2 \rightarrow Z_3 \rightarrow Z_4$).

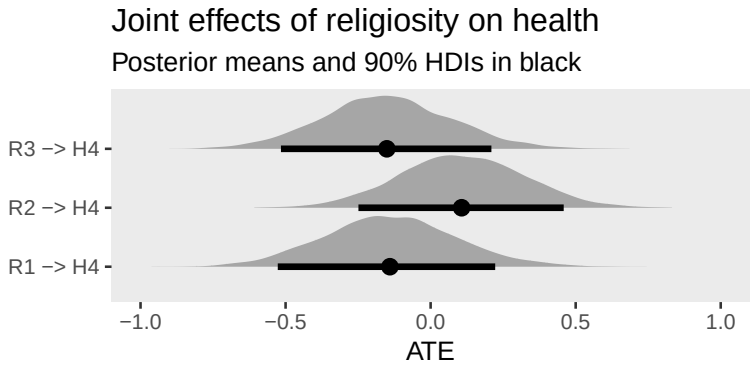
```
## Outcome model for religion_t2
r2mod <- brm(H4 ~ G + A + E + I + R1 + R2 + R3 + H1,
             data = d,
             cores = 4,
             iter = iter,
             control = list(adapt_delta = 0.99))

# Min. religious: E[Y{R_1=1, C=c, R_2=r_2, R_3=r_3, H_1=h_1}]
ER20 <- add_epred_draws(r2mod, newdata = transform(d, R2=1))

# Max. religious: E[Y{R_1=4, C=c, R_2=r_2, R_3=r_3, H_1=h_1}]
ER21 <- add_epred_draws(r2mod, newdata = transform(d, R2=4))

# Contrast: E[Y{R_1=4}] - E[Y{R_1=1}]
ateR2 <- data.frame(ER20 = ER20$.epred,
                    ER21 = ER21$.epred,
                    draw = ER21$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(ER21 - ER20))

# compare with marginaleffects
median_qi(ateR2$ate)
marginaleffects::avg_comparisons(r2mod,
                                 variables = list(R2 = "minmax"))
```

Finally, the model for R_3 adjusts for the demographic covariates as well as R_2 (because it's a confounder on the path $R_3 \leftarrow R_2 \rightarrow Z_4$) and Z_2 (because it's a confounder on the path $R_3 \leftarrow Z_2 \rightarrow Z_4$ through Z_3). Notice that, had we assumed a causal relationship $R_4 \rightarrow Z_4$, R_4 would have to be included, too.

```
# Outcome model for religion_t3
r3mod <- brm(H4 ~ G + A + E + I + R2 + R3 + H2,
             data = d,
             cores = 4,
             iter = iter,
             control = list(adapt_delta = 0.99))

# Min. religious: E[Y{R_1=1, C=c, R_2=r_2, R_3=r_3, H_2=h_2}]
ER30 <- add_epred_draws(r3mod, newdata = transform(d, R3=1))
```

```r
# Max. religious: E[Y{R_1=4, C=c, R_2=r_2, R_3=r_3, H_2=h_2}]
ER31 <- add_epred_draws(r3mod, newdata = transform(d, R3=4))

# Contrast: E[Y{R_3=4}] - E[Y{R_3=1}]
ateR3 <- data.frame(ER30 = ER30$.epred,
                    ER31 = ER31$.epred,
                    draw = ER30$.draw) |>
  # for each posterior draw...
  group_by(draw) |>
  # ... calculate ATE
  summarise(ate = mean(ER31 - ER30))

# compare with marginaleffects
median_qi(ateR3$ate)
marginaleffects::avg_comparisons(r3mod,
                                 variables = list(R3 = "minmax"))
```

The objects `ateR1`, `ateR2` and `ateR3` store the posterior means and HPDIs of the joint effects of religiosity on subjective health. Inspect the plot below to see that the contrasts huddle mostly around zero, meaning that we find little evidence for the notion that religion on average works as a protective factor during the pandemic: At least in these (self-reported) data, more religious individuals were not more healthy. If anything, these data point to the opposite relationship, in that most of the joint posterior mass is negative.



One last comment, before we move on: In our estimation above, we silently cut some corners, for the sake of illustrating the basic principles of standardisation in a context with time-varying confounders. There are some additional modeling complexities that we could (and perhaps should) incorporate: Monotonicity in predictors; ordinal (not gaussian) outcome; random intercepts for individuals; Bayesian imputation; non-default priors.

In the next section, we discuss assumptions that are required for a causal interpretation of estimates obtained via standardization and IPTW.

4.4 It's assumptions all the way down

Causal inference in an observational setting generally requires several key conditions for a causal interpretation of the main exposure (e.g. Hernan and Robins 2020, ch. 13; Naimi, Cole, and Kennedy 2017). One way to think of this exercise is that when assumptions are (assumed) satisfied, an observational study will have emulated a randomized study.

First, we assume that the potential outcomes under varying levels of exposure are independent from the observed outcomes (*conditional exchangeability*). In a perfectly randomized trial, this is the case since randomization ensures that the probability of treatment is independent of the outcome — we say that the treatment and control groups are "exchangeable". However, in an observational setting, there can be countless

factors that both influence exposure levels and the outcome. It's a main goal of a statistical model to adjust for these confounding factors, in order to obtain conditional exchangeability. For instance, as we've seen, a DAG is useful for guiding statistical adjustment. In essence, we aim to statistically block all backdoor paths that both influence our exposure M and outcome Y of interest.

Second, and relatedly, we assume *no model misspecification*, which entails that our model is specified correctly (e.g., in terms of functional relationships, no omitted confounding variables, etc.). However, whether any given statistical model and adjustment sets are sufficient to ensure these conditions hold is generally not empirically testable.

Third, and similarly, we assume that our variables are *measured without error*, another difficult-to-verify assumption, unless there are known sources of measurement error.

Fourth, we assume that an individual's observed outcome under a given exposure is equivalent to the potential outcome that would've been observed under that exposure (*counterfactual consistency*). In other words, in the context of our g-computation procedure, when we obtain expected values for participants setting $X = x$, we assume that we obtain the values that we in fact would've observed, if those participants had been observed under $X = x$.

Fifth, we assume *positivity*, which implies that individuals have similar exposure levels within all confounder levels. While this is empirically unlikely to hold in particular (for instance, when we have several covariates, including continuous ones, several groups, etc.), lack of positivity can be ignored to the extent that we're willing to assume that estimates for the strata with zero observations can be extrapolated from the model fitted on the observed strata (Hernan and Robins 2020, 162).

Sixth, we assume no-interference, such that the potential outcomes of an individual is assumed to be unaffected by the treatment assignment of other individuals (also known as "the stable unit treatment value assumption" (SUTVA)).

Seventh, and more trivially, we assume temporal ordering of relevant variables, for instance such that exposures and mediating variables occur before an outcome.

5 Missing data

Causal inference is fundamentally a missing data problem. All real-world analyses are marred by missingness one way or the other. In the most basic sense, this directly relates to what we above called the fundamental problem of causal inference: We're always missing the counter-factual, so we impute it through assumption-laden estimation or by finding an appropriate comparison group.

But even setting this issue aside, and even if any particular dataset is "complete," that dataset will almost always only represent a sample from a bigger population. Often, it's the population, not the sample, that we're interested in making inferences about. So in that light, we are missing data on most of the population of interest.

Add to this the very plausible scenarios that our sample might not be a random subset of the population, that some of our variables are measured with error, or that our statistical models fail to capture key aspects of the data-generating process. These a tough but inevitable considerations. But all hope is not lost.

A comprehensive overview of data analysis in the presence of missing data is obviously outside the scope of this brief primer. Instead, I want to give a compact run down of ways to think about missing data and some practical pointers to handling it.

As touched upon just above, we'll distinguish between missingness relating to **external** validity (i.e., is our treatment effect estimate unbiased in the target population?) and **internal** validity (i.e., is our treatment effect estimate unbiased in the sample?) (Lesko et al. 2017).

5.1 External validity

Once again, DAGs can help structure our thinking. Say we're interested in estimating the effect of a health intervention on some outcome (Figure 6). Consider too that, even if we randomize the treatment assignment – i.e., there's no arrows pointing into "Treatment" – our sample might not reflect the target population that we're ultimately interested in. When it comes to external validity, we're interested in not only backdoor paths between the treatment and the outcome but also between the outcome and the sampling indicator ("Sampled" in Figure 6). Here, there is such a path, through "Health seeking". That is, if you're on the lookout for treatment of some ailment, you're probably more likely to both enroll in a health study and also have health conditions that are different from non-health seeking individuals (Westreich 2019, 208–9).

We box the "Sampled" node to emphasize that we've conditioned on the sampled individuals. This is similar to how we condition on variables by statistical adjustment, only here the conditioning happens at the design or sampling stage.

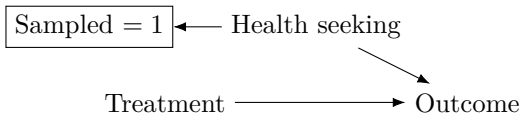


Figure 6: **Sampling selection bias**. There's an open backdoor from the outcome to the sampling indicator.

But it's not only in the health sciences that such causal structures tend to crop up. Consider, say, survey studies in general (Schuessler and Selb 2019). On the most basic level, if you plan to run a survey study, you shouldn't expect to hear from people who really don't want to participate in survey studies (Figure 7).

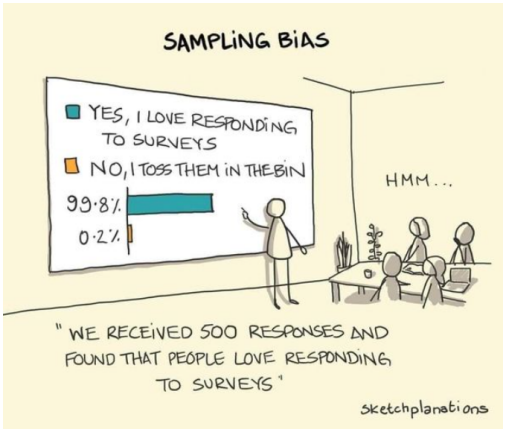


Figure 7: **Sampling selection bias**. If you run a survey study, you'll rarely hear from people who don't want to participate in survey studies.

Once you're habituated to thinking about external validity – or "sampling selection bias", to use another fancy term – this way, you'll see it everywhere. To give just one example, a recent survey study aimed at estimating how widespread spiritual needs are in the general Danish population (Stripp et al. 2023). This was an impressive data collection effort in that the researchers drew a random population-based sample of over 100,000 adult Danes, which was then linked up with national register-based data. This representative sample of Danes then received an invitation to participate in a survey. According to the invitation, the survey would include questions about existential and spiritual matters, such as "satisfaction with life", "meaning" and "faith". One in four responded.

Setting aside important definitional quibbles about what we even mean by "spirituality," here's the catch (as discussed in Mauritsen and Bendixen 2023): What if people's spiritual needs influence the likelihood that they respond to a survey on existential and spiritual matters? Then we have a scenario as in Figure 8, where people self-select into the study based not only on their demographic characteristics but also on their

(non-)spirituality. This in turn means that we're sampling on the outcome of interest and that the sample is therefore no longer representative of the target population.

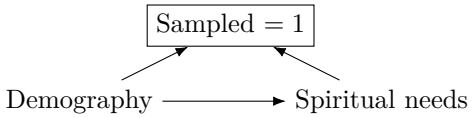


Figure 8: **"Spiritual needs are common among Danes, who respond to surveys on spiritual needs"**. Sampling selection bias in the wild.

When the outcome is associated with the sampling indicator, as in Figure 8, there is no straightforward way to obtain an unbiased estimate of the population-level distribution of the outcome – unless we're willing to make strong (and often unverifiable) assumptions about the exact sampling mechanisms at hand. That's the bad news.

But there are good news, too. Under less restrictive causal structures, we can in fact recover a population estimate even if our sample is only partially or not at all overlapping with the population of interest. We'll tackle an applied case in the next section.

5.1.1 Generalizability, Transportability, Poststratification

As we saw above, whether an empirical estimate from one sample can be generalized to another population depends on explicit causal assumptions. These assumptions can be encoded in DAGs (e.g., Pearl and Bareinboim 2022; Deffner, Rohrer, and McElreath 2022; Schuessler and Selb 2019). If the outcome of interest is associated with sampling, we've selected on our outcome and there's little we can do, without further assumptions (think of Figure 8). However, if we can break the association between the sampling indicator and the outcome through statistical adjustment, there's hope.

Figure 9 shows one such example. Say we're interested in estimating the effect of an experimental prosociality prime on choice in an behavioral economic game across different field sites. The study and data come from House et al. (2020) and the analysis example from Deffner, Rohrer, and McElreath (2022).

We suspect that our outcome is associated with sampling through age. This scenario could arise for at least two main reasons. First, different field sites may simply have different underlying demographic compositions, such that for instance one population is older or younger than the other. In this case, and assuming our sample is a sufficiently large random sample of each field, our sample estimates will reflect the true population distributions. If we're only interested in the empirical differences between our sampled populations, no further action is needed.

But what if different sampling procedures caused our samples to differ in age? Perhaps some field researchers sampled children more than adults, because they were primarily interested in developmental aspects of prosociality. Indeed, children were preferentially sampled by House et al. (2020). In that case, we cannot simply compare the empirical population differences, because our sample is not a random sample of the target populations. We need a statistical procedure – so-called **poststratification** – to adjust for such a scenario.

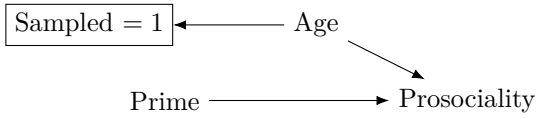


Figure 9: **Transportability example**.

In general, poststratification involves three main steps (Kennedy and Gelman 2021):

- Fit a theoretically informed statistical model on the sample adjusting for relevant variables to block backdoor paths between the outcome and exposure and the outcome and sampling indicator.

- Source external data on the demographic composition of a target population (e.g., census data).
- Obtain predictions from the fitted model and reweight the fitted values with weights computed from the external target population data.

In the following, we'll go through each step using as an illustrative example, adapted from House et al. (2020) and Deffner, Rohrer, and McElreath (2022).

5.1.1.1 Fit statistical model

First, we prepare the data from House et al. (2020). There were three experimental priming conditions but we look at only two here – "Generous" and "Selfish" – and a binary outcome, where the "Prosocial choice" = 1. Age ranges from 4-15 years but is recoded to a 1-12 range for practical indexing purposes (see source document for data preparation).

Then, we fit a statistical model that adjusts for priming condition and age. Now, this model is a little bit more involved than previous models. Since these complications pertain more to estimation than causality *per se*, I'll just very briefly break it down bit by bit and include references for further reading. I'll also use default priors here, although in an applied case, we'd want to think carefully about defining appropriate prior distributions and checking their implications.

```
model <- brm(outcome ~ 1 + condition*mo(age) + (1 + condition*mo(age) | pop_id),
             family = bernoulli,
             data = d_list,
             cores = 4, iter = 1000)
```

First, since the outcome is binary, we use logistic (Bernoulli) regression, `family = bernoulli` (McElreath 2020, ch. 11). This means that we're predicting the probability of choosing the prosocial option, conditionally on the covariates. Second, we interact priming condition and age (indicated by the `*`), as we're interested in how age modifies the effect of priming (McElreath 2020, ch. 8). That is, is priming more influential in certain age groups? Third, we model age with *monotonic effects* (indicated by `mo()`). This allows for a more flexible relationship between age and the outcome. Specifically, each year of age is allowed to have its own differential effect on the outcome, while assuming shared directionality across all age groups (Bürkner and Charpentier 2020; McElreath 2020, ch. 12). Finally, we're allowing each field site to have its own intercept and slope to capture cultural-specific factors, `(... | pop_id)`. This multilevel specification ensures that each field site is simultaneously *informing* and *informed by* all other field sites, making for more stable predictions, particularly for groups and combinations of covariates with few data points (McElreath 2020, ch. 13-14). For this reason, multilevel modeling is often combined with poststratification (Kennedy and Gelman 2021).

5.1.1.2 Source external data

Next step in our poststratification procedure is to settle on an appropriate target population and obtain external data on that population. Since we're not guided by any particular theoretical research program here, we'll simply follow Deffner, Rohrer, and McElreath (2022) and poststratify to the Wichí in the sample, an indigenous group of Argentina.

Another perhaps more common situation is that we want poststratification to a more representative population. We could then look for census data for this general population of interest. For European countries, one could use Eurostat[3], which is a datahub for European census data. US census data are also readily available[4].

Returning to our present analysis, we arrange the Wichí data in a format that resembles census data (Table 1). Census data are often structured such that, for each combination of demographic variables of interests, we get the population frequency or proportion. Once we've calculated the proportional age distribution for the Wichí, we copy that to the remaining sites.

```
# data resembling census format
psw <- model$data |>

  # calculate number of individuals in each age group and site
```

[3] https://ec.europa.eu/eurostat/databrowser/view/EDAT_LFS_9901___custom_5352473/default/table?lang=en
[4] https://usa.ipums.org/usa/index.shtml

Table 1: Census-like data format for the Wichí.

Age	Frequency	Proportion
1	0	0.00
2	0	0.00
3	1	0.02
4	4	0.07
5	5	0.09
6	11	0.19
7	6	0.10
8	9	0.16
9	16	0.28
10	5	0.09
11	1	0.02
12	0	0.00

```
  group_by(as.factor(pop_id), as.factor(age), .drop = F) |>
  summarise(freq = n()) |>
  ungroup() |>

  # rename columns
  rename("pop_id" = "as.factor(pop_id)",
         "age" = "as.factor(age)") |>

  # select only Wichi (target pop)
  filter(pop_id == 6) |>

  # get proportions for each combination of demographic variables
  mutate(prop = freq/sum(freq)) |>
  ungroup() |>

  # copy age proportions to all other sites
  mutate(pop_id = NULL,
         age = as.integer(age)) |>
  expand_grid(pop_id = as.factor(1:6))
```

5.1.1.3 Obtain and reweight fitted values Finally, we obtain predictions from the fitted model and reweigh the fitted values with the weights (i.e., proportions) computed from the external target population data in the previous code chunk. For computing counterfactual predictions and marginal effects, we follow the – by now – familiar g-computation workflow from above.

```
# set X=1, Z=z
psX1 <- expand_grid(psw, condition=1)

# set X=0, Z=z
psX0 <- expand_grid(psw, condition=0)

# E[Y{Z=z, X=1}]
psEX1 <- add_epred_draws(model, newdata = psX1,
                         value = "estimate") |>
```

```
  # weight predictions with ps weights
  mutate(estimate_prop = estimate*prop) |>

  # for each draw and site, sum up weighted predictions
  group_by(pop_id, .draw) |>
  summarise(estimate_sum = sum(estimate_prop)) |>
  rename(.epred = estimate_sum)

# E[Y{Z=z, X=0}]
psEX0 <- add_epred_draws(model, newdata = psX0,
                         value = "estimate") |>

  # weight predictions with ps weights
  mutate(estimate_prop = estimate*prop) |>

  # for each draw and site, sum up weighted predictions
  group_by(pop_id, .draw) |>
  summarise(estimate_sum = sum(estimate_prop)) |>
  rename(.epred = estimate_sum)

# E[Y{X=1} - Y{X=0}]
poststratified_ate <- data.frame(X1 = psEX1$.epred,
                                 X0 = psEX0$.epred,
                               draw = psEX0$.draw,
                             pop_id = psEX0$pop_id) |>
  # for each posterior draw and site...
  group_by(pop_id, draw) |>
  # ... calculate ATE
  summarise(ate = mean(X0 - X1))
```

Finally, we plot the estimates. When working with poststratification, it's often useful to have alternative, non-poststratified estimates to plot against. This allows us to assess the impact of the poststratification procedure on our inferences.

We calculate such a simplified model in the following code chunk and obtain marginal effects of the priming condition for each site. Then, in the next code chunk, we plot the two results against each other.

```
# Empirical estimate
fixmod <- brm(outcome ~ 1 + condition*pop_id,
              family = bernoulli,
              data = d_list,
              cores = 4, iter = 1000,
              backend = "cmdstan")

# set X=1, Z=z
X1 <- transform(d_list, condition=1)

# set X=0, Z=z
X0 <- transform(d_list, condition=0)

# E[Y{Z=z, X=1}]
EX1 <- add_epred_draws(fixmod, newdata = X1, value = ".epred")

# E[Y{Z=z, X=0}]
EX0 <- add_epred_draws(fixmod, newdata = X0, value = ".epred")
```

```r
# E[Y{X=1} - Y{X=0}]
sample_ate <- data.frame(X1 = EX1$.epred,
                         X0 = EX0$.epred,
                       draw = EX0$.draw,
                     pop_id = EX0$pop_id) |>
  # for each posterior draw and site...
  group_by(pop_id, draw) |>
  # ... calculate ATE
  summarise(ate = mean(X0 - X1))
```

```r
# Combine and compare the two sets of estimates with an indicator for poststratified
poststratified_ate$ps <- "Poststratified"
sample_ate$ps <- "Empirical"
compare <- rbind(sample_ate,
                 poststratified_ate) |>
  mutate(ps = as.factor(ps))

# facet labels
pop_label <- c("Berlin (GER)","La Plata (ARG)","Phoenix (USA)", "Pune (IND)", "Shuar (ECU)", "Wichí (AR
names(pop_label) <- 1:6

# plot!
compare |>
  ggplot(aes(x = ate, group = ps, color = ps, fill = ps)) +
  facet_wrap(~ pop_id,
             labeller = labeller(pop_id = pop_label)) +
  geom_line(aes(y=after_stat(scaled)), alpha = 0.5, size = 2, stat = "density") +
  geom_vline(xintercept = 0, linetype = "dotted") +
  scale_fill_manual(values = c("gray40", "#3182BD")) +
  scale_color_manual(values = c("gray40", "#3182BD")) +
  theme_test(base_size = 8) +
  theme(axis.text.y=element_blank(),
      axis.ticks=element_blank(),
      axis.title.y=element_blank()) +
  ylab(NULL) +
  xlab("Effect of prime on prosocial choice in dictator game") +
  labs(title = "Empirical vs. Poststratified estimates (census weights-approach)",
       subtitle = "Poststratied to the Wichí") +
  guides(color=guide_legend(title=NULL))
```

As it turns out (Figure 10), the "poststratified estimates" do differ somewhat from the "empirical estimates", except for Wichí. This is not surprising since, after all, we poststratified to the age distribution of Wichí. In all other sites but for Phoenix (USA), the poststratified effect is larger than the empirical estimates. For Phoenix, it's smaller. This is because all other sites are nudged toward the Wichí estimate, which is around an 0.5, and only Phoenix's empirical estimate is above that.

5.1.1.4 Taking stock: Simulating counterfactual populations It might be helpful at this point to pause for a moment and consider what we've actually achieved here. The empirical estimates are the average treatment effects *in the sample* for each site. However, we suspect that the in-sample effects across field sites might differ *from the population effects* both because of unobserved cultural factors and because of different sampling procedures. So, we conduct a counterfactural experiment: *What if all field sites had an age distribution identical to the Wichí?* This procedure, in effect, holds age constant such that the residual variation in the treatment effect that might still exist between sites can now only be due to site-specific cultural factors – of course, given the usual identification assumptions.

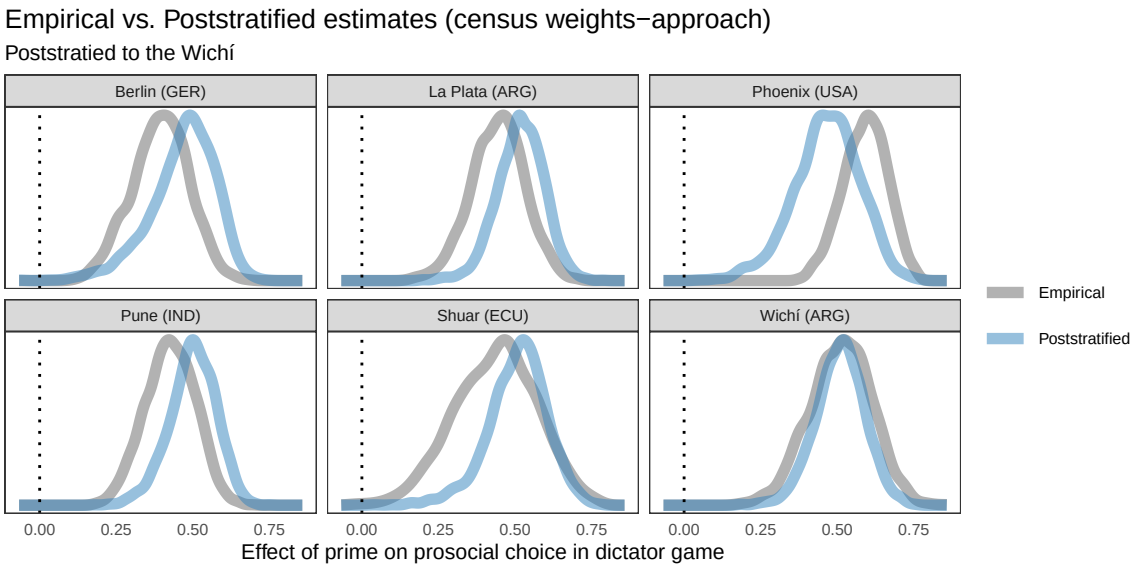


Figure 10: Poststratification example.

To think of poststratification this way – as a counterfactual demographic experiment – follows neatly in line with the potential outcomes framework that guides this primer. Throughout, we've used potential outcomes to identify an average treatment effect, simulating counterfactual exposures to treatment. In the poststratification procedure, we instead simulate counterfactual demographic compositions that might modify the outcome of interest. This makes for a powerful workflow for *estimating* a treatment effect in a sample and *generalizing* that effect to a population of interest.

Sometimes, these two estimands are referred to, respectively, as the **sample average treatment effect (SATE)** and the **(population) average treatment effect (ATE)** (Figure 2). So far in this primer, we've computed only the SATE. But, our sample is not a random subset of the population of interest, poststratification is strictly needed to get at the (P)ATE.

5.1.1.5 Extensions, assumptions and caveats The workflow presented above is, of course, simplified to drive home both the conceptual and computational points. However, it's easily scaleable. For instance, say we wanted to not only look at age but also gender. Assuming gender is measured in-sample, we simply include gender in the model, calculate the proportions of the target population with all possible combinations of age and gender from external data, and use those proportions to weight model predictions. In the source document, I show a complete example.

Second, if – as is the case in our working example – we have access to individual-level data, and not census-like aggregated data, for the target population, we can simplify the computational workflow somewhat. Instead of calculating proportions as weights, we simply assemble a data frame with the individual-level data of the target population. We then use that data frame to get fitted values from our model and calculate marginal effects, as usual. Below, I show this modified procedure but leave it to the reader to verify that this approach yields identical results to our census-like approach above.

```
### Poststratification with individual-level data

# Filter for target population
target_pop <- model$data[model$data[["pop_id"]]==6,] |>

  # and then copy target age distribution to all other sites
  mutate(pop_id = NULL) |>
  expand_grid(pop_id = as.factor(1:6))
```

```r
# set X=1, Z=z
psX1 <- transform(target_pop, condition=1)

# set X=0, Z=z
psX0 <- transform(target_pop, condition=0)

# E[Y{Z=z, X=1}]
psEX1 <- add_epred_draws(model, newdata = psX1, value = ".epred")

# E[Y{Z=z, X=0}]
psEX0 <- add_epred_draws(model, newdata = psX0, value = ".epred")

# E[Y{X=1} - Y{X=0}]
poststratified_ate <- data.frame(X1 = psEX1$.epred,
                                 X0 = psEX0$.epred,
                               draw = psEX0$.draw,
                             pop_id = psEX0$pop_id) |>
  # for each posterior draw and site...
  group_by(pop_id, draw) |>
  # ... calculate ATE
  summarise(ate = mean(X0 - X1))
```

Another important point to emphasize is that our poststratified estimates are only valid under certain assumptions. As it turns out, identification assumptions for internal validity, as discussed in Section 4.4, have counterparts to identification assumptions for external validity. Quoting from Lesko et al. (2017), we assume that:

- "The participants included in the study sample are exchangeable with members of the target population who were not sampled, perhaps conditional on pre-treatment characteristics W (conditional exchangeability between those sampled and those not sampled)"
- "Within strata of W, all subjects in the target population have some non-zero probability of being selected into the sample (analogous to positivity)"
- "The same distribution of versions of treatment [are administered] in the study sample and the target population"
- "[There is] no interference in the target population and the study sample (although these results can be extended to scenarios where the pattern of interference is the same in the target population and the study sample)" (similar to "the stable unit treatment value assumption").
- "We assume no measurement error [and] also require correct model(s) specification for any parametric or semi-parametric models used to describe associations between covariates and outcome or any models used to describe the sampling mechanism"

In a final quote, Lesko et al. (2017) sums up the connection between internal and external validity as follows:

> "Assumptions sufficient for identification of a causal effect in the target population may, at first glance, look similar to those required for identification of a causal effect in the study sample. However, assumptions about the relationships between the potential outcomes and the sampling mechanism are sufficient for external validity, compared to the case of internal validity for which assumptions about the relationships between the potential outcomes and the treatment assignment mechanism are sufficient. As assumptions sufficient for internal validity are met in expectation when treatment is randomized, assumptions sufficient for external validity will be met in expectation if the study sample is a simple random sample of the target population."

To connect these insights to our working example, consider that in Figure 9, we assumed that the exposure

Table 2: Overview of some treatment effects.

Effect	Description	Target
ATE (population average treatment effect)	Average	Population
SATE (sample average treatment effect)	Average	Sample
ITT (population intention to treat effect)	Average among units assigned to treatment	Population
SITT (sample intention to treat effect)	Average among units assigned to treatment	Sample
ATT (population average treatment effect on treated)	Average among the actual treated	Population
SATT (sample average treatment effect on the treated)	Average among the actual treated	Sample
LATE (population local average treatment effect	Average among the compliers	Population
SLATE (sample local average treatment effect	Average among the compliers	Sample
ITE (individual treatment effect)	Indvidual unit effect	Unit

of interest – the prime – was experimentally manipulated. This meant that we didn't have to worry about backdoor paths between the outcome and exposure: Assuming perfect randomization, perfect measurement and no model misspecification, our estimate therefore represents a valid causal estimate in the sample and – to the extent that Figure 9 captures the true causal structure of sampling – in the population of interest.

However, in observational settings, we'd have to worry about those backdoor paths, too. Our poststratified estimate will only have a valid causal interpretation, if the in-sample estimate is identified. Conversely, if we're using poststratification but our poststratification adjustment misses key variables, our poststratified estimate will arguably be worse than the empirical estimate, which at least holds in-sample, given satisfied internal validity assumptions (Schuessler and Selb 2019).

5.1.1.6 Generalizability vs. transportability: What's the difference?

According to Lesko et al. (2017):

> "Generalizability is concerned with making inference from a possibly biased sample of the target population back to the full target population (including the study sample), while transportability concerns making inference for a target population when the study sample and the target population are partially or completely non-overlapping."

5.2 Internal validity

In the previous section, we focused on external validity. That is, what can we do about the fact that our sample will almost always only be a subset – and often a *non-random* one – of the population of interest. In other words, most of the data will be missing most of the time. We could term this *unit missingness* – we're missing data on entire units (e.g., individuals).

But, above and beyond unit missingness, we'll also very often have missing values certain variables for even sampled units – *variable missingness*. This kind of missingness is what most people think of as missing data. And it is indeed ubiquitous. But as with unit missingness, there are principled strategies for thinking about variable missingness that in turn can guide analysis. The key idea here too is that we aim for the missing values to be conditionally exchangeable with the observed values. For this purpose, we'll again use DAGs and missingness indicators to encode our assumptions about the data-generating process (**thoemmes2015graphical?**).

5.2.1 Missing (completely) at random

5.2.2 Missing not at random

6 To do:

- Plot results in a simple style
- double check notation throughout – within-document consistency and external consistency with e.g., What if? book
- a little bit more formal detail on the IPTW weights and the difference btw. unstabilized and stabilized weigths and why we can't rely on the standard errors (and therefore must resort to bootstrapping)
- doubly robust estimators
- Missing data MCAR, MAR, MNAR. See Morris et al. (2022, appendix) for discussion of missingness in X and/or Y in the context of g-comp.

Above, we conducted a complete cases analysis, under the assumption that missingness is unsystematic. One way to investigate this assumption a little bit further is to ask, *does health predict dropout at the final measurement*? This is not a bulletproof "test" by any means, but it's a start.

```
veluxData <- read.csv("velux_data.csv")

veluxData <- with(veluxData,
          data.frame(
            id,
            R1 = religion_1,
            R2 = religion_2,
            R3 = religion_3,
            H1 = health_1,
            H2 = health_2,
            H3 = health_3,
            H4 = health_4,
            G = gender,
            A = age.c,
            E = education,
            I = household_income)
)

# 1 if health is missing at t4, 0 otherwise
veluxData$H4miss <- ifelse(is.na(veluxData$H4), 1, 0)
```

```
# What predicts missingness at final follow-up?
hNAmod <- brm(H4miss ~ mo(H1) + mo(H2) + mo(H3) + G + A + mo(E) + mo(I),
                       data = veluxData,
                       family = "bernoulli",
                       cores = 4)
```

```
# Does religion at previous measurements predict missingness at final follow-up?
rNAmod <- brm(H4miss ~ mo(R1) + mo(R2) + mo(R3),
                       data = veluxData,
                       family = "bernoulli",
                       cores = 4)
```

See end of Standardization chapter in What if?

- Econometric techniques

Regression discontinuity, diff-in-diff, instrumental variables, synthetic controls

- Threats to identification and some solutions

Attrition (simple MNAR analysis), selection bias, spill-over, non-compliance (IV of randomization), etc.

- Transportability of treatment effects (see What if? chapter); also extends naturally from standardisation/g-computation.
- Red herrings Testing for covariate imbalances
- Dictionary for common terminology in the causal inference literature

D-separation, Do-calculus,

- Principles for randomized trials/experiments

Adjusting for covariates? Fixed effects or multilevel modeling? Randomization strata?

- Per-protocol vs. as-treated vs. intention to treat (see e.g. Hernan and Robins 2020, ch. 22)

7 Appendix: Velux IPTW

I said above that we'd leave IPTW behind., Well, not quite. Here's how to use IPTWs and a MSM to analyze the panel data on health and religiosity.

Note that we use only the complete cases. Also, this illustration differs from the IPTW above, because now we have a continuous, not a binary, exposure (well, it's categorical, but we assume gaussian for the sake of simplicity). The IPTWs are calculated slightly differently, when the exposure is continuous. All this said, the results are qualitatively (although not numerically, likely due to difference in estimation and also different sample subsets) similar to our standardization approach, in that we find no noteworthy effect of religiosity on health.

However, try and run the bootstrap and subsequent steps *without* the weights (i.e., delete the `weights` argument from `iptw_fun()`); you'll see that the model then picks up a "near-statistically significant" negative association between R_3 and the outcome, which we can only guess is spurious.

```
diptw <- d[complete.cases(d),] # complete cases

## Computing IPTW with a continuous exposure: https://www.andrewheiss.com/blog/2020/12/01/ipw-binary-co

## the stabilized weights for religion_t1 is 1 (since the numerator and the denominator is the same)
## cf., VanderWeele et al. (online appendix)

## numerator model for religion_t2
r2exppn <- glm(religion_2 ~ gender + age.c + education + household_income + religion_1, data = diptw)
diptw$pn_r2 <- dnorm(diptw$religion_2,
             predict(r2exppn),
             sd(r2exppn$residuals))

## denominator model for religion_t2
r2exppd <- glm(religion_2 ~ gender + age.c + education + household_income + religion_1 + health_1, data
diptw$pd_r2 <- dnorm(diptw$religion_2,
             predict(r2exppd),
             sd(r2exppd$residuals))

## numerator model for religion_t3
r3exppn <- glm(religion_3 ~ gender + age.c + education + household_income + religion_2, data = diptw)
diptw$pn_r3 <- dnorm(diptw$religion_3,
             predict(r3exppn),
             sd(r3exppn$residuals))
```

```r
## denominator model for religion_t3
r3exppd <- glm(religion_3 ~ gender + age.c + education + household_income + religion_2 + health_1 + heal
diptw$pd_r3 <- dnorm(diptw$religion_3,
             predict(r3exppd),
             sd(r3exppd$residuals))

## calculate weights
diptw$sw2 <- with(diptw, pn_r2/pd_r2)
diptw$sw3 <- with(diptw, pn_r3/pd_r3)
diptw$sw <- with(diptw, sw2*sw3)

# MSM without and with stabilized weights
velux_uw <- glm(health_4 ~ gender + age.c + education + household_income + religion_1 + religion_2 + rel
summary(velux_uw)

velux_msm <- glm(health_4 ~ gender + age.c + education + household_income + religion_1 + religion_2 + re
summary(velux_msm)

### bootstrapping, when weights are calculated
library(boot)

# same function as above
iptw_fun <- function(formula, data, indices) {
  d <- data[indices,]
  fit <- glm(formula, family="gaussian", weights = sw, data=d)
  return(coef(fit))
}

iptw.velux.result <- boot(data = diptw,
                 statistic = iptw_fun,
                 R = 1e4,
                 formula = health_4 ~ gender + age.c + education + household_income + religion_1 + relig

# bootstrapped point estimate for religion_t1
iptw.velux.r1point <- iptw.velux.result$t0[6]

# bootstrapped point estimate for religion_t2
iptw.velux.r2point <- iptw.velux.result$t0[7]

# bootstrapped point estimate for religion_t3
iptw.velux.r3point <- iptw.velux.result$t0[8]

# bootstrapped interval for religion_t1
iptw.velux.r1interval <- boot.ci(iptw.velux.result,
                         type = "norm",
                         index = 6)$normal

# bootstrapped interval for religion_t2
iptw.velux.r2interval <- boot.ci(iptw.velux.result,
                         type = "norm",
                         index = 7)$normal

# bootstrapped interval for religion_t3
```

```
iptw.velux.r3interval <- boot.ci(iptw.velux.result,
                         type = "norm",
                         index = 8)$normal
```

References

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22 (4): 327–39.

Ahern, Jennifer, Alan Hubbard, and Sandro Galea. 2009. "Estimating the Effects of Potential Public Health Interventions on Population Disease Burden: A Step-by-Step Illustration of Causal Inference Methods." *American Journal of Epidemiology* 169 (9): 1140–47.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Austin, Peter C, and Elizabeth A Stuart. 2015. "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine* 34 (28): 3661–79.

Bendixen, Theiss. 2023. "100,000 Regressions Do Not Make for Causal Inference." *Science Advances*. https://www.science.org/doi/10.1126/sciadv.abn3517#elettersSection.

Bürkner, Paul-Christian, and Emmanuel Charpentier. 2020. "Modelling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models." *British Journal of Mathematical and Statistical Psychology* 73 (3): 420–51. https://doi.org/https://doi.org/10.1111/bmsp.12195.

Chesnaye, Nicholas C, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. 2022. "An Introduction to Inverse Probability of Treatment Weighting in Observational Research." *Clinical Kidney Journal* 15 (1): 14–20.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. "A Crash Course in Good and Bad Controls." *Sociological Methods & Research*, 00491241221099552.

Deffner, Dominik, Julia M Rohrer, and Richard McElreath. 2022. "A Causal Framework for Cross-Cultural Generalizability." *Advances in Methods and Practices in Psychological Science* 5 (3): 25152459221106366.

Hernan, MA, and J Robins. 2020. "Causal Inference: What If?"

House, Bailey R, Patricia Kanngiesser, H Clark Barrett, Tanya Broesch, Senay Cebioglu, Alyssa N Crittenden, Alejandro Erut, et al. 2020. "Universal Norm Psychology Leads to Societal Diversity in Prosocial Behaviour and Development." *Nature Human Behaviour* 4 (1): 36–44.

Kennedy, Lauren, and Andrew Gelman. 2021. "Know Your Population and Know Your Model: Using Model-Based Regression and Poststratification to Generalize Findings Beyond the Observed Sample." *Psychological Methods* 26 (5): 547.

Lesko, Catherine R, Ashley L Buchanan, Daniel Westreich, Jessie K Edwards, Michael G Hudgens, and Stephen R Cole. 2017. "Generalizing Study Results: A Potential Outcomes Perspective." *Epidemiology (Cambridge, Mass.)* 28 (4): 553.

Lübke, Karsten, Matthias Gehrke, Jörg Horst, and Gero Szepannek. 2020. "Why We Should Teach Causal Inference: Examples in Linear Regression with Simulated Data." *Journal of Statistics Education* 28 (2): 133–39.

Mauritsen, Anne L, and Theiss Bendixen. 2023. *Are Spiritual Needs Ubiquitous? Conceptual, Statistical, and Sampling Biases in a Recent Study on Spirituality and Health in Denmark*. PsyArXiv.

Mauritsen, Anne L, Theiss Bendixen, and Henrik R Christensen. 2022. "Does a Pandemic Increase Religiosity in a Secular Nation? A Longitudinal Examination." PsyArXiv. https://doi.org/10.31234/osf.io/qsgej.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Second. CRC Press.

Morgan, Stephen L, and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Second. Cambridge University Press.

Naimi, Ashley I, Stephen R Cole, and Edward H Kennedy. 2017. "An Introduction to g Methods." *International Journal of Epidemiology* 46 (2): 756–62.

Pearl, Judea, and Elias Bareinboim. 2022. "External Validity: From Do-Calculus to Transportability Across Populations." In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 451–82.

Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Purzycki, Benjamin Grant, Theiss Bendixen, and Aaron D. Lightner. 2022. "Coding, Causality, and Statistical Craft: The Emergence and Evolutionary Drivers of Moralistic Supernatural Punishment Remain Unresolved." *Religion, Brain & Behavior*. https://doi.org/10.1080/2153599X.2022.2065349.

Robins, James M. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained

Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7 (9-12): 1393–1512.

Robins, James M, Miguel A Hernán, and Larry Wasserman. 2015. "On Bayesian Estimation of Marginal Structural Models." *Biometrics* 71 (2): 296.

Rohrer, Julia M. 2018. "Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42.

Saarela, Olli, David A Stephens, Erica EM Moodie, and Marina B Klein. 2015. "On Bayesian Estimation of Marginal Structural Models." *Biometrics* 71 (2): 279–88.

Schuessler, Julian, and Peter Selb. 2019. "Graphical Causal Models for Survey Inference." SocArXiv. https://doi.org/10.31235/osf.io/hbg3m.

Snowden, Jonathan M, Sherri Rose, and Kathleen M Mortimer. 2011. "Implementation of g-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique." *American Journal of Epidemiology* 173 (7): 731–38.

Stripp, Tobias Anker, Sonja Wehberg, Arndt Büssing, Harold G Koenig, Tracy A Balboni, Tyler J VanderWeele, Jens Søndergaard, and Niels Christian Hvidt. 2023. "Spiritual Needs in Denmark: A Population-Based Cross-Sectional Survey Linked to Danish National Registers." *The Lancet Regional Health–Europe*.

VanderWeele, Tyler J, John W Jackson, and Shanshan Li. 2016. "Causal Inference and Longitudinal Data: A Case Study of Religion and Mental Health." *Social Psychiatry and Psychiatric Epidemiology* 51: 1457–66.

Vansteelandt, Stijn, and Niels Keiding. 2011. "Invited Commentary: G-Computation–Lost in Translation?" *American Journal of Epidemiology* 173 (7): 739–42.

Wang, Aolin, and Onyebuchi A Arah. 2015. "G-Computation Demonstration in Causal Mediation Analysis." *European Journal of Epidemiology* 30: 1119–27.

Westreich, Daniel. 2019. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press.

Westreich, Daniel, and Sander Greenland. 2013. "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients." *American Journal of Epidemiology* 177 (4): 292–98.

Wysocki, Anna C, Katherine M Lawson, and Mijke Rhemtulla. 2022. "Statistical Control Requires Causal Justification." *Advances in Methods and Practices in Psychological Science* 5 (2): 25152459221095823.