Introduction to Structural Models for Health Disparities

Ashley I Naimi

Jan 2023

Contents

1 Introduction

Social epidemiologists are often interested in evaluating the multifaceted interrelationships between social, political, and/or economic constructs, and health related constructs. Every so often, researchers in social epidemiology will rely on analytic methods that are more commonly used in the social sciences. Among these methods include structural equations models, which consist of a set of equations defining the **structural, or causal** relationships between variables in a given system of interest, combined into a single model. At times, latent (or unmeasured) variables are included in the model under a set of assumptions governing their relations with other measured variables in the system. Structural equations models are often used due to their perceived ability to decompose a set of relations between variables into their component mechanisms.

Several examples / implementations of structural equations models targeting health disparities and/or social epidemiology questions are available in the literature. SEMs have been used to evaluate the relationship and mechanisms between socioeconomic status and smoking (Martinez et al., 2018), race/ethnicity and childhood asthma (Sidora-Arcoleo et al., 2012), social and behavioral variables on a range of health outcomes (Hartwell et al., 2019), and the relationship between weekly working hours and the incidence of injury (Arlinghaus et al., 2012), among other things.

> **Structural Models**:
> While the word “structural” is often used to connote causal when used in the context of modeling data. However, there are generally two phases where the word “structural” has been used for data analysis. I refer to these phases as pre- and post-counterfactual. Linear structural equation models, for example, have been around and in use since the early 20th century, originating in the work of the American Geneticist Sewell Wright. Unfortunately, causal inference at the turn of the 19th century was highly undeveloped. It was only in the early 1980s where many of the theories around the conditions needed to estimate causal effects using observed data were developed. Structural models developed since then are characterized by a much more complete understanding of the conditions needed to interpret model coefficients causally.

2 Structural Equation Models

To give you an example of what a structural equation model entails, consider the SEM depoloyed by Arlinghaus to evaluate the relationship between work hours and injury. The model can be represented by the diagram in Figure 1:

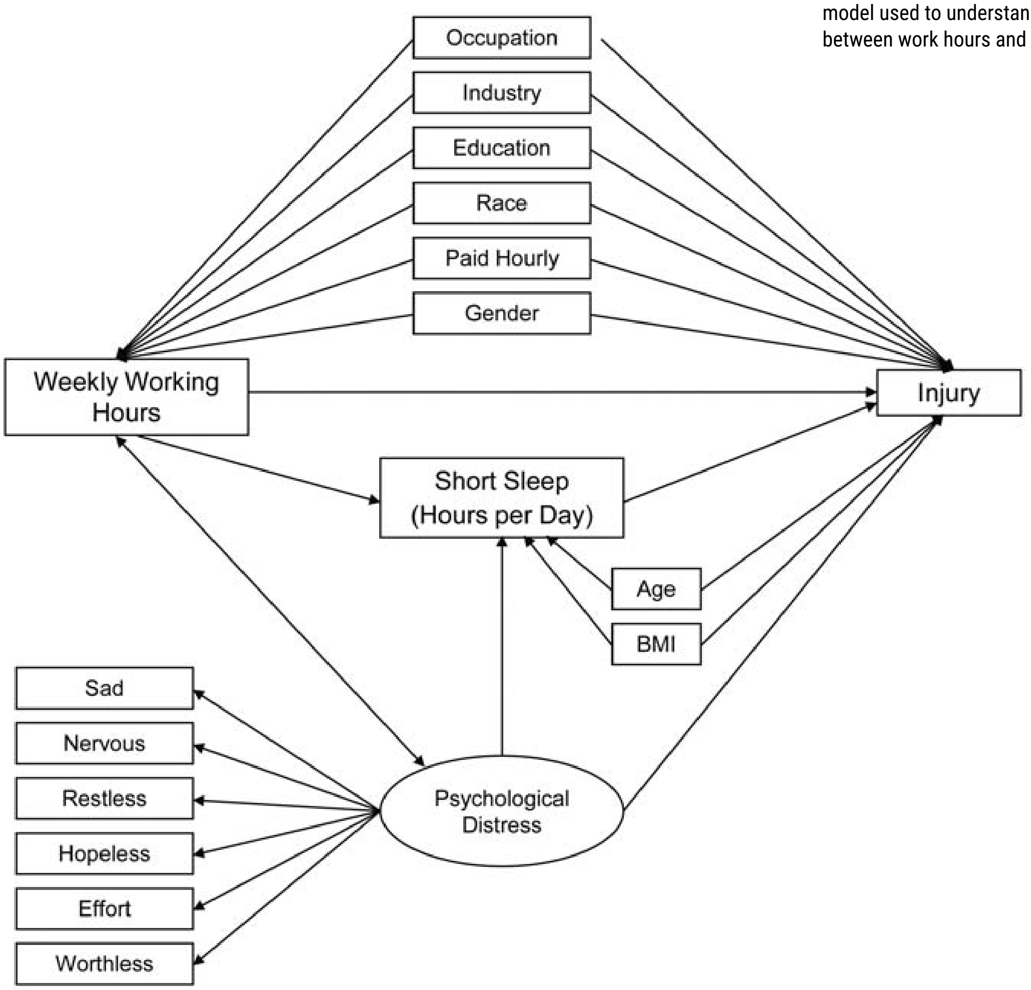


Figure 1: Illustration of a structural equation model used to understand the relationship between work hours and injury.

This model can be used to evaluate a number of different questions about the relation between the variables depicted in Figure 1. For example, we may ask:

- How much of the association between BMI and Injury is mediated by Short Sleep?
- Is the pathway from Race to Work Hours to Psychological Distress to Injury

stronger than the pathway from Race to Work Hours to Short Sleep to Injury?

- What is the relation between Industry and Injury?
- Does BMI affect Short Sleep?
- Is the relation between Psychological Distress and Injury stronger through the Short Sleep pathway?

This is often seen as a strength of structural equation models: several questions can be answered with the same model fit. Unfortunately, this feature of SEMs comes at the cost of making several fairly hefty assumptions about the *nature* of the relationship between all of the variables in the system (VanderWeele, 2012).

3 Problems with Structural Equation Models

Linearity, **extensive** linearity: One assumption characterizing SEMs is that the relationship between all of the variables in thy system are assumed to be linear. What this means is that a single unit increase in each parent variable leads to a single unit increase in each descendant variable. This can be a problem if there's a non-proportionate increase in one or more of the descendant variables (Seber and Wild, 1989).

> **The Cost of Assumptions**:
> Assumptions are embedded into all of science, particularly when we use statistical methods to analyze data. Some of these assumptions are commonly understood (or misunderstood). However, the making of assumptions can sometimes be a bit cavalier in empirical research. While the particular impact of a given assumption or a given set of assumptions can be difficult to specify, the general idea that should be understood is: the more assumptions you make, the higher the chance that one of those assumptions isn't true, and the more likely it is that you get the wrong answer from your particular analysis. Some assumptions are worse than others, in that they are difficult to justify AND they can more than likely lead to a full reversal of the association being studied.

Arrow Absence: In any kind of structural equation model a relationship between two variables is depicted by a directed edge (arrow). The presence of an arrow between two variables means that there is direct causal relationship. However, the absence of an arrow implies that there is no relationship between two variables. Generally, assuming that there is no relationship (no arrow) between two variables is stronger than assuming that a relationship exists (arrow present). This is because assuming no arrow forces the relationship between two variables to be excactly zero. However, including an arrow between two variables allows the relationship to be any number, including zero (Greenland et al., 1999).

Variable Absence: Excluding a variable from a structural equation model can have important implications for the accuracy and bias of the algorithm. In particular, it's important to include enough variables in the system such that there is no unmeasured confounding. However, when interest lies in a large set of variables in a complex system such as the SEM above, it becomes inordinately difficult to consider all of the relevant variables needed to adjust for confounding. For example, in the SEM presented by Arlinghaus above, it is likely that the age and or number of children confounds the relationship between work hours and sleep, but it is not included in the SEM. Exercise and physical fitness are likely confounders of the relationship between hours of sleep and injury, but there are also not included in the SEM (Greenland et al., 1999).

Sensitivity Analysis for Unmeasured Confounding: Of course, excluding such confounders is a common occurrence in epidemiologic studies. However, for more traditional (single outcome, single exposure) regression analyses, there are a host of methods to explore the sensitivity of results to the presence of such unmeasured confounding. Unfortunately, no such methods exist for SEMs, leaving us in the dark about how sensitive our results may be to the absence of potential confounders (McCandless and Gustafson, 2017).

Interactions, particularly for mediation contrasts: SEMs are often used for mediation analysis, where it's important account for exposure-mediator interactions when they are present. Several mediation techniques exist based on counterfactuals that can be used to quantify well-defined mediation effects when exposure-mediator interactions are present. However, it is not as easy to account for such exposure-mediator interactions in the context of a linear SEM (VanderWeele, 2016).

Effect interpretation, particularly mediation contrasts: The last 20 years of literature on methods for mediation analysis has focused extensively on precisely how these effects can be defined and interpreted. Much of the work in this area demonstrates that it is not very straightforward nor intuitive. However, mediation effects from SEMs have not received the same degree of attention with respect to the definition of the effect and its interpretation. Thus, there is sufficient reason for concern about precisely what these effects quantify (Robins and Greenland, 1992).

Mediator-outcome confounders, mediator-outcome confounders affected by exposure: Estimating mediated effects without bias requires adjusting for potential confounders of the mediator outcome association. However, sometimes, these confounders may also be affected by the exposure of interest. When these confounders are present (see Margin Figure), SEMs will fail to quantify an unbiased estimate the mediated effect of the exposure (VanderWeele et al., 2014).

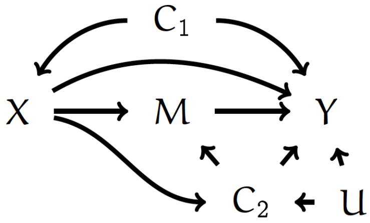


Figure 2: Mediator outcome confounding affected by the exposure.

Finally, it's important to realize that these assumptions are not simply made with respect to one specific exposure-outcome relation of interest, but rather for all variables in the SEM. There is usually a considerable degree of concern when such assumptions are made for a single exposure-outcome relation (VanderWeele, 2012).

4 A Practice Example

To provide a more practical perspective on some of the problems that we can encounter with SEMs, consider the following practice example. To start, we'll simulate some data from a data generating mechanism that looks like Figure 2 in the margin above. These data include one exposure (x), one outcome (y), one mediator of interest (m), one baseline confounder $(c1)$, and one mediator-outcome confounder affected by the exposure $(c2)$.

```
## simulate some data inverse logit
## function
expit <- function(x) {
    1/(1 + exp(-x))
}

# data gen for SEM example
```

```
n <- 2e+06
c1 <- rnorm(n)
x <- rbinom(n, 1, expit(-2 + log(1.5) * c1))
c2 <- rbinom(n, 1, expit(-2 + log(2.5) *
    x))
m <- rbinom(n, 1, expit(-2 + log(1.5) * c1 +
    log(2.5) * x + log(2.5) * c2))
y <- 120 + 1.5 * c1 + 1.5 * x + 2.5 * c2 +
    1.5 * m + rnorm(n)
a <- tibble(c1, c2, x, m, y)

a %>%
    print(n = 3)
```

```
## # A tibble: 2,000,000 x 5
##       c1    c2     x     m     y
##    <dbl> <int> <int> <int> <dbl>
## 1  0.615     0     0     0  119.
## 2 -0.167     0     1     0  120.
## 3 -1.24      0     0     0  117.
## # ... with 1,999,997 more rows
```

We can start by using the `lavaan` package to fit a SEM to these data. We will use this to estimate the effect of the exposure on the outcome that does not occur through the mediator of interest. By "effect", I mean specifically what would happen if everyone in this dataset were exposed (i.e., `x = 1`), versus if everyone in this dataset were unexposed (`x = 0`), all while keeping the mediator value fixed[1]:

```
library(lavaan)

sem_model <- " # direct effect
               y ~ a0*1 + a1*x + a2*m + a3*c2 + a4*c1
               # mediator
               m ~ b0*1 + b1*x + b2*c1 + b3*c2
```

[1] This effect will come up again and again in our short course. It is actually a key parameter of interest in social epidemiology, as it is often used to quantify the extent to which a health disparity is explained by a third variable.

```
               # indirect effect (a*b)
               ab := a1*b1
             "
fit <- sem(sem_model, data = a)
```

We can extract the coefficient for the exposure effect from this model specifically:

```
summary(fit)$pe[2, ]
```

```
##   lhs op rhs label exo     est          se        z pvalue
## 2   y  ~   x    a1   0 1.50041 0.002193008 684.1789      0
```

Here we see that the average difference in the outcomes that would be observed if everyone were exposed versus unexposed if the mediator were held fixed is equivalent to 1.5 units.

Unfortunately, this answer is wrong. The correct answer is actually 1.85 units.

The problem with the answer we obtained from the SEM is that it completely ignores the fact that x also has an effect on $c2$, which has it's own effect on y. In the real world, if we were to set everyone to be exposed and then unexposed, this path from x to $c2$ to y would be engaged, and we would end up with a difference in means that would reflect the contribution of this path.

We will see this problem come up again through several examples in the remainder of this section.

5 The Target Parameter Framework

Mediation analysis methods continue to gain in popularity. These methods are particularly commmon in social epidemiology, as they are regularly used to assess the extent to which an exposure-outcome relation is attributable to a third variable. Commonly used racial/ethnic classifications, measures of socioeconomic position, or characterizations of the neighborhood environment are all associated with several health outcomes throughout the lifecourse (Gee et al., 2012, Williams et al. (2008)). Variables representing these constructs are often taken to designate "fundamental" (Link and Phelan, 1995) or "upstream"

(Gehlert et al., 2008) causes that shape the distribution of more proximal risk factors leading to health disparities.

There is a reasonably strong interest in social epidemiology to attempt to quantify how more proximal risk factors explain social disparities in health.

Examples are numerous and include:

- serum potassium concentrations in the relation between race and diabetes (Chatterjee et al., 2011)
- gestational age at birth and birth weight in the relation between race and fetal death (Lorch et al., 2012)
- cancer stage at diagnosis in the relation between socioeconomic position and mortality (Ibfelt et al., 2013)
- systolic blood pressure in the relation between race and stroke (Howard et al., 2011)
- tobacco consumption in the relation between neighborhood socioeconomic status and lung cancer (Hystad et al., 2013)
- racial disparity in infant mortality explained by breastfeeding prior to discharge from the place of birth (Naimi, 2016)

Most of these studies rely on a procedure for mediation (the difference method) that are based on structural equation modeling concepts, and that yields valid causal inferences under rather strict conditions (Jiang and VanderWeele, 2015, Naimi (2015)).

In this section, we'll introduce key methods for mediation analysis, and discuss how they can be put to use to answer questions in social epidemiology.

In particular, we focus on the use of methods and interpretation of results that are still valid when mediator-outcome confounders are associated with the exposure.

We outline two key methods. The strength of these methods are many, and include the fact that they can accommodate exposure-mediator interactions and mediator-outcome confounders associated with the exposure. These include inverse probability weighting (VanderWeele, 2009) and an outcome modeling approach known as the "structural transformation method" (also known as sequential g-estimation) (Vansteelandt, 2009).

In an associated manuscript, we also provide a similar outline of two "double robust" methods, namely g-estimation of a direct effect structural nested model (Robins, 1999) and targeted minimum loss based estimation (van der

Laan and Gruber, 2012).

6 Motivating Example

To get us started as we learn about key methods to analyze health disparities data, let's begin with a hypothetical example dataset on the role of diet in the racial preterm birth disparity.

These data are available in the online repository: `vegetable_data.csv`

```
a <- read_csv(here("data", "vegetable_data.csv"))
```

These data contain information on 7,653 pregnancies. The outcome of interest is whether the pregnancy ended preterm (i.e., whether the gestational age of the infant was < 37 weeks). The data include information on diet. The primary variable of interest is whether the woman consumed at least 1.5 cups of green leafy vegetables per day per 1000 kcals over the course of her pregnancy.

A second primary variable of interest is whether the woman self-classifies as non-Hispanic Black versus non-Hispanic White.[2]

We also collected information on several confounders of potential interest. These include:

- pre pregnancy BMI
- any prior preterm birth
- pre pregnancy smoking status
- maternal education
- participation in WIC
- maternal age
- overall diet quality

Here are what the data look like:

```
# first five columns, first three rows
a %>%
    select(1:5) %>%
    print(n = 3)
```

[2] We will explore how to handle more complex multi-category exposures, such as a more complete measure of race/ethnicity, in the in person session.

```
## # A tibble: 7,653 x 5
##      ID   ptb green_veg overall_diet maternal_age
##   <dbl> <dbl>     <dbl>        <dbl>        <dbl>
## 1     1     1         1        14.1          32.1
## 2     2     0         1         5.49         26.5
## 3     3     0         1         9.05         25.7
## # ... with 7,650 more rows
```

```
# names of all the variables in the
# dataset
names(a)
```

```
##  [1] "ID"             "ptb"            "green_veg"       "overall_diet"
##  [5] "maternal_age"   "bmi"            "prepreg_smoking" "prior_ptb"
##  [9] "high_school"    "wic"            "race"
```

> **Self Study**:
> Construct a project folder based on the material in the previous session. Name the main project folder as `preterm_disparity`, and include the following sub-folders: `data`, `code`, `figures`, `misc`, `sandbox`, `reports`. With this folder structure, create an RStudio project, and conduct a very brief preliminary exploratory data analysis of the `vegetable_data.csv` file, focusing on the outcome (`ptb`), race/ethnicity (`race`), the mediator (`green_veg`), and BMI. Use a table for the outcome, race/ethnicity, and the mediator. Use a histogram for BMI. Save the EDA report in the `reports` folder, and the BMI histogram in the `figures` folder.
>
> Next, share your entire project folder with a classmate in the short course. Is your classmate able to reproduce your report exactly?

Once we have these data loaded, we are ready to start writing code we need for our analysis. But before we can proceed, we need to clearly articulate what we want to quantify, precisely. There are many possible quantities that we can pursue. Here, we will explore methods that can be used to answer questions such as:

How much of the racial/ethnic disparity in preterm birth is explained by the consumption of green leafy vegetables?

We can use counterfactual disparity measures, introduced in 2016 (Naimi, 2016) to quantify a parameter that correponds to this question.

7 Traditional Mediation Analysis

Figures 1a and 1b are causal diagrams (Pearl, 1995), where X represents an exposure, M a mediator, and Y an outcome of interest. We represent exposure-outcome confounders as C_{XY} and mediator-outcome confounders as C_{MY}.

Traditional mediation analysis methods typically answer questions about extent to which the effect of a particular exposure is mediated or transmitted through a second, mediating variable. These questions are usually framed in terms of the "direct" and/or "indirect" effect of an exposure on an outcome.

Four assumptions may be required to estimate direct and indirect causal exposure effects (VanderWeele and Vansteelandt, 2009):

1) No uncontrolled exposure-outcome confounding
2) No uncontrolled mediator-outcome confounding
3) No mediator-outcome confounders affected by the exposure
4) No exposure-mediator interaction on the scale of interest

Assumptions 1 to 3 are encoded in Figure 1a, which shows that adjusting for C_{XY} and C_{MY} leaves no open back-door path from X to Y (Assumptions 1 & 2), and where there is no arrow from X to C_{MY} (Assumption 3). If the stable unit treatment value assumption (Rubin, 1986) is met for both the exposure and the mediator, there is no selection or information bias, and assumptions 1 to 4 hold, estimating the direct and indirect exposure effects can be done with a wide variety of methods, including simple regression, as well as SEMs.

If assumptions 3 and 4 are violated, then simpler methods such as the "difference" method the "generalized product" method, simple regression approaches, or standard structural equation models, cannot be used.

Many exposure variables of common interest in social epidemiology will often be associated with mediator-outcome confounders.

Let's consider a few examples from the list we introduced above. In each case, it is easy to argue that variables confounding the mediator-outcome relation are also associated with the exposure:

- serum potassium concentrations in the relation between race and diabetes (Chatterjee et al., 2011)

In this case, the analysts assumed that relevant confounders for the relation

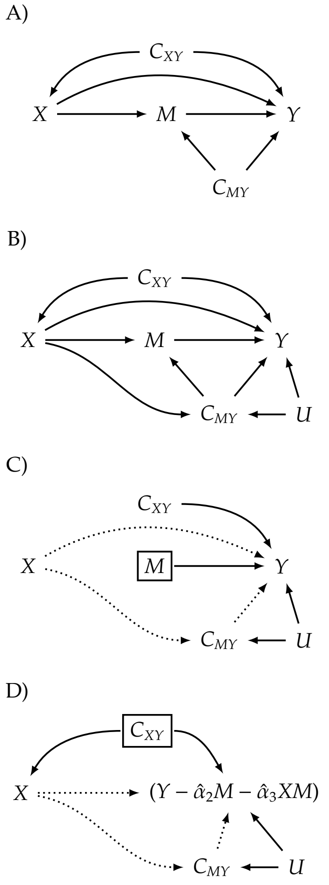


Figure 3: Mediation diagrams: (a) standard setting with an exposure X, mediator M, outcome Y, measured confounders of the exposure-outcome C_{XY} and mediator-outcome C_{MY}. This diagram assumes the exposure X does not affect confounders of the mediator-outcome relation C_{MY}; (b) setting in which the exposure affects confounders of the mediator-outcome relation. In this diagram, conditioning on C_{MY} using standard regression methods will (i) induce collider bias between X and U, and block part of the effect of interest; (c) scenario encountered after applying inverse probability weights for the mediator and exposure; (d) scenario encountered after a structural transformation by subtracting the effect of the mediator from the outcome.

between serum potassium concentrations and diabetes include leisure-time physical activity, hypertension, income, and education. However, each of these is arguably associated with race/ethnicity (Naimi et al., 2011).

- systolic blood pressure in the relation between race and stroke (Howard et al., 2011)

In this case, the analysts adjusted for antihypertensive medication use, diabetes mellitus, atrial fibrillation, heart disease, and cigarette smoking. Again, it is easy to argue that each of these is (sometimes strongly) associated with race and ethnicity.

- tobacco consumption in the relation between neighborhood socioeconomic status and lung cancer (Hystad et al., 2013)

Here, the analysts adjusted for individual level SES, smoking status, diet, physical activity, alcohol consumption, occupational exposures to cancer causing substances, and exposure to environmental hazards such as ambient air pollution. Once again, each of these confounders may be strongly associated with neighborhood-level SES, which is the primary exposure of interest.

This is a common theme in such analyses in social epidemiology, and there are two takeaways: The first is that improperly accounting these mediator-outcome confounders can lead to biased results. This bias can be very large. For example, we have shown that when improper methods are used, the key associations of interest can be reversed (other side of the null, Naimi (2016)). The second takeaway is that we need analytic tools that will allow us to quantify the effects of interest without incurring such biases.

8 Counterfactual Disparity Measures: Definition

With the methods we will be using, there is one key assumption required for the valid estiamtion of counterfactual disparity measures:

- No uncontrolled mediator-outcome confounding

Fewer assumptions are required to estimate counterfactual disparity measures (CDM) because of the information these quantities provide (Vanderweele and Robinson, 2014).

In our motivating example, our “exposure” X is an indicator of maternal race (1 if non-Hispanic Black, 0 otherwise), M denotes whether a woman consumed at least 1.5 cups of green leafy vegetables per day per 1000 kcals (1 if no, 0 if yes)[3], and Y denotes whether the pregnancy ended preterm or not (1 if yes, 0 if no).

We can define a counterfactual disparity measure of this association on the difference scale as:

$$CDM(m=0) \equiv E\big[Y(m=0) \mid X=1\big] - E\big[Y(m=0) \mid X=0\big]$$

where $Y(m = 0)$ is the potential outcome that would be observed if, possibly contrary to fact, a woman consumed at least 1.5 cups of green leafy vegetables per day (Rubin, 2005). We can also define the CDM on the risk ratio scale, such as:

$$CDM(m=0) \equiv \frac{E\big[Y(m=0) \mid X=1\big]}{E\big[Y(m=0) \mid X=0\big]}$$

In both equations (8) and (8), $CDM(m = 0)$ represents the magnitude of the racial disparity in preterm birth that would be observed if all women consumed 1.5 cups per day.

Next, we’ll use the `vegetable_data.csv` dataset to quantify the above CDM using two analytic techniques.

9 Inverse Probability Weighted Marginal Structural Models

Inverse probability weighted marginal structural models (VanderWeele, 2009) can be used to estimate counterfactual disparity measures. The approach proceeds by modeling the mediator and (possibly) exposure, generating inverse probability weights from these models, and fitting a weighted regression model of the outcome against the exposure, the mediator, and their interaction. Mathematically, these weights can be obtained as:

$$sw = \frac{f(X)}{f(X \mid C_{XY})} \times \frac{f(M)}{f(M \mid X, C_{XY}, C_{MY})}$$

where $f(X)$ and $f(M)$ are the probability density functions for X and M, respectively. As explained elsewhere (Hernán et al., 2000, VanderWeele

[3] We’ll soon see the importance of appropriate variable coding for the mediator and exposure.

(2009)) correct specification of the models for the denominator of sw yields an unbiased estimate of the parameter of interest. To correctly estimate the CDM, this approach relies on the assumption that the mediator model is correctly specified as a function of all mediator-outcome confounders.

To do this in R is relatively straightfoward with the standard GLM function:

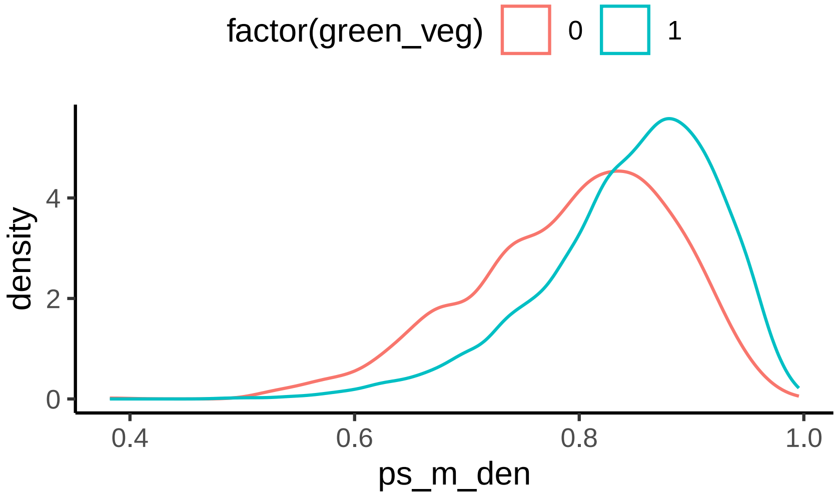
```
a <- read_csv(here("data", "vegetable_data.csv"))

# code snippet 1 mediator models

a$ps_m_num <- glm(green_veg ~ 1, data = a,
    family = binomial("logit"))$fitted.values

a$ps_m_den <- glm(green_veg ~ overall_diet +
    maternal_age + bmi + prepreg_smoking +
    prior_ptb + high_school + wic + race,
    data = a, family = binomial("logit"))$fitted.values

ggplot(a) + geom_density(aes(x = ps_m_den,
    group = factor(green_veg), color = factor(green_veg)))
```



```
a <- a %>%
    mutate(sw_m = green_veg * (ps_m_num/ps_m_den) +
        (1 - green_veg) * ((1 - ps_m_num)/(1 -
             ps_m_den)))

summary(a$sw_m)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2639  0.9143  0.9708  0.9997  1.0510  6.1097
```

```
a %>%
    select(ID, green_veg, ps_m_num, ps_m_den,
        sw_m) %>%
    print(n = 5)
```

```
## # A tibble: 7,653 x 5
##       ID green_veg ps_m_num ps_m_den  sw_m
##    <dbl>     <dbl>    <dbl>    <dbl> <dbl>
## 1      1         1    0.837    0.964 0.868
## 2      2         1    0.837    0.871 0.961
## 3      3         1    0.837    0.907 0.923
## 4      4         1    0.837    0.885 0.945
## 5      5         1    0.837    0.858 0.976
## # ... with 7,648 more rows
```

Now that we've constructed the stabilized weights that we need, the next step is to fit the model we need to estimate the counterfactual disparity measure. For the IP weighting approach, this is the step where we decide if we want to quantify this measure on the difference or ratio scales (or both):

```
library(sandwich)
library(lmtest)

# Overall Risk Difference
tot_rd <- glm(ptb ~ race, data = a, family = binomial("identity"))

summary(tot_rd)$coefficients[2, ]
```

```
##     Estimate   Std. Error      z value     Pr(>|z|)
## 1.335140e-01 7.355454e-03 1.815169e+01 1.245010e-73
```

```
tot_rr <- glm(ptb ~ race, data = a, family = poisson("log"))

tot_vcov_rr <- vcovCL(tot_rr, cluster = a$ID,
    type = "HC3", sandwich = T)
coeftest(tot_rr, vcov = tot_vcov_rr, type = "HC3")
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -3.142108   0.070980 -44.268 < 2.2e-16 ***
## race         1.408838   0.080441  17.514 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# CDM Risk Difference
cdm_rd <- glm(ptb ~ race + green_veg + race *
    green_veg, weights = sw_m, data = a,
    family = binomial("identity"))

cdm_vcov_rd <- vcovCL(cdm_rd, cluster = a$ID,
    type = "HC3", sandwich = T)
coeftest(cdm_rd, vcov = cdm_vcov_rd, type = "HC3")
```

```
##
## z test of coefficients:
##
##                 Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    0.0177513  0.0048621  3.6509 0.0002613 ***
## race           0.0554004  0.0148845  3.7220 0.0001976 ***
## green_veg      0.0295734  0.0060237  4.9095 9.131e-07 ***
## race:green_veg 0.0891131  0.0170083  5.2394 1.611e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(cdm_rd, vcov = cdm_vcov_rd, type = "HC3")
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    0.0177513  0.0048621  3.6509 0.0002613 ***
## race           0.0554004  0.0148845  3.7220 0.0001976 ***
## green_veg      0.0295734  0.0060237  4.9095 9.131e-07 ***
## race:green_veg 0.0891131  0.0170083  5.2394 1.611e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cdm_rr <- glm(ptb ~ race + green_veg + race *
    green_veg, weights = sw_m, data = a,
    family = poisson("log"))

cdm_vcov_rr <- vcovCL(cdm_rr, cluster = a$ID,
    type = "HC3", sandwich = T)
coeftest(cdm_rr, vcov = cdm_vcov_rr, type = "HC3")
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)    -4.031294   0.273902 -14.7180 < 2.2e-16 ***
## race            1.416075   0.334674   4.2312 2.324e-05 ***
## green_veg       0.980571   0.284022   3.4525 0.0005555 ***
## race:green_veg -0.016455   0.345180  -0.0477 0.9619782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now that we've used IP weighting to compute the CDM on the risk difference and risk ratio scales, we must now interpret it.

First, we'll note that the overall disparity in preterm birth between non-Hispanic Black and non-Hispanic White women was 13 per 100 pregnancies (95% CIs: 12, 15). That is, for every 100 pregnancies, non-Hispanic Black women experienced 13 more preterm births than non-Hispanic White women.

We can now compare this to the counterfactual disparity measure. On the difference scale, the counterfactual disparity measure yielded an estimate of 6 more preterm births per 100 pregnancies (95% CIs: 3, 8). This suggests that, overall, getting all women to consume at least 1.5 cups of green leafy vegetables per 1000 kcals per day would reduce the disparity by approximately 8 preterm births for every 100 pregnancies.

Note that, while my preference is to present these results on the absolute (i.e., risk difference) scales, it is possible to estimate risk ratios as well. With IP weighting, this simply requires changing the link functions in the distribution specification, as we did in the above code. On the ratio scale, the total racial disparity in preterm birth is 4.09 (95% CIs: 3.49, 4.79). After accounting for vegetable consumption, we obtain a counterfactual disparity measure (as a risk ratio) of 4.12 (95% CIs: 2.14, 7.94).

10 The Structural Transformation Method

Inverse probability weighting can be used to calculate counterfactual disparity measures. They are based on modeling the "mediator" to construct weights, and then estimating the CDM by using a weighted regression model for the outcome against the "exposure" and "mediator" (and their interaction). However, there is another approach that we can use to calculate the counterfactual disparity measure that does not require constructing weights. This appraoch has been referred to as the "sequential g estimation" method in the biostatistics literature (Goetgeluk et al., 2008), which is unfortunate, because it is quite distinct from the estimation approach referred to as "g estimation" (see Naimi et al., 2017). For this reason, I often refer to this approach as the *structural transformation* method.

This method starts with a full model that regresses the outcome against the exposure, the mediator, the exposure-mediator interaction, as well as all the confounders needed to adjust for the mediator outcome relation. To estimate the counterfactual disparity measure on the difference scale, we start with generalized linear models with a gaussian distribution and an identity link function, or, equivalently, the `lm` function to implement ordinary least squares:

```
struct_trans1 <- glm(ptb ~ race + green_veg +
    race * green_veg + overall_diet + maternal_age +
    bmi + prepreg_smoking + prior_ptb + high_school +
    wic + race, data = a, family = gaussian("identity"))
struct_trans1 <- lm(ptb ~ race + green_veg +
    race * green_veg + overall_diet + maternal_age +
    bmi + prepreg_smoking + prior_ptb + high_school +
    wic + race, data = a)
```

From this model, we then extract the coefficients for the mediator, and the mediator-exposure interaction. All we need at this step are the point estimates. We can discard the standard errors and all other statistics:

```
med_estimates <- summary(struct_trans1)$coefficients[c("green_veg",
    "race:green_veg"), "Estimate"]

med_estimates
```

```
##      green_veg race:green_veg
##     0.01947926     0.09932870
```

We now create a transformed outcome using these point estimate as follows:

```
a <- a %>%
    mutate(ptb_tilde = ptb - med_estimates[1] *
        green_veg - med_estimates[2] * green_veg *
        race)
```

At this point, let's take a moment to understand what is going on. First, note that the `med_estimates` object contains estimates of the effect of NOT consuming at least 1.5 cups of green leafy vegetables per day per 1000 kcals. Thus, when we create the `ptb_tilde` variable, we are effectively constructing the outcome that would be observed if we remove the contribution that NOT eating green leafy vegetables has to the overall risk.

Because of the way we coded green leafy vegetable consumption, we can interpret the average of `ptb_tilde` as the preterm birth rate that would be observed if everyone consumed at least 1.5 cups of green leafy vegetables. We can then regress this outcome against race to evaluate the disparity that we'd see if everyone consumed at least 1.5 cups of green leafy vegetables:

```
summary(lm(ptb ~ race, data = a))$coefficients
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 0.04319163 0.004412115  9.789326 1.695449e-22
## race        0.13351396 0.006766342 19.732075 1.412808e-84
```

```
struct_trans2 <- lm(ptb_tilde ~ race, data = a)

summary(struct_trans2)$coefficients
```

```
##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 0.02734786 0.004393504 6.224611 5.082076e-10
## race        0.04614013 0.006737800 6.847952 8.071634e-12
```

One challenge with using the structural transformation method is that the standard errors contained in the model output are incorrect. To address this, it is best to use the bootstrap (Efron and Tibshirani, 1993). Here, we present the simplest bootstrap we can use: the normal-interval bootstrap[4]:

```
cdm_boot <- NULL

for (i in 1:500) {
    # set the seed
    set.seed(i)

    # resample the data
    index <- sample(1:nrow(a), nrow(a), replace = T)
```

[4] There are several versions of the bootstrap, and most can be implemented using the `boot` package in R. However, here, I'm showing you how to implement a simple normal-interval bootstrap without using the `boot` package.

```
    boot_dat <- a[index, ]

    # estimate the models
    struct_trans1_boot <- lm(ptb ~ race +
        green_veg + race * green_veg + overall_diet +
        maternal_age + bmi + prepreg_smoking +
        prior_ptb + high_school + wic + race,
        data = boot_dat)
    med_estimates_boot <- summary(struct_trans1)$coefficients[c("green_veg",
        "race:green_veg"), "Estimate"]
    boot_dat <- boot_dat %>%
        mutate(ptb_tilde = ptb - med_estimates_boot[1] *
            green_veg - med_estimates_boot[2] *
            green_veg * race)
    struct_trans2_boot <- lm(ptb_tilde ~
        race, data = boot_dat)

    cdm_boot <- rbind(cdm_boot, coef(struct_trans2_boot)[2])

}

head(cdm_boot)
```

```
##            race
## [1,] 0.04359148
## [2,] 0.03991927
## [3,] 0.04933918
## [4,] 0.04833409
## [5,] 0.06256254
## [6,] 0.04077166
```

We can obtain a standard error for the counterfactual disparity measure estimate from the `struct_trans2` object above by computing the standard deviation of the bootstrap estimates in the `cdm_boot` object. We can then use this standard error estimate in the standard Wald equation:

```
sd(cdm_boot)
```

```
## [1] 0.007618946
```

```
summary(struct_trans2)$coefficients[2, 1] *
    100
```

```
## [1] 4.614013
```

```
(summary(struct_trans2)$coefficients[2, 1] -
    1.96 * sd(cdm_boot)) * 100
```

```
## [1] 3.1207
```

```
(summary(struct_trans2)$coefficients[2, 1] +
    1.96 * sd(cdm_boot)) * 100
```

```
## [1] 6.107327
```

We can also compute risk ratios using the structural transformation approach. To do this, we need to modify the above procedure by placing key elements on the log scale. For example:

```
struct_trans1 <- glm(ptb ~ race + green_veg +
    race * green_veg + overall_diet + maternal_age +
    bmi + prepreg_smoking + prior_ptb + high_school +
    wic + race, data = a, family = poisson("log"))

med_estimates <- summary(struct_trans1)$coefficients[c("green_veg",
    "race:green_veg"), "Estimate"]

a <- a %>%
    mutate(ptb_tilde = ptb * exp(-med_estimates[1] *
        green_veg - med_estimates[2] * green_veg *
        race))
```

```
struct_trans2 <- glm(ptb_tilde ~ race, data = a,
    family = quasipoisson("log"))

round(exp(summary(tot_rr)$coefficients[2,
    1]), 2)
```

```
## [1] 4.09
```

```
round(exp(summary(struct_trans2)$coefficients[2,
    1]), 2)
```

```
## [1] 3.18
```

Confidence intervals for this risk ratio should also be obtained using the bootstrap.

References

Anna Arlinghaus, David A Lombardi, Joanna L Willetts, Simon Folkard, and David C Christiani. A structural equation modeling approach to fatigue-related risk factors for occupational injury. *Am J Epidemiol*, 176(7):597–607, 2012.

Ranee Chatterjee, Hsin-Chieh Yeh, Tariq Shafi, Cheryl Anderson, James S Pankow, Edgar R Miller, David Levine, Elizabeth Selvin, and Frederick L Brancati. Serum potassium and the racial disparity in diabetes risk: the atherosclerosis risk in communities (aric) study. *Am J Clin Nutr*, 93(5): 1087–1091, 2011.

Bradley Efron and Robert Tibshirani. *Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993.

Gilbert C Gee, Katrina M Walsemann, and Elizabeth Brondolo. A life course perspective on how racism may be related to health inequities. *Am J Public Health*, 102(5):967–974, May 2012. DOI: 10.2105/AJPH.2012.300666.

Sarah Gehlert, Dana Sohmer, Tina Sacks, Charles Mininger, Martha McClintock, and Olufunmilayo Olopade. Targeting health disparities: a model linking upstream determinants to downstream interventions. *Health Aff (Millwood)*, 27(2):339–349, Mar-Apr 2008. ISSN 1544-5208 (Electronic); 0278-2715 (Linking). DOI: 10.1377/hlthaff.27.2.339.

Sylvie Goetgeluk, Stijn Vansteelandt, and Els Goetghebeur. Estimation of controlled direct effects. *J R Stat Soc Series B Stat Methodol*, 70(5):1049–1066, 2008.

Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Stat Sci*, 14(1):29–46, 1999.

Micah L Hartwell, Jam Khojasteh, Marianna S Wetherill, Julie M Croff, and Denna Wheeler. Using structural equation modeling to examine the influence of social, behavioral, and nutritional variables on health outcomes based on nhanes data: Addressing complex design, nonnormally distributed variables, and missing information. *Current Developments in Nutrition*, 3(5):nzz010, 11/30/2022 2019.

Miguel Ángel Hernán, Babette Brumback, and James M. Robins. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. *Epidemiol*, 11(5):561–570, 2000.

George Howard, Mary Cushman, Brett M Kissela, Dawn O Kleindorfer, Leslie A McClure, Monika M Safford, J David Rhodes, Elsayed Z Soliman, Claudia S Moy, Suzanne E Judd, and Virginia J Howard. Traditional risk factors as the underlying cause of racial disparities in stroke: lessons from the half-full (empty?) glass. *Stroke*, 42(12):3369–3375, Dec 2011. DOI: 10.1161/STROKEAHA.111.625277.

Perry Hystad, Richard M Carpiano, Paul A Demers, Kenneth C Johnson, and Michael Brauer. Neighbourhood socioeconomic status and individual lung cancer risk: evaluating long-term exposure measures and mediating mechanisms. *Soc Sci Med*, 97:95–103, Nov 2013. DOI: 10.1016/j.socscimed.2013.08.005.

E H Ibfelt, S K Kjar, C Hogdall, M Steding-Jessen, T K Kjar, M Osler, C Johansen, K Frederiksen, and S O Dalton. ’socioeconomic position and survival after cervical cancer: influence of cancer stage, comorbidity and smoking among danish women diagnosed between 2005 and 2010. *Br J Cancer*, 109(9): 2489–2495, 2013.

Z Jiang and TJ VanderWeele. When is the difference method conservative for assessing mediation? *Am J Epidemiol*, 182(2):105–8, 2015.

B G Link and J Phelan. Social conditions as fundamental causes of disease. *J Health Soc Behav*, 35((Extra Issue)):80–94, 1995.

Scott A. Lorch, Charlan D. Kroelinger, Corinne Ahlberg, and Wanda D. Barfield. Factors that mediate racial/ethnic disparities in us fetal death rates. *American Journal of Public Health*, 102(10):1902–1910, 2012.

Sydney A. Martinez, Laura A. Beebe, David M. Thompson, Theodore L. Wagener, Deirdra R. Terrell, and Janis E. Campbell. A structural equation modeling approach to understanding pathways that connect socioeconomic status and smoking. *PLOS ONE*, 13(2):e0192451–, 2018.

Lawrence C. McCandless and Paul Gustafson. A comparison of bayesian and monte carlo sensitivity analysis for unmeasured confounding. *Statistics in Medicine*, 36(18):2887–2901, 2017.

A I Naimi. The Counterfactual Implications of Fundamental Cause Theory. *Curr Epidemiol Reports*, In Press, 2016.

Ashley I Naimi. Invited commentary: Boundless science-putting natural direct and indirect effects in a clearer empirical context. *Am J Epidemiol*, 182(2): 109–114, 2015.

Ashley I Naimi, Jay S Kaufman, Chanelle J Howe, and Whitney R Robinson. Mediation considerations: serum potassium and the racial disparity in diabetes risk. *Am J Clin Nutr*, 94(2):614–616, 2011.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An Introduction to G Methods. *Int J Epidemiol*, 46(2):756–62, 2017.

J Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In C Glymour and GF Cooper, editors, *Computation, Causation, and Discovery*, pages 349–405. AAAI Press/The MIT Press, Menlo Park, CA / Cambridge, MA, 1999.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiol*, 3(2):143–155, 1992.

Donald B Rubin. Comment: Which ifs have causal answers. *J Am Stat Assoc*, 81(396):961–962, 1986.

Donald B Rubin. Causal inference using potential outcomes. *J Am Stat Assoc*, 100(469):322–331, 2005.

G. A. F. Seber and C. J. Wild. *Nonlinear regression*. Wiley, New York, 1989.

Kimberly Sidora-Arcoleo, Jonathan M Feldman, Denise Serebrisky, and Amanda Spray. A multi-factorial model for examining racial and ethnic disparities in acute asthma visits by children. *Ann Behav Med*, 43(1):15–28, 2012.

Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat*, 8(1): Article 9, 2012.

T. J. VanderWeele. Marginal structural models for direct and indirect effects (erratum in *Epidemiology* 2009; 20(4):629). *Epidemiol*, 20(1):18–26, 2009.

Tyler J VanderWeele. Invited commentary: structural equation models and epidemiologic analysis. *Am J Epidemiol*, 176(7):608–612, Oct 2012.

Tyler J. VanderWeele. Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1):17–32, 2016.

Tyler J Vanderweele and Whitney R Robinson. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiol*, 25(4):473–84, 2014.

Tyler J VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*, 2(4):457–468, 2009.

Tyler J. VanderWeele, Stijn Vansteelandt, and James M. Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiol*, 25(2):300–306, 2014.

Stijn Vansteelandt. Estimating direct effects in cohort and case–control studies [erratum in: *Epidemiol* 2010:21(2)]. *Epidemiol*, 20(6):851–860, 2009.

David R Williams, Manuela V Costa, Adebola O Odunlami, and Selina A Mohammed. Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities. *J Public Health Manag Pract*, 14 Suppl:S8–17, Nov 2008. DOI: 10.1097/01.PHH.0000338382.36695.42.