The parametric g-formula

Lab 3

EPI 207

Harvard University

Outline

Today we will:

1. Review the derivation of the g-formula density, the assumptions under which it recovers the counterfactual density we're interested in, and how we've estimated it to date.

2. Introduce the parametric g-formula as an alternative way to estimate the g-formula density.

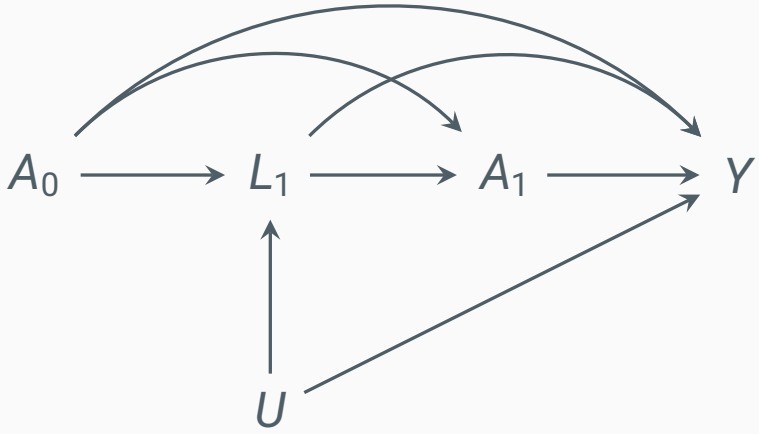
3. Work through a step by step example of the parametric g-formula using R.

4. Introduce the `gfoRmula` package in R, which makes estimating the parametric g-formula much easier in practice.

Setup

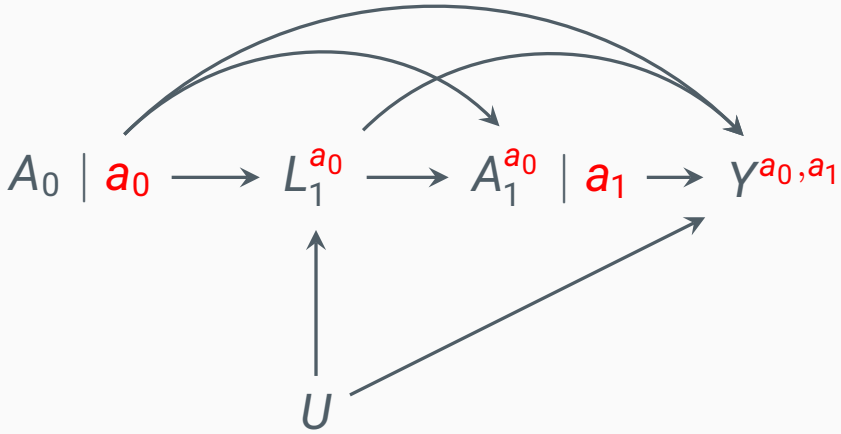
A_0	L_1	A_1	N	$\mathbb{E}[Y \mid A_0, L_1, A_1]$
0	0	0	6000	60
0	0	1	2000	60
0	1	0	2000	210
0	1	1	6000	210
1	0	0	3000	240
1	0	1	1000	240
1	1	0	3000	120
1	1	1	9000	120

Table: Homework 3 Frequency Table



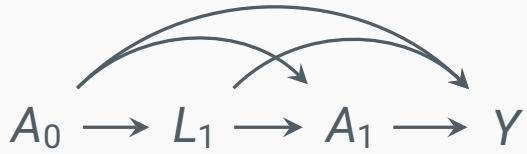
$$Y^{a_0,a_1} \perp\!\!\!\perp A_0$$

$$Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$$



The g-formula density "recipe"

1. Draw the statistical DAG for the observed variables



2. Write down the usual factorization

$$f(a_0, l_1, a_1, y) = f(a_0)f(l_1 \mid a_0)f(a_1 \mid a_0, l_1)f(y \mid a_0, l_1, a_1)$$

3. Leave out terms for the treatment variables given their parents

$$f(l_1 \mid a_0)f(y \mid a_0, l_1, a_1)$$

4. Whenever a treatment variable appears as a parent, set it equal to the value specified by the regime

$$f^{G,g=(a_0,a_1)}(y, l_1) = f(l_1 \mid a_0)f(y \mid a_0, l_1, a_1)$$

Using the g-formula density

Once we have the joint density we can use basic probability theory to get other quantities of interest.

Marginal distribution for Y

$$f^{G,g=(a_0,a_1)}(y) = \sum_{l_1} f^{G,g=(a_0,a_1)}(y, l_1)$$
$$= \sum_{l_1} f(l_1 \mid a_0) f(y \mid a_0, l_1, a_1)$$

Marginal mean of Y

$$\mathbb{E}^{G,g=(a_0,a_1)}[Y] = \sum_{y} y \cdot f^{G,g=(a_0,a_1)}(y)$$
$$= \sum_{y} y \cdot \sum_{l_1} f(l_1 \mid a_0) f(y \mid a_0, l_1, a_1)$$
$$= \sum_{l_1} f(l_1 \mid a_0) \sum_{y} y \cdot f(y \mid a_0, l_1, a_1)$$
$$= \sum_{l_1} \mathbb{E}[Y \mid A_0 = a_0, L_1 = l_1, A_1 = a_1] f(l_1 \mid a_0)$$

The g-formula and causality

Under the following assumptions:

1. Conditional/sequential exchangeability

$$Y^{a_0,a_1} \perp\!\!\!\perp A_0$$

$$Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_1^{a_0}, A_0$$

2. Consistency

$$Y^a = Y \mid A = a$$

3. Positivity

$$0 < \Pr(A_0 = 1) < 1$$

$$0 < \Pr(A_1 = 1 \mid A_0, L_1) < 1$$

The g-formula density for $g = (a_0, a_1)$ is equal to the distribution in the counterfactual world where everyone follows $g = (a_0, a_1)$.

That is:

$$\underbrace{f_{Y^{a_0,a_1}}(y)}_{\substack{\text{Marginal density for } Y^{a_0,a_1} \\ \text{in counterfactual world}}} = \underbrace{f^{G,g=(a_0,a_1)}(y)}_{\substack{\text{G-formula density for Y} \\ \text{in the factual world}}}$$

ONLY if 1, 2, and 3 above are true

Plug-in estimation of the g-formula

Recall that last week we estimated the g-formula by replacing its component expectations and probabilities with their sample equivalents:

$$\widehat{\mathbb{E}}^{G,g=\bar{a}_K}[Y] = \sum_{l_k} \overbrace{\widehat{\mathbb{E}}[Y \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k]}^{\text{plug in sample mean of } Y \text{ within strata of } \bar{A}_k \text{ and } \bar{L}_k} \prod_{k=1}^{K} \underbrace{\widehat{f}(l_k \mid \bar{l}_{k-1}, \bar{a}_k)}_{\substack{\text{plug in sample probability that } L_k = l_k \\ \text{within strata of } \bar{L}_{k-1} \\ \text{and } \bar{A}_k}}$$

This is sometimes called the plug-in estimator of the g-formula.

Parametric g-formula

Instead of using sample means and probabilities, an alternative plug-in estimator, the parametric g-formula, replaces the components of the g-formula with parametric models of the outcome and the covariate histories:

$$\widehat{\mathbb{E}}^{G,g=\bar{a}_k}[Y] = \sum_{l_k} \overbrace{\widehat{\mathbb{E}}[Y \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k]}^{\text{plug in estimates from parametric outcome model}} \prod_{k=1}^{K} \underbrace{\widehat{f}(l_k \mid \bar{l}_{k-1}, \bar{a}_k)}_{\text{plug in estimates from parametric covariate models}}$$

Reminder: what is a parametric model?

A parametric model restricts the joint distribution of the data, often through limiting assumptions on the shape of the mean function and/or the form of the conditional distribution of the outcome.

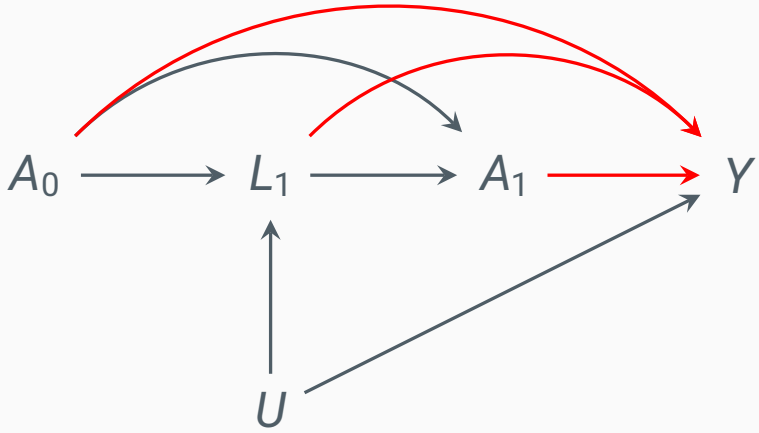
$$\mathbb{E}[Y \mid A_0 = a_0, L_1 = l_1, A_1 = a_1] = \underbrace{\mu(a_0, l_1, a_1)}_{\substack{\text{mean function} \\ \mu}}$$

$$Y \mid A_0, L_1, A_1 \sim \underbrace{\mathcal{P}_\theta}_{\substack{\text{conditional} \\ \text{distribution}}}$$

A common choice is a generalized linear model:

$$\text{generalized linear models} \begin{cases} \mu(a_0, l_1, a_1) = \beta_0 + \beta_1 a_0 + \beta_2 l_1 + \beta_3 a_1 \\ Y \mid A_0, L_1, A_1 \sim \text{Normal}(\mu(a_0, l_1, a_1), \sigma^2) \\ \\ \mu(a_0, l_1, a_1) = \text{logit}^{-1}(\beta_0 + \beta_1 a_0 + \beta_2 l_1 + \beta_3 a_1) \\ Y \mid A_0, L_1, A_1 \sim \text{Bernoulli}(\mu(a_0, l_1, a_1)) \\ \\ \mu(a_0, l_1, a_1) = \log^{-1}(\beta_0 + \beta_1 a_0 + \beta_2 l_1 + \beta_3 a_1) \\ Y \mid A_0, L_1, A_1 \sim \text{Poisson}(\mu(a_0, l_1, a_1)) \end{cases}$$

Outcome model

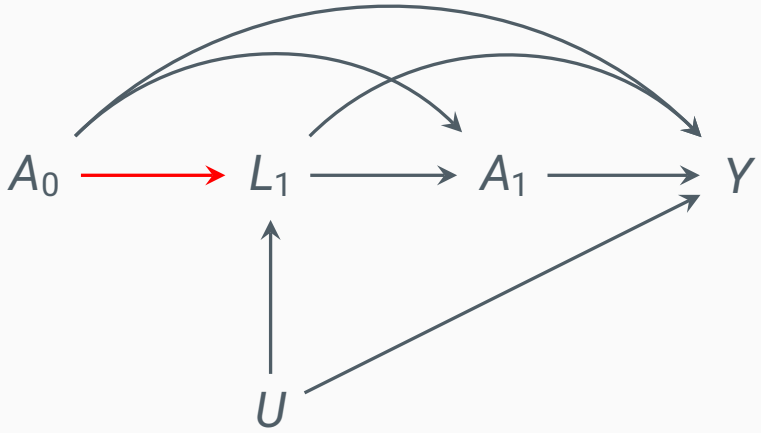


Given Y is continuous a reasonable choice is the following linear regression model:

$$\mathbb{E}[Y \mid A_0, L_1, A_1] = \beta_0 + \beta_1 A_0 + \beta_2 L_1 + \beta_3 A_1$$

$$Y \mid A_0, L_1, A_1 \sim \text{Normal}(\beta_0 + \beta_1 A_0 + \beta_2 L_1 + \beta_3 A_1, \sigma^2)$$

Covariate model(s)



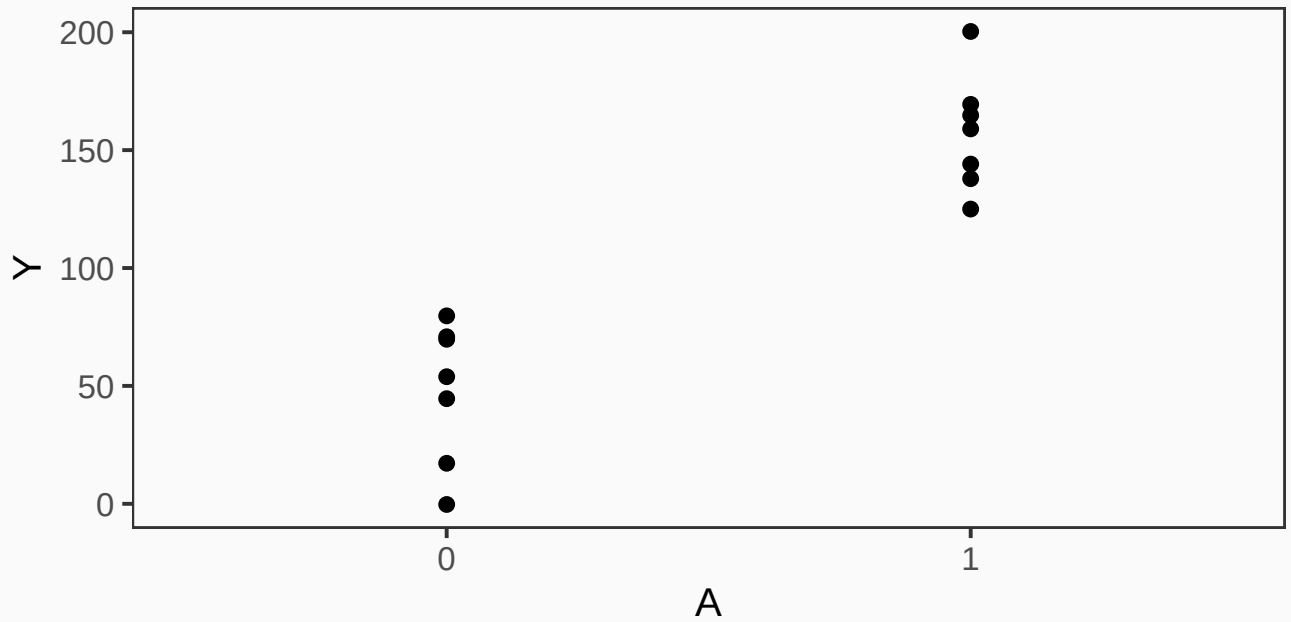
Given L_1 is binary a reasonable choice is the following logistic regression model:

$$\Pr(L_1 \mid A_0) = \text{logit}^{-1}(\beta_0 + \beta_1 A_0)$$
$$L_1 \mid A_0 \sim \text{Bernoulli}(\beta_0 + \beta_1 A_0)$$

Reminder: saturated models

If the number of parameters is equivalent to the number of possible values the conditional mean of the distribution supports then we say our model is "saturated".

$$\mathbb{E}[Y \mid A] = \beta_0 + \beta_1 A$$



Recall that a "saturated" model isn't really a model at all — it imposes no real restrictions on the joint distribution of the data. In this case, the estimates from a saturated model will exactly coincide with the plug-in estimates.

Modeling assumptions and bias-variance trade-off

As we introduce more modeling assumptions, we are reducing the number of parameters that we need to estimate which will tend to lower the variance of our estimates. However, by introducing the possibility that our models are misspecified we are also increasing the potential for bias.

$$\begin{aligned} \text{More assumptions} &= \uparrow \text{Bias} + \downarrow \text{Variance} \\ \text{Less assumptions} &= \downarrow \text{Bias} + \uparrow \text{Variance} \end{aligned}$$

The gamble: if our parametric models are correctly specified (i.e. if our assumptions are correct) we get lower variance and therefore tighter confidence intervals at no cost. However, absent knowledge of the true relationships we can never be certain that our assumptions are correct.

Using simulation to approximate the sum/integral

We now have estimates of the components of the g-formula; however to complete our estimation of the g-formula we still need to take the sum/integral over the distribution of L_k:

For a discrete L_k

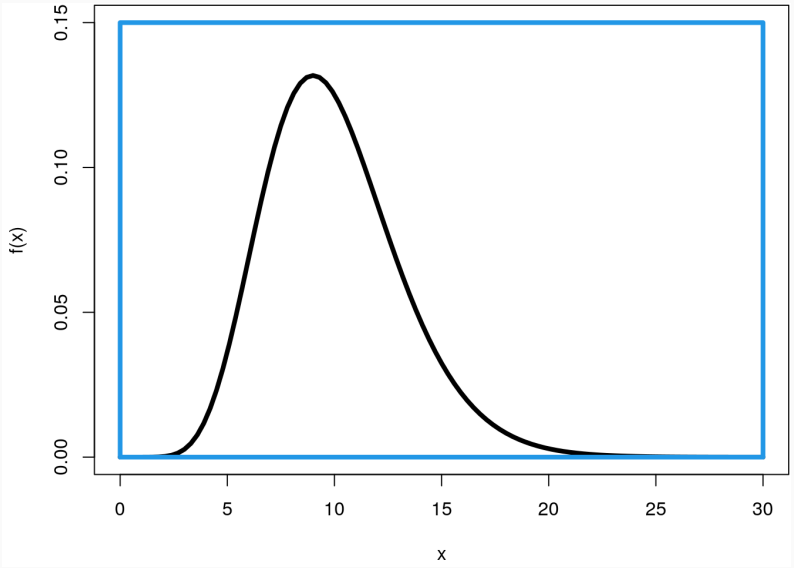
$$\widehat{\mathbb{E}}^{G,g=\bar{a}_k}[Y] = \sum_{l_k} \widehat{\mathbb{E}}[Y \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k] \prod_{k=1}^{K} \widehat{f}(l_k \mid \bar{l}_{k-1}, \bar{a}_k)$$

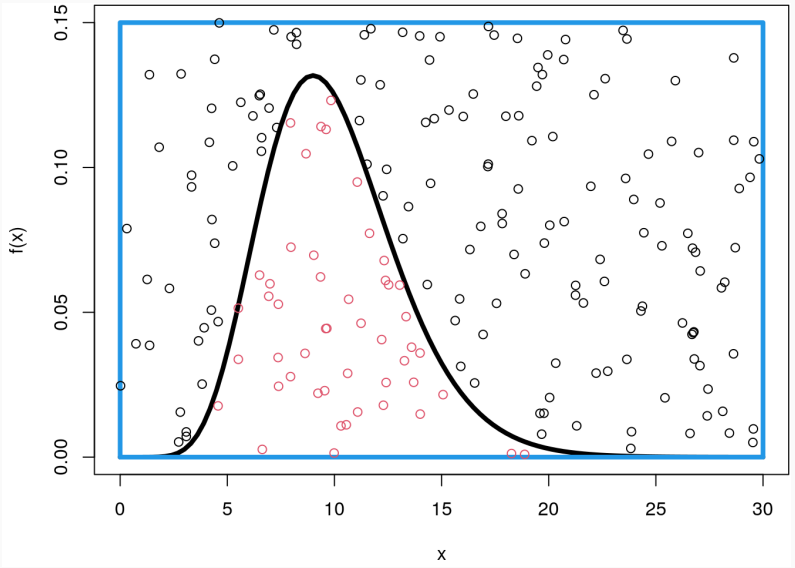
For a continuous L_k

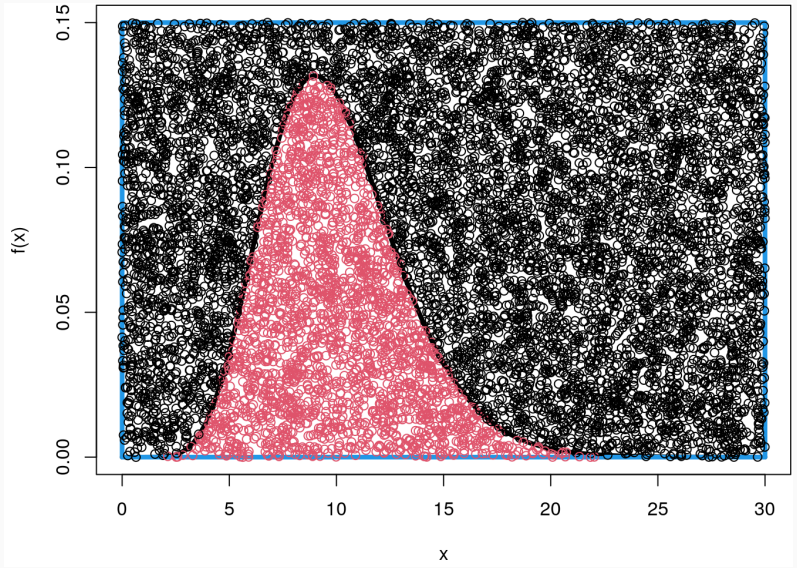
$$\widehat{\mathbb{E}}^{G,g=\bar{a}_k}[Y] = \int_{l_k} \widehat{\mathbb{E}}[Y \mid \bar{A}_k = \bar{a}_k, \bar{L}_k = \bar{l}_k] \prod_{k=1}^{K} \widehat{f}(l_k \mid \bar{l}_{k-1}, \bar{a}_k) dL_k$$

It turns out that we can approximate this sum/integral using Monte Carlo simulation.

Monte Carlo integration







Simulation steps

The steps of the parametric g-formula:

1. Fit parametric models for the outcome and covariate history.
2. To approximate the sum/integral, starting at time 0 draw a random sample of starting values from observed distribution at baseline. For each time point k from 0 to K:
   (a) Intervene to set values of treatment variables to be consistent with regime.
   (b) Use fitted covariate models to estimate mean covariate values at time $k + 1$ based on current values at time k.
   (c) Simulate realizations of the covariates at $k + 1$ using estimated mean and residual variance.
3. Repeat steps (a) through (c) using the simulated realizations of the covariates at k as the entries for the parameteric model in all future $k^* > k$.
4. At the final time also estimate the outcome using the fitted outcome model and all simulated covariate values.
5. Calculate the mean of all outcome estimates to get the final marginal estimate.

Example

Let's use the parametric g-formula to estimate the marginal mean of Y for the strategy $g = (1, 1)$, i.e. we want

$$\mathbb{E}^{G,g=(1,1)}[Y]$$

Begin with the data in long format (i.e. each observation is person-time)

id	time	A	L	Y
1	0	0	NA	NA
1	1	1	1	60
2	0	1	NA	NA
2	1	1	0	210
⋮	⋮	⋮	⋮	⋮
1000	0	0	NA	NA
1000	1	0	1	120

Example

1. Fit parametric models for the outcome and covariate history.

$$\Pr(L_1 \mid A_0) = \text{logit}^{-1}(\beta_0 + \beta_1 A_0)$$
$$L_1 \mid A_0 \sim \text{Bernoulli}(\mu(a_0))$$

$$\mathbb{E}[Y \mid A, L,] = \beta_0 + \beta_1 A_0 + \beta_2 L_1 + \beta_3 A_1$$
$$Y \mid A_0, L_1, A_1 \sim \text{Normal}(\mu(a_0, l_1, a_1), \sigma^2)$$

2. Draw a random sample of starting values from observed distribution at baseline.

id	time	A	L	Y
315	0	0	NA	NA
⋮	⋮	⋮	⋮	⋮

Example

(a) Intervene to set values of treatment variables to be consistent with regime.

id	time	A	L	Y
315	0	1	NA	NA
⋮	⋮	⋮	⋮	⋮

(b) Use fitted covariate models to estimate mean covariate values at time $k+1$ based on current values at time k.

id	time	A	L	Y	$\widehat{\Pr}(L \mid A)$
315	0	1	NA	NA	0.6
⋮	⋮	⋮	⋮	⋮	⋮

(c) Simulate realizations of the covariates at $k+1$ using estimated mean and residual variance.

id	time	A	L	Y	$\widehat{\Pr}(L \mid A)$
315	0	1	NA	NA	0.6
315	1	1	1	NA	NA
⋮	⋮	⋮	⋮	⋮	⋮

Example

3. Repeat steps (a) through (c) using the simulated realizations of the covariates at k as the entries for the parameteric model in all future $k^* > k$.

4. At the final time also estimate the outcome using the fitted outcome model and all simulated covariate values.

id	time	A	L	Y	$\widehat{\Pr}(L \mid A)$	$\widehat{\mathbb{E}}[Y \mid A, L]$
315	0	1	NA	NA	0.6	NA
315	1	1	1	75.5	NA	75.5
⋮	⋮	⋮	⋮	⋮	⋮	

5. Calculate the empirical mean of all outcome estimates to get the final marginal estimate.

Practice

Let's practice calculating the parametric g-formula by hand using R!

```
library(tidyverse)
library(gfoRmula)

# re-create the frequency table for homework 3
hw3_freq <- tribble(
  ~A_0, ~L_1, ~A_1, ~N, ~Y_1,
  0, 0, 0, 6000, 60,
  0, 0, 1, 2000, 60,
  0, 1, 0, 2000, 210,
  0, 1, 1, 6000, 210,
  1, 0, 0, 3000, 240,
  1, 0, 1, 1000, 240,
  1, 1, 0, 3000, 120,
  1, 1, 1, 9000, 120
)
```

Practice

```
# expand the frequency table to 1 row per person
# (first just assign everyone the average Y value)
dat <- uncount(hw3_freq, N) %>%
  # give those rows an id number
  rowid_to_column(var = "id") %>%
  # for each of the variables except for id, split it into two
  # rows, one for each time point
  pivot_longer(-id,
               names_to = c(".value", "time"),
               names_sep = "_"
  ) %>%
  # make sure that time is read as a number
  mutate(time = parse_number(time),
         # add some random error to Y
         # (nrow(.) means a different value for each row of the dataset
         Y = Y + rnorm(nrow(.), 0, 10))
```

Practice

```
> dat
# A tibble: 64,000 x 5
      id  time     A     L     Y
   <int> <dbl> <dbl> <dbl> <dbl>
 1     1     0     0    NA  NA
 2     1     1     0     0  67.7
 3     2     0     0    NA  NA
 4     2     1     0     0  65.0
 5     3     0     0    NA  NA
 6     3     1     0     0  59.3
 7     4     0     0    NA  NA
 8     4     1     0     0  62.7
 9     5     0     0    NA  NA
10     5     1     0     0  44.4
# ... with 63,990 more rows
```

Practice

```r
# create lagged variables
t1 <- dat %>%
  mutate(lag_A = lag(A)) %>%
  filter(time == 1)
```

Practice

Try coding the following on your own:

1. fit models for covariate and outcome given past on `dat`
2. create new data set with space for 10,000 simulated entries
3. fix intervention values in the new data set to be consistent with the regime
4. predict covariate means at time 1
5. simulate covariate values at time 1
6. predict outcome at time 1 based on simulated covariates and fixed treatments
7. take mean across all simulations to get g-formula estimate!

Statistical inference

To get standard errors and confidence intervals for our estimates we can use the bootstrap.

For $b = 1, \ldots, B$:

1. Draw a sample of size n with replacement from the observed data.

2. Calculate the g-formula estimate $\widehat{\psi}_b$ in the simulated bootstrap sample.

Use the resulting estimates $(\widehat{\psi}_1, \widehat{\psi}_2, \ldots, \widehat{\psi}_B)$ to approximate the sampling distribution for $\widehat{\psi}$.

$$SE_{boot}(\widehat{\psi}) = sd(\widehat{\psi}_b)$$

$$95\%CI_{boot}(\widehat{\psi}) = \left(q_{2.5}(\widehat{\psi}_b), q_{97.5}(\widehat{\psi}_b)\right)$$

The `gfoRmula` package in R

Feeling like the parametric g-formula is a lot of work? Fortunately, there's a package for that.

gfoRmula: An R package for estimating effects of general time-varying treatment interventions via the parametric g-formula

Victoria Lin[*]
School of Computer Science
Carnegie Mellon University

Sean McGrath[*]
Harvard T. H. Chan School of Public Health

Zilu Zhang
Harvard Medical School
Dana Farber Cancer Institute

Lucia C. Petito
Feinberg School of Medicine
Northwestern University

Roger W. Logan
Harvard T. H. Chan School of Public Health

Miguel A. Hernán[†]
Harvard T. H. Chan School of Public Health

Jessica G. Young[†]
Harvard Medical School
Harvard Pilgrim Health Care Institute

Practice

Read through the code. Add comments to the lines that start with # to briefly explain what the line below means. You may want to run `?gformula` or read the paper about the package. Then try running the code.

Questions:

1. How many models were fit?

2. How can you tell which of the data in the `sim_data` object was simulated or predicted from a model?

3. Do these results match your answers to the earlier questions, and to Homework 3? Why or why not?