

Discussion. Parametric g-formula: Outcome modeling

Cornell STSCI / INFO / ILRST 3900

Fall 2025

causal3900.github.io

08 Oct 2025

Statistical modeling

Under exchangeability,

$$E(Y^a \mid \vec{L} = \vec{\ell}) = E(Y^a \mid A = a, \vec{L} = \vec{\ell})$$

Under consistency,

$$E(Y^a \mid A = a, \vec{L} = \vec{\ell}) = E(Y \mid A = a, \vec{L} = \vec{\ell})$$

To estimate, we have been taking the subgroup mean

$$\hat{E}(Y \mid A = a, \vec{L} = \vec{\ell}) = \frac{1}{n_{a, \vec{\ell}}} \sum_{i: A_i = a, \vec{L}_i = \vec{\ell}} Y_i$$

When subgroups are empty, we need a model. Example:

$$\hat{E}(Y \mid A = a, \vec{L} = \vec{\ell}) = \hat{\alpha} + A\hat{\beta} + \vec{L}'\hat{\gamma} + A\vec{L}'\hat{\eta}$$

Parametric g-formula: Outcome modeling

1. Learn a model to predict Y given $\{A, \vec{L}\}$
2. For each i , predict
 - ▶ $\{A = 1, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under treatment
 - ▶ $\{A = 0, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under control
3. Take the difference for each unit
4. Average over the units

G-formula: Data example

Estimate a model based on the true data

```
# A tibble: 10 x 4
```

	a	y	sex	race
	<chr>	<lgl>	<chr>	<fct>
1	college	FALSE	Female	Non-Hispanic Non-Black
2	college	FALSE	Female	Non-Hispanic Non-Black
3	college	TRUE	Male	Non-Hispanic Non-Black
4	college	TRUE	Male	Non-Hispanic Non-Black
5	no_college	FALSE	Male	Hispanic
6	no_college	FALSE	Female	Hispanic
7	no_college	TRUE	Male	Hispanic
8	no_college	FALSE	Female	Hispanic
9	no_college	FALSE	Male	Hispanic
10	no_college	FALSE	Female	Hispanic

Predict values - control

Predict the counterfactuals when everybody is in the control group

```
# A tibble: 10 x 3
```

	a	sex	race
	<chr>	<chr>	<fct>
1	no_college	Female	Non-Hispanic Non-Black
2	no_college	Female	Non-Hispanic Non-Black
3	no_college	Male	Non-Hispanic Non-Black
4	no_college	Male	Non-Hispanic Non-Black
5	no_college	Male	Hispanic
6	no_college	Female	Hispanic
7	no_college	Male	Hispanic
8	no_college	Female	Hispanic
9	no_college	Male	Hispanic
10	no_college	Female	Hispanic

Predict values - treatment

Predict the counterfactuals when everybody is in the treatment group

```
# A tibble: 10 x 3
```

	a	sex	race
	<chr>	<chr>	<fct>
1	college	Female	Non-Hispanic Non-Black
2	college	Female	Non-Hispanic Non-Black
3	college	Male	Non-Hispanic Non-Black
4	college	Male	Non-Hispanic Non-Black
5	college	Male	Hispanic
6	college	Female	Hispanic
7	college	Male	Hispanic
8	college	Female	Hispanic
9	college	Male	Hispanic
10	college	Female	Hispanic

1. Learn a model to predict Y given $\{A, \vec{L}\}$

```
fit <- lm(y ~ a + sex + race + mom_educ + dad_educ +  
          log_parent_income +  
          log_parent_wealth +  
          test_percentile,  
          data = d)
```

2. Predict conditional average potential outcomes for every unit

```
conditional_average_outcomes <- d %>%  
  mutate(yhat1 = predict(fit,  
                        newdata = d %>%  
                          mutate(a = "college")),  
  yhat0 = predict(fit,  
                 newdata = d %>%  
                   mutate(a = "no_college")))
```


3. Difference to estimate conditional average effects

```
conditional_average_effects <-  
  conditional_average_outcomes %>%  
  mutate(effect = yhat1 - yhat0)
```

4. Average over units

```
conditional_average_effects %>%  
  select(yhat1, yhat0, effect) %>%  
  summarize_all(.funs = mean)
```

```
# A tibble: 1 x 3  
  yhat1 yhat0 effect  
  <dbl> <dbl> <dbl>  
1 0.427 0.164 0.263
```

Recap. Parametric g-formula: Outcome modeling

1. Learn a model to predict Y given $\{A, \vec{L}\}$
2. For each i , predict
 - ▶ $\{A = 1, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under treatment
 - ▶ $\{A = 0, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under control
3. Take the difference for each unit
4. Average over the units

Extension 1: Conditional average effects

Modify the procedure above to estimate the average effect in subgroups defined by mom's education:

1. those with `sex == Male`
2. those with `sex == Female`

Extension 1: Conditional average effects

Modify the procedure above to estimate the average effect in subgroups defined by mom's education:

1. those with `sex == Male`
2. those with `sex == Female`

One way to code it:

```
conditional_average_effects %>%  
  group_by(sex) %>%  
  select(sex, yhat0, yhat1, effect) %>%  
  summarize_all(.funs = mean)
```

```
# A tibble: 2 x 4  
  sex      yhat0 yhat1 effect  
  <chr>   <dbl> <dbl>   <dbl>  
1 Female 0.125 0.388 0.263  
2 Male   0.203 0.466 0.263
```

Extension 2: Logistic regression

Repeat the steps above with logistic regression

$$\log \left(\frac{\hat{P}(Y | A = a, \vec{L} = \vec{\ell})}{1 - \hat{P}(Y | A = a, \vec{L} = \vec{\ell})} \right) = \hat{\alpha} + A\hat{\beta} + \vec{L}'\hat{\gamma} + A\vec{L}'\hat{\eta}$$

Helpful hints:

- ▶ read about using `glm()` to estimate logistic regression
- ▶ when using `predict()`, search to find out how to predict probabilities

Extension: Logistic regression

Fit a model

```
fit <- glm(y ~ a*(sex + race + mom_educ + dad_educ +  
              log_parent_income +  
              log_parent_wealth +  
              test_percentile),  
          data = d,  
          family = binomial)
```

Extension: Logistic regression

Predict and summarize to estimate the average effect

```
d %>%  
  mutate(yhat1 = predict(fit,  
                          newdata = d %>%  
                            mutate(a = "college"),  
                          type = "response"),  
  yhat0 = predict(fit,  
                  newdata = d %>%  
                    mutate(a = "no_college"),  
                  type = "response"),  
  effect = yhat1 - yhat0) %>%  
  select(yhat1,yhat0,effect) %>%  
  summarize_all(.funs = mean)
```

```
# A tibble: 1 x 3  
  yhat1 yhat0 effect  
  <dbl> <dbl> <dbl>  
1 0.406 0.165 0.241
```


Recap. Parametric g-formula: Outcome modeling

1. Learn a model to predict Y given $\{A, \vec{L}\}$
2. For each i , predict
 - ▶ $\{A = 1, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under treatment
 - ▶ $\{A = 0, \vec{L} = \vec{\ell}_i\}$, the conditional average outcome under control
3. Take the difference for each unit
4. Average over the units