

# Discussion. Parametric g-formula: Outcome modeling

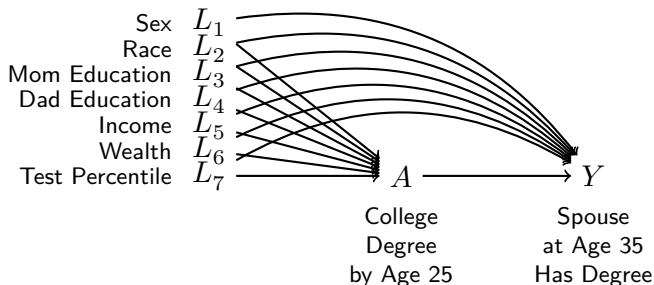
Cornell STSCI / INFO / ILRST 3900

Fall 2023

[causal3900.github.io](https://causal3900.github.io)

27 Sep 2023

Where lecture ended



**100%** of the sample

is in a subgroup with either 0 treated or 0 untreated units

# Setup

Follow the instructions on Ed to download the data!

# Statistical modeling

Under exchangeability,

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

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

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,  
         family = binomial)
```

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 0.404 0.163 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

## 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`

If you finish, choose a subgroup of interest to you and summarize.

## 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`

If you finish, choose a subgroup of interest to you and summarize.

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 0.129 0.348 0.219  
2 Male   0.196 0.460 0.263
```

## Extension 2: Logistic regression

In groups: 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