Discussion. Parametric g-formula: Outcome modeling

Cornell STSCI / INFO / ILRST 3900
Fall 2025
causal3900.github.io

08 Oct 2025

Statistical modeling

Under exchangeability,

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

Under consistency,

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

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:

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

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 $\{A, \vec{L}\}$

```r
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

```r
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: IPW Estimator

We can also estimate the causal effect using an IPW estimator

$$\pi_i = P(A_i = 1 \mid L = \ell_i)$$

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

```
# turn outcome into 0/1
d <- d %>%
  mutate(a_binary = a == "college")

# Estimate propensity score using logistic regression
propensity_model <- glm(a_binary ~ sex + race +
                          mom_educ + dad_educ +
                          log_parent_income +
                          log_parent_wealth +
                          test_percentile,
                        family = binomial, data = d)
```

IPW: Estimate the causal effect

Use the estimated propensty scores to calculate the ACE

$$\hat{E}(Y^1) = \frac{1}{n}\sum_i \frac{A_i Y_i}{\hat{\pi}_i}$$

$$\hat{E}(Y^0) = \frac{1}{n}\sum_i \frac{(1-A_i) Y_i}{1-\hat{\pi}_i}$$

```
# Add propensity score
d <- d %>%
  mutate(ps = predict(propensity_model,
                      newdata = d, type = "response"))

##
d <- d %>%
  mutate(weightedMean = a_binary*y / ps - (1-a_binary)*y /(
```

IPW: Estimate the causal effect

```
d %>% select(weightedMean) %>%
  summarize_all(.funs = mean)
```

```
# A tibble: 1 x 1
  weightedMean
         <dbl>
1        0.210
```

Extension 2: 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 2: 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

Since out outcome is binary, it's more appropriate to use logistic regression. Repeat the steps above with logistic regression

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

Helpful hints:

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