

Discussion. Parametric g-formula: Outcome modeling

Cornell STSCI / INFO / ILRST 3900

Fall 2026

causal3900.github.io

21 Oct 2026

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

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

What happens when a subgroup is empty, we have continuous treatment variables or continuous covariates?

We need a model!

Outcome modeling: Parametric and Non-Parametric g-formula:

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

Estimating a Model - Example

- ▶ Question: Does studying more than 7 hours increase the probability of passing an exam?
- ▶ Define the variables:
 - ▶ Y : 1 if pass, 0 if fail
 - ▶ A : 1 if studied more than 7 hours, 0 otherwise
 - ▶ L : grade in Intro to stat

Example

Y	A	intro_grade
0	0	2.10
1	1	1.78
1	1	3.56
0	0	1.89
1	0	3.87
1	1	1.30
0	0	1.56
1	0	4.03
0	1	2.03
1	1	2.67

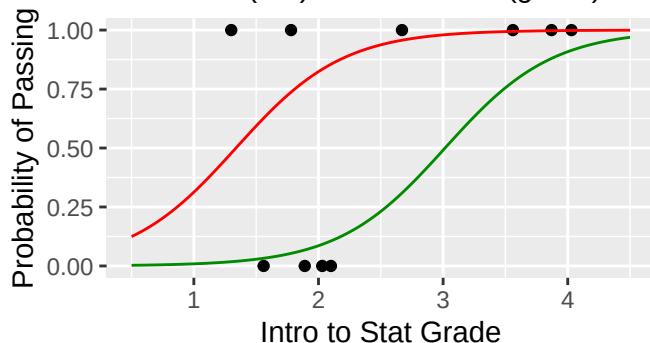
Example

Learn a logistic regression model with glm:

```
model <- glm(Y ~ A + intro_grade ,  
             data = exam_data,  
             family = binomial(link = "logit"))
```

Logistic Regression

Curves shows probabilities predicted by glr
for treated (red) and untreated (green)



Example

We will now estimate the potential probability of passing the exam under treatment and no treatment for each individual.

Y	A	intro_grade	P_yhat1	P_yhat0
0	0	2.10	0.855	0.106
1	1	1.78	0.737	0.053
1	1	3.56	0.994	0.780
0	0	1.89	0.783	0.068
1	0	3.87	0.997	0.880
1	1	1.30	0.478	0.018
0	0	1.56	0.627	0.033
1	0	4.03	0.998	0.914
0	1	2.03	0.834	0.091
1	1	2.67	0.957	0.309

Example

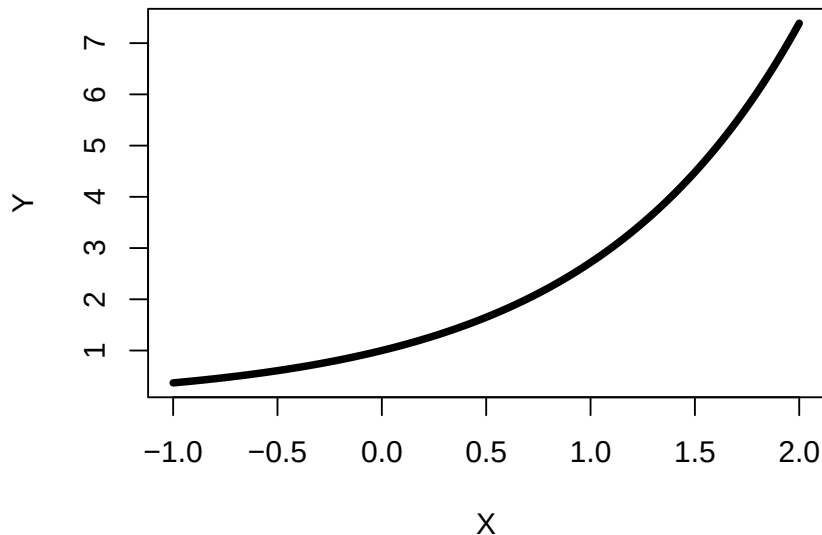
Finally to estimate the average treatment effect we will take the difference for each unit and average over them

```
exam_data %>%  
  mutate(ATE = P_yhat1 - P_yhat0) %>%  
  select(P_yhat1,P_yhat0,ATE) %>%  
  summarise_all(.funs = mean) %>% kbl(digits = 3)
```

P_yhat1	P_yhat0	ATE
0.826	0.325	0.501

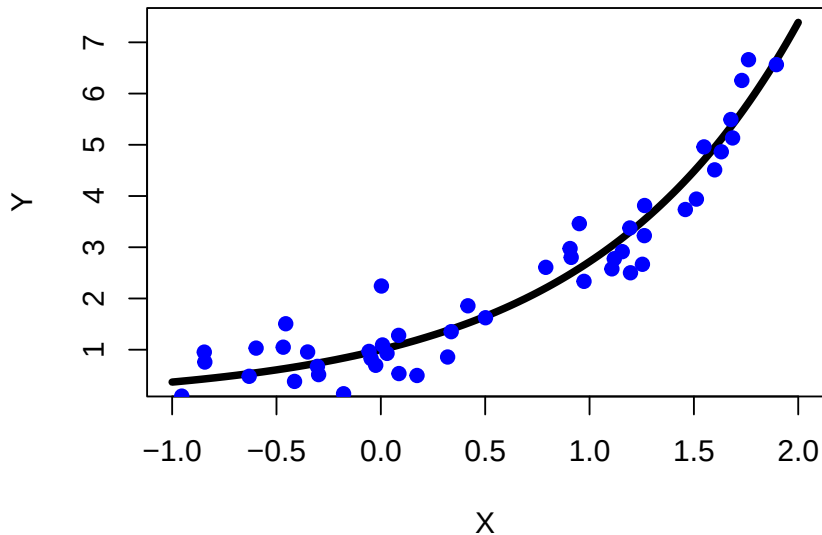
Bias Variance trade-off

Assume first a true model where $Y = e^X$



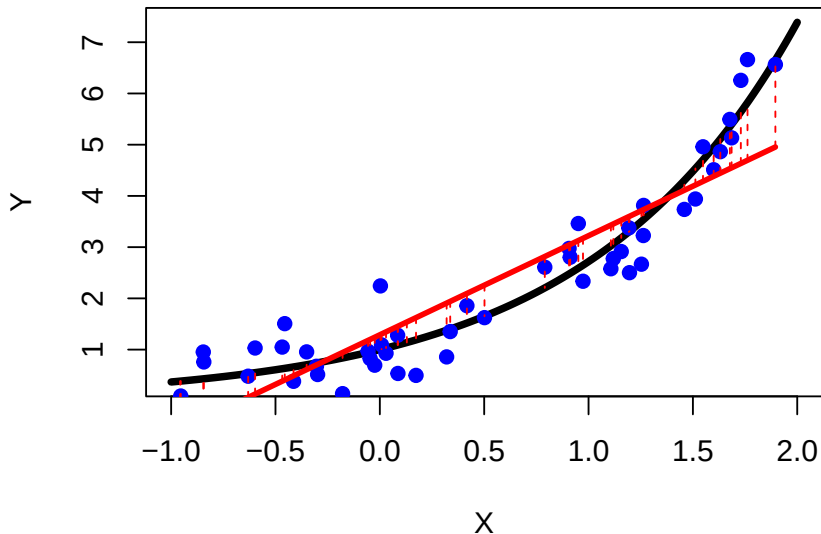
Bias Variance trade-off

We sample $n = 50$ data points from this true model



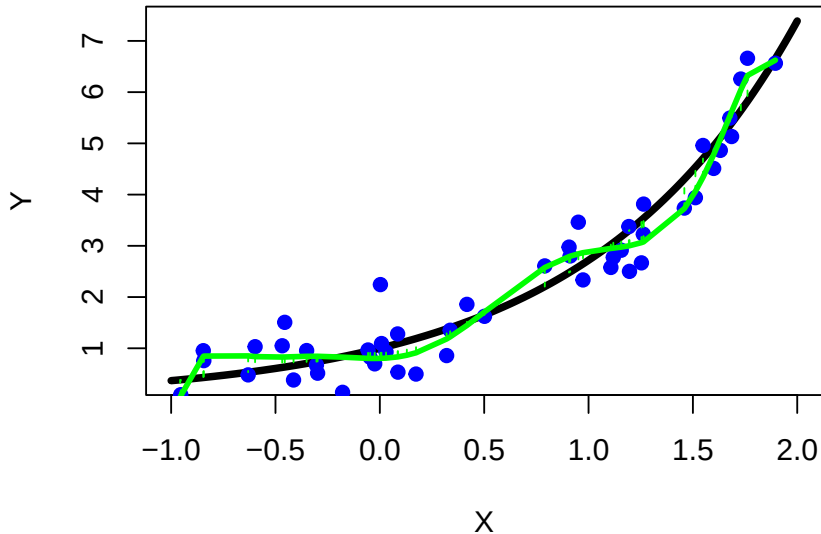
Bias Variance trade-off

Using these 50 data points, we can estimate a linear model, the red line, where the bias is visualized through the dotted vertical lines



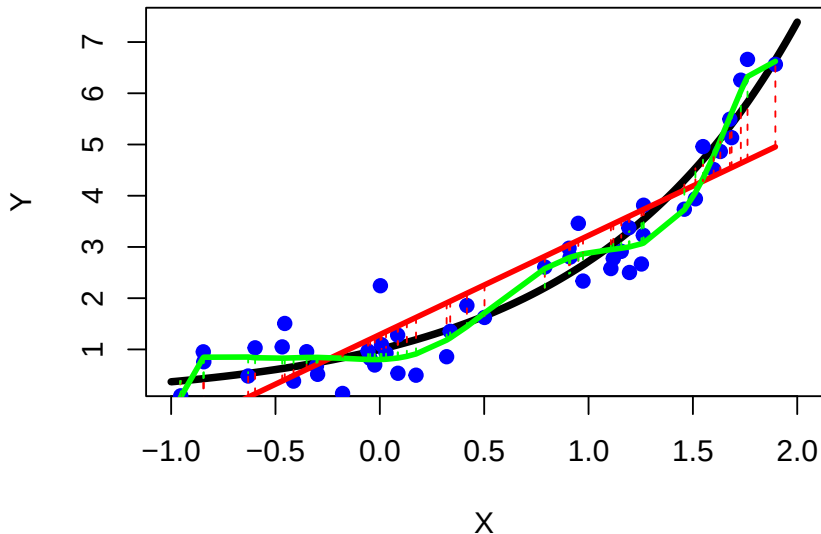
Bias Variance trade-off

Using these 50 data points, we can estimate a more complex model including polynomial of degree 10, the green line, where the bias is visualized through the dotted vertical lines



Bias Variance trade-off

You can see that the linear model (red line) has higher bias than the polynomial (green line).



Bias Variance trade-off

To show the bias variance trade-off we sample n points from the true model, $Y = e^X$, 5000 times to evaluate the models

n	Bias-Linear	Bias-10deg	Var-Linear	Var-10deg
20	0.6674	0.0047	0.0253	0.1382
50	0.5587	0.0020	0.0099	0.0554
100	0.5710	0.0021	0.0050	0.0272
500	0.5555	0.0006	0.0010	0.0055
1000	0.5738	0.0005	0.0005	0.0028

Research question

To what degree does finishing college increase the probability of having a spouse who finished college?

Data. National Longitudinal Survey of Youth 1997

- ▶ Probability sample of U.S. non-institutional civilian youth age 12–16 on Dec 31 1996
- ▶ Surveyed annually 1997–2011, then biennially
- ▶ $n = 7,771$

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

► Parametric model:

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

► Non-Parametric model:

```
fit_rf <- randomForest(  
  y ~ a + sex + race + mom_educ + dad_educ +  
    log_parent_income + log_parent_wealth +  
    test_percentile,  
  data = d %>% mutate(a = as.factor(a),  
                      y = as.factor(y)),  
  ntree = 500)
```

2. Predict conditional average potential outcomes

► Parametric model:

```
average_outcomes_glm <- d %>%  
  mutate(yhat1 = predict(fit_glm,  
                          newdata = d %>%  
                            mutate(a = "college"),  
                          type = "response"),  
  yhat0 = predict(fit_glm,  
                  newdata = d %>%  
                    mutate(a = "no_college"),  
                  type = "response"))
```

2. Predict conditional average potential outcomes

Notice how the predict function adapts to the model specified, pay attention to its argument and output

► Non-Parametric model:

```
average_outcomes_rf <- d %>%  
  mutate(yhat1 = predict(fit_rf,  
                        newdata = d %>%  
                          mutate(a = factor("college",  
                                           levels = c("college", "no_college"))),  
                          type="prob")[, "TRUE"],  
         yhat0 = predict(fit_rf,  
                        newdata = d %>%  
                          mutate(a = factor("no_college",  
                                           levels = c("college", "no_college"))),  
                          type="prob")[, "TRUE"])
```

3. Difference to estimate conditional average effects

► Parametric model:

```
average_effects_glm <-  
  average_outcomes_glm %>%  
  mutate(effect = yhat1 - yhat0)
```

► Non-Parametric model:

```
average_effects_rf <-  
  average_outcomes_rf %>%  
  mutate(effect = yhat1 - yhat0)
```

4. Average over units

► Parametric model:

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

```
# A tibble: 1 x 3  
  yhat1 yhat0 effect  
  <dbl> <dbl> <dbl>  
1 0.381 0.163 0.218
```

► Non-Parametric model:

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

```
# A tibble: 1 x 3  
  yhat1 yhat0 effect  
  <dbl> <dbl> <dbl>  
1 0.394 0.123 0.270
```

G-formula: Parametric vs. Non-Parametric

- ▶ Parametric models have lower variance but probably high bias
- ▶ Non-parametric models have lower bias but higher variance
- ▶ Non-parametric methods usually require more observations
- ▶ Note that each time you run `randomForest` the results might slightly change

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