

Starting to think about data

Cornell STSCI / INFO / ILRST 3900

Fall 2026

causal3900.github.io

15 Oct 2026

Logistics

- ▶ Project details on canvas
- ▶ Part I due 11/6
- ▶ Full project due 12/7
- ▶ Quiz 3: 11/5
- ▶ Quiz 4: 12/2 (in discussion section)

Arc of the course

We began by asking causal questions

- ▶ Defining counterfactuals

Then we discussed causal assumptions

- ▶ Exchangeability and experiments
- ▶ Consistency and positivity
- ▶ Directed Acyclic Graphs

Arc of the course

We began by asking causal questions

- ▶ Defining counterfactuals

Then we discussed causal assumptions

- ▶ Exchangeability and experiments
- ▶ Consistency and positivity
- ▶ Directed Acyclic Graphs

First 7 weeks:

- ▶ What is our target? (causal effect vs association)
- ▶ When can it be recovered from infinite data? (conditional exchangeability)
- ▶ Causal assumptions only (positivity, consistency)
- ▶ No statistical/parametric models (linear regression, logistic regression, etc)

Part III

Remainder of the course:

- ▶ How can we estimate causal effects when given data?
- ▶ Causal assumptions + statistical models
- ▶ Given the target we defined in Part I and II, how can we hit that target?

Identification vs Estimation

So far, everything we've done so far involves objects like

- ▶ Average of outcomes for people with $A = 1, L = 1$

$$E(Y \mid A = 1, L = 1)$$

- ▶ Average over the entire population; i.e., infinite data
- ▶ **Identification:** given infinite data, how can we recover the quantity of interest

Identification vs Estimation

So far, everything we've done so far involves objects like

- ▶ Average of outcomes for people with $A = 1$, $L = 1$

$$E(Y \mid A = 1, L = 1)$$

- ▶ Average over the entire population; i.e., infinite data
- ▶ **Identification:** given infinite data, how can we recover the quantity of interest

In practice, we only have a finite sample

- ▶ Average of outcomes in the sample for people with $A = 1$, $L = 1$

$$\frac{1}{n} \sum_{i:A_i=1, L_i=1} Y_i$$

- ▶ Will sometimes use $\hat{E}(Y \mid A = 1, L = 1)$:
- ▶ **Estimation:** given the data we have, how can we get close to the quantity of interest

Example: causal effect of job training

Identification:

- ▶ Average of outcomes for people in US who attend job training ($A = 1$) and are young ($L = 1$)

$$E(Y \mid A = 1, L = 1)$$

- ▶ Average over the entire population (including people who might appear later)
- ▶ Fixed but unknown quantity

Example: causal effect of job training

Identification:

- ▶ Average of outcomes for people in US who attend job training ($A = 1$) and are young ($L = 1$)

$$E(Y \mid A = 1, L = 1)$$

- ▶ Average over the entire population (including people who might appear later)
- ▶ Fixed but unknown quantity

Estimation:

- ▶ We have a sample of 1000 people
- ▶ Average of outcomes for people in our who attend job training ($A = 1$) and are young ($L = 1$)

$$\frac{1}{n} \sum_{i: A_i=1, L_i=1} Y_i$$

- ▶ Will change if I get a new sample of 1000 people

Learning goals for today

At the end of class, you will be able to

- ▶ explain the curse of dimensionality
- ▶ recognize the possible futility of nonparametric estimation

Motivating a research question¹

Income inequality across households depends on

1. inequality across individuals
2. how individuals pool into households

A college degree affects (1) and (2)

¹[Mare 1991](#), [Schwartz 2013](#)

Research question

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

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
- ▶ $n = 8,984$

Research question

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

Research question

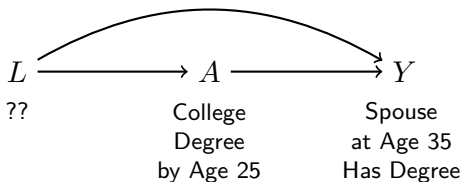
To what degree does finishing college increase the probability of having a spouse who finished college?

- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA

Research question

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

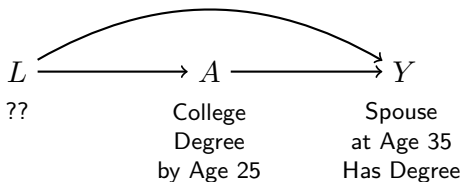
- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA



Research question

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

- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA



Research question

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

Research question

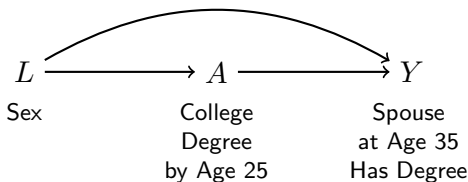
To what degree does finishing college increase the probability of having a spouse who finished college?

- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA

Research question

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

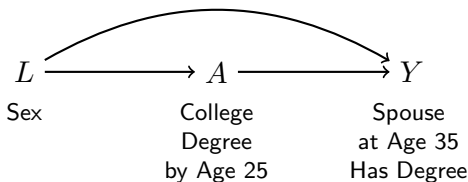
- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA



Research question

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

- ▶ Treatment A : Finished BA by age 25
- ▶ Outcome Y : Spouse or partner at age 30–40 holds a BA
 - ▶ 0 if no spouse or partner, or partner with no BA
 - ▶ 1 if spouse or partner holds a BA



Adjustment procedure

- 1) Estimate within subgroups defined by $\{\text{sex}\}$
- 2) Aggregate over the subgroups

Data

```
d %>%  
  select(sex, a, y) %>%  
  print(n = 8)
```

```
# A tibble: 7,771 x 3  
  sex      a      y  
  <chr> <chr>   <lgl>  
1 Female college FALSE  
2 Male   no_college FALSE  
3 Female no_college FALSE  
4 Male   no_college TRUE  
5 Female no_college FALSE  
6 Male   no_college FALSE  
7 Female college FALSE  
8 Male   college TRUE  
# i 7,763 more rows
```

1) Estimate in subgroups

```
ybar_in_subgroups <- d %>%  
  # Group by confounders and treatment  
  group_by(sex, a) %>%  
  # Summarize mean outcomes and nber of cases  
  summarize(ybar = mean(y),  
            n = n(),  
            .groups = "drop") %>%  
  print()
```

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

	sex	a	ybar	n
	<chr>	<chr>	<dbl>	<int>
1	Female	college	0.467	896
2	Female	no_college	0.102	2953
3	Male	college	0.614	637
4	Male	no_college	0.174	3285

1) Estimate in subgroups

```
# A tibble: 4 x 4
  sex      a      ybar      n
<chr> <chr>   <dbl> <int>
1 Female college 0.467   896
2 Female no_college 0.102 2953
3 Male   college 0.614   637
4 Male   no_college 0.174 3285
```

1) Estimate in subgroups

```
# A tibble: 4 x 4
  sex      a      ybar      n
  <chr> <chr>    <dbl> <int>
1 Female college  0.467   896
2 Female no_college 0.102  2953
3 Male   college  0.614   637
4 Male   no_college 0.174  3285
```

```
pivoted <- ybar_in_subgroups %>%
  pivot_wider(names_from = a,
              values_from = c("ybar","n")) %>%
  print()
```

```
# A tibble: 2 x 5
  sex      ybar_college ybar_no_college n_college n_no_college
  <chr>          <dbl>          <dbl>      <int>      <int>
1 Female          0.467          0.102      896      2953
2 Male            0.614          0.174      637      3285
```

1) Estimate in subgroups

```
# A tibble: 2 x 5
  sex      ybar_college ybar_no_college n_college n_no_college
<chr>      <dbl>         <dbl>      <int>      <int>
1 Female    0.467         0.102      896      2953
2 Male     0.614         0.174      637      3285
```

1) Estimate in subgroups

```
# A tibble: 2 x 5
```

	sex	ybar_college	ybar_no_college	n_college	n_no_college
	<chr>	<dbl>	<dbl>	<int>	<int>
1	Female	0.467	0.102	896	2953
2	Male	0.614	0.174	637	3285

```
cate <- pivoted %>%  
  mutate(conditional_effect = ybar_college - ybar_no_college,  
         n_in_stratum = n_college + n_no_college) %>%  
  select(sex, conditional_effect, n_in_stratum) %>%  
  print()
```

```
# A tibble: 2 x 3
```

	sex	conditional_effect	n_in_stratum
	<chr>	<dbl>	<int>
1	Female	0.365	3849
2	Male	0.440	3922

2) Aggregate over subgroups

```
# A tibble: 2 x 3
  sex      conditional_effect n_in_stratum
<chr>          <dbl>          <int>
1 Female          0.365            3849
2 Male            0.440            3922
```

2) Aggregate over subgroups

```
# A tibble: 2 x 3
  sex      conditional_effect n_in_stratum
<chr>          <dbl>          <int>
1 Female          0.365            3849
2 Male            0.440            3922
```

```
cate %>%
  summarize(population_average_effect = weighted.mean(
    conditional_effect,
    w = n_in_stratum
  ))
```

```
# A tibble: 1 x 1
  population_average_effect
          <dbl>
1              0.403
```

Recap: Intuition

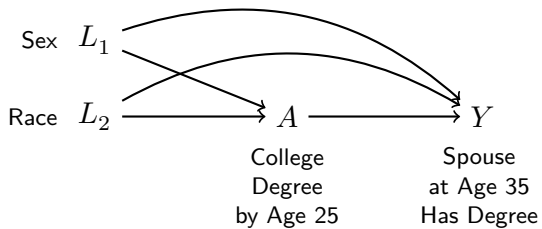
<div data-bbox="358 241 428 262">College</div> <div data-bbox="344 581 441 601">No College</div>	<div data-bbox="937 218 1007 239">College</div> <div data-bbox="923 558 1020 579">No College</div>
Female	Male

Recap: In code

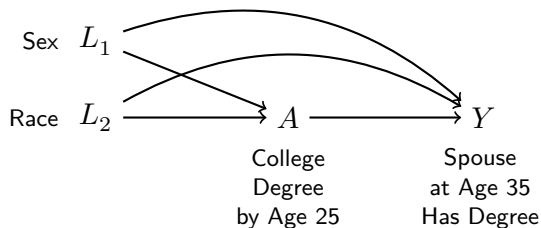
```
d %>%  
  # Group by confounders and treatment  
  group_by(sex, a) %>%  
  # Estimate within subgroups  
  summarize(ybar = mean(y),  
            n = n(),  
            .groups = "drop") %>%  
  pivot_wider(names_from = a,  
              values_from = c("ybar", "n")) %>%  
  mutate(conditional_effect = ybar_college - ybar_no_college,  
         n_in_stratum = n_college + n_no_college) %>%  
  # Aggregate over subgroups  
  summarize(population_average_effect = weighted.mean(  
    conditional_effect,  
    w = n_in_stratum  
  ))
```

```
# A tibble: 1 x 1  
  population_average_effect  
                <dbl>  
1                0.403
```

Adjust for sex and race



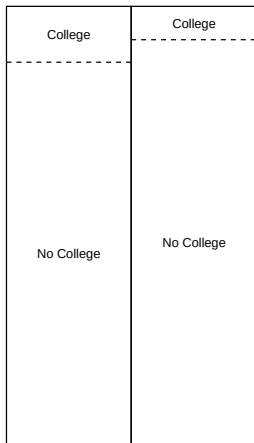
Adjust for sex and race



- 1) Estimate effects within subgroups defined by {sex, race}
- 2) Aggregate over subgroups

Adjust for sex and race

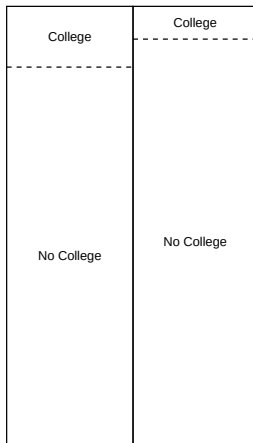
Hispanic



Female

Male

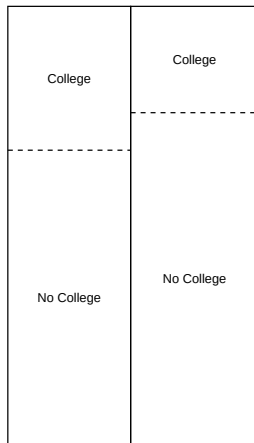
Non-Hispanic Black



Female

Male

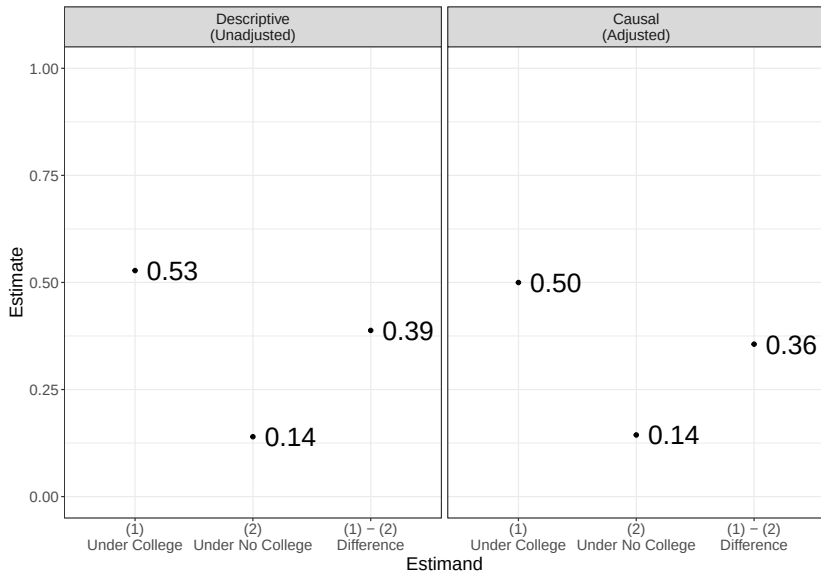
Non-Hispanic Non-Black



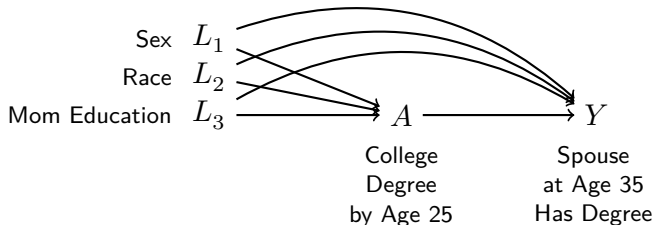
Female

Male

Adjust for sex and race

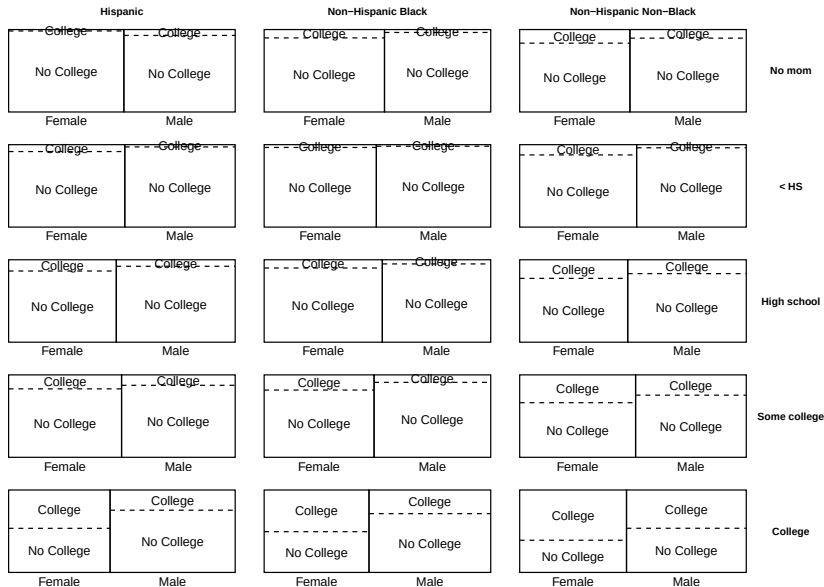


Adjust for sex, race, mom education

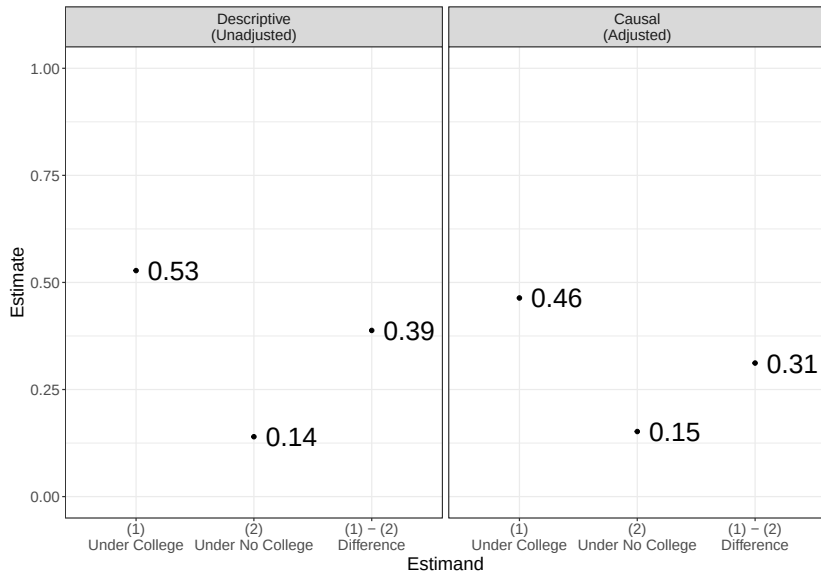


- 1) Estimate effects within subgroups defined by $\{\text{race, sex, mom education}\}$
- 2) Aggregate over subgroups

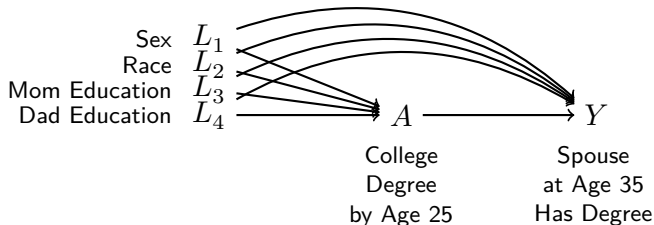
Adjust for sex, race, mom education



Adjust for sex, race, mom education



Adjust for sex, race, mom education, dad education

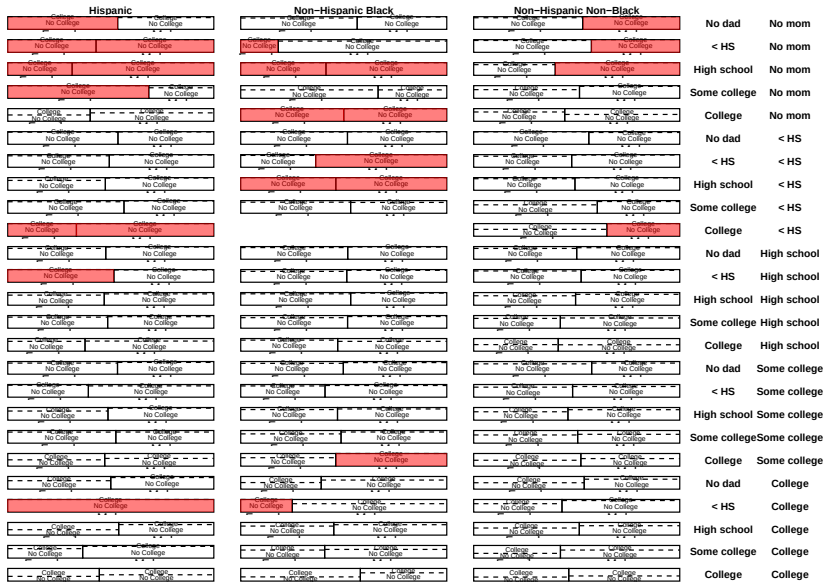


- 1) Estimate effects within subgroups defined by $\{\text{race, sex, mom education, dad education}\}$
- 2) Aggregate over subgroups

Adjust for sex, race, mom education, dad education

Hispanic	Non-Hispanic Black	Non-Hispanic Non-Black		
			No dad	No mom
			< HS	No mom
			High school	No mom
			Some college	No mom
			College	No mom
			No dad	< HS
			< HS	< HS
			High school	< HS
			Some college	< HS
			College	< HS
			No dad	High school
			< HS	High school
			High school	High school
			Some college	High school
			College	High school
			No dad	Some college
			< HS	Some college
			High school	Some college
			Some college	Some college
			College	Some college
			No dad	College
			< HS	College
			High school	College
			Some college	College
			College	College

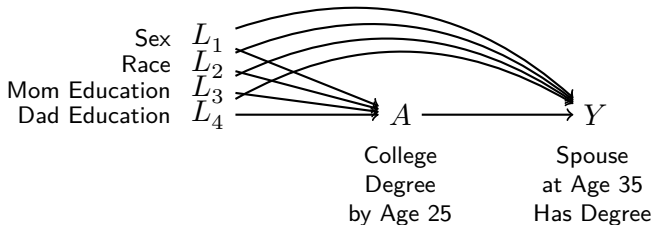
Adjust for sex, race, mom education, dad education



Curse of dimensionality: Unpopulated cells

```
# A tibble: 147 x 6
  sex      race mom_educ dad_educ n_college n_no_college
  <chr> <chr>   <fct>   <fct>      <int>      <int>
1 Female H      No mom    No dad      NA          32
2 Female H      No mom    < HS        NA           6
3 Female H      No mom    High school NA           5
4 Female H      No mom    Some college NA          13
5 Female H      < HS     College     NA           1
6 Female H    High school < HS        NA          34
7 Female Non-H B No mom    < HS        NA           2
8 Female Non-H B No mom    High school NA          12
9 Female Non-H B No mom    College     NA           4
10 Female Non-H B < HS     High school NA          24
# i 137 more rows
```

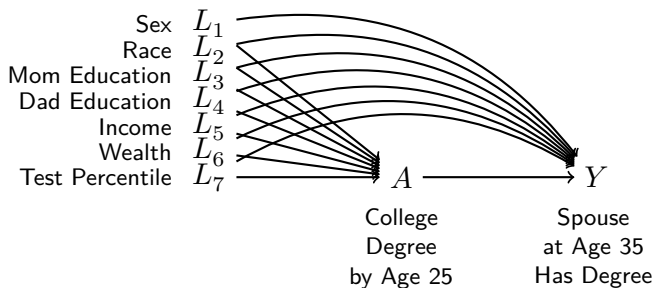
Curse of dimensionality



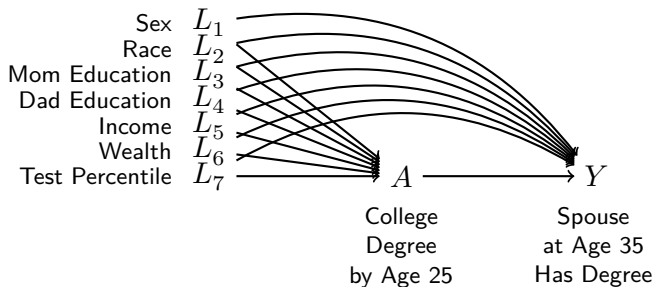
4.2% of the sample

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

Curse of dimensionality



Curse of dimensionality



100% of the sample

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

What's the problem?

- ▶ Empty cells are almost like (but not exactly) a positivity violation

What's the problem?

- ▶ Empty cells are almost like (but not exactly) a positivity violation
- ▶ In the population, it's not impossible to observe treated/untreated people for each level of the covariates
- ▶ Problem is that we only observe a finite sample

Potential Solution: statistical assumptions

No assumptions (non-parametric):

- ▶ Average for individuals with test percentile of 60% tells us nothing about individuals with test percentile of 61%

Potential Solution: statistical assumptions

No assumptions (non-parametric):

- ▶ Average for individuals with test percentile of 60% tells us nothing about individuals with test percentile of 61%

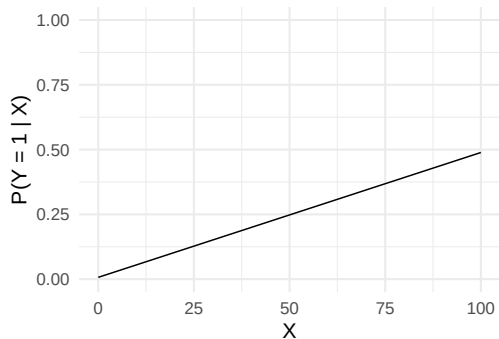
Reasonable assumptions:

- ▶ Average for individuals with with test percentile of 60% should be close to average for people with income of with test percentile of 61%
- ▶ Probability of $Y = 1$ should increase with test percentile
- ▶ Probability for 61% should be larger than probability for 60%

Statistical Assumptions

Linear regression assumes:

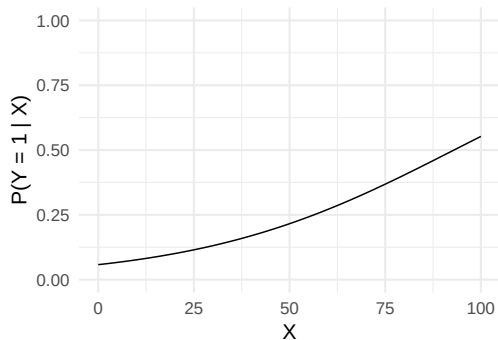
$$E(Y \mid L = \ell) = \beta_0 + \beta_1 \ell$$



Statistical Assumptions

Logistic regression assumes:

$$E(Y \mid L = \ell) = \frac{\exp(\beta_0 + \beta_1 \ell)}{1 + \exp(\beta_0 + \beta_1 \ell)}$$



Assumptions

- ▶ Assumptions can be helpful when correct (or very close to correct)
- ▶ Assumptions can hurt when incorrect (or obviously incorrect)

Assumptions

- ▶ Assumptions can be helpful when correct (or very close to correct)
- ▶ Assumptions can hurt when incorrect (or obviously incorrect)
- ▶ We need to assume something, but want to avoid assuming wrong things
- ▶ Goal: assume enough to be useful, but not too much to be wrong

Learning goals for today

At the end of class, you will be able to

- ▶ explain the curse of dimensionality
- ▶ recognize the possible futility of nonparametric estimation

After class, you should

- ▶ read [Hernán & Robins Ch 11](#)
- ▶ attend discussion: you will learn to use models!