

# Why model?

Cornell STSCI / INFO / ILRST 3900

Fall 2023

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

26 Sep 2023

# 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

5 weeks

# 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

5 weeks

0 statistical models

# 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 question<sup>1</sup>

Income inequality across households depends on

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

A college degree affects (1) and (2)

---

<sup>1</sup>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
- ▶  $n = 8,984$



# Data access

To access these data, first

- ▶ set your working directory where you will be working
- ▶ download two supporting files from us
  1. [nlsy97.NLSY97](#) is a tagset file containing the variable names
  2. [prepare\\_nlsy97.R](#) is an R script to prepare the data

## Data access

Now go to the data distributor

1. [Register](#) with the survey
2. [Log in](#) to the NLS Investigator
3. Choose the NLSY97 study
4. Upload the tagset [nlsy97.NLSY97](#) that you downloaded from us
5. In the Investigator, download the data. Type to change the file name from default to `nlsy97`
6. Unzip the file. Drag `nlsy97.dat` into the folder you will work in
7. In your R console, run the line of code below
  - ▶ this will take about 30 seconds to run
  - ▶ you will need these R packages: `tidyverse` and `Amelia`

```
source("prepare_nlsy97.R")
```

In the future, you can now load the data with

```
d <- readRDS("d.RDS")
```

Register with the survey

# NLS Investigator

**Tell us about yourself** - Only email is required

First name:	<input type="text"/>
Last name:	<input type="text"/>
Organization:	<input type="text"/>
Email: *	<input type="text"/>
Confirm Email: *	<input type="text"/>

**Enter your username and password** - All fields are required

Username: *	<input type="text"/>
Password: *	<input type="text"/>
Confirm password: *	<input type="text"/>

Username is automatically filled in from email field.

Password must be 8 characters or more and contain at least one numeric and one non numeric character.  
In addition the password must not be based on username.

I agree to the NLS Investigator [Privacy Policy](#).

\* Required field

Register

Choose the NLSY97 study

# NLS Investigator

**Select the study you want to work with:**

NLSY97 (National Longitudinal Survey of Youth 1997) ▾

**Select a substudy:**

NLSY97 1997-2019 (rounds 1-19) ▾

Released November 01, 2021

# Upload our tagset

Choose Tagsets

Variable Search

Review Sele

**Upload Tagset** (from PC):

Choose File No file chosen

Upload

# Download the data

Choose Tagsets

Variable Search

Review Selected Variables (6)

Codebook

Save / Download

Save Tagset

Basic Download

Advanced Download

Manage Downloads

## Customize your advanced download:

### Create Download of Data

- Tagset (list of selected variables)
- SAS® control file (includes the datafile of selected variables)
- SPSS® control file (includes the datafile of selected variables)
- STATA® dictionary file of selected variables
- R® Source code (includes the datafile of selected variables)
- Codebook of selected variables
- Short Description File
- Comma-delimited datafile of selected variables (to be read in Excel, etc.)  
Column headers -- Use  Reference Number  Question Name (does not guarantee uniqueness)

### Create Frequency / Table

- Apply Universe Restrictors** ([How to use Universe Restrictors](#))
- Notify me by email when download is complete.

Filename:

Filename must only contain alpha, numeric, hyphen or underscore characters.



## Run our code

This code prepares the data file (one time, takes about 30 seconds)

```
source("prepare_NLSY97.R")
```

This code loads the prepared data (after the above, very fast)

```
d <- readRDS("d.RDS")
```

## Research question

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



# Research question

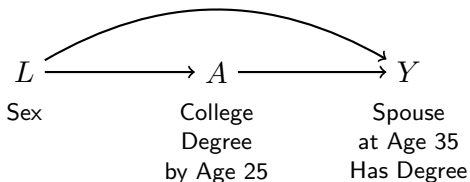
To what degree does finishing college increase the probability of having a spouse who finished college?

- ▶ 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?

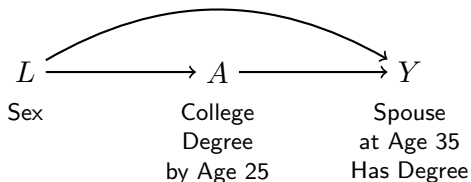
- ▶ Treatment  $A$ : Finished BA by age 25
- ▶ Outcome  $Y$ : Spouse or partner at age 30–40 holds a BA
  - ▶ 0 if no spouse or partner, or partner with no BA
  - ▶ 1 if spouse or partner holds a BA



# 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
```

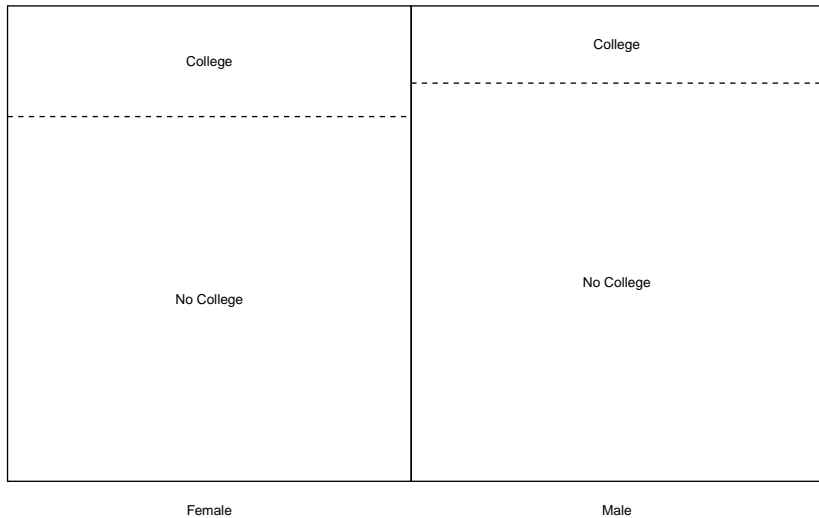
## 2) Aggregate over subgroups

```
# 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

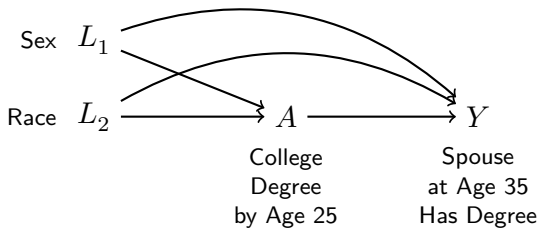


## Recap: In code

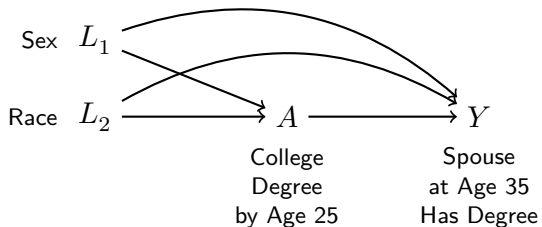
```
d %>%
  # Group by confounders and treatment
  group_by(sex, a) %>%
  # Estimate within subgroups
  summarize(ybar = mean(y),
            n = n(),
            .groups = "drop") %>%
  pivot_wider(names_from = a,
              values_from = c("ybar","n")) %>%
  mutate(conditional_effect = ybar_college - 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



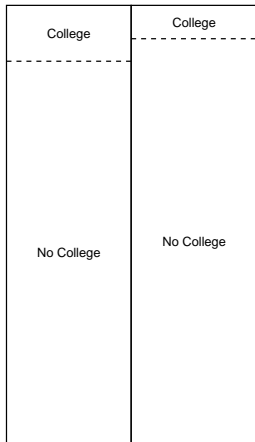
## Adjust for sex and race



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

# Adjust for sex and race

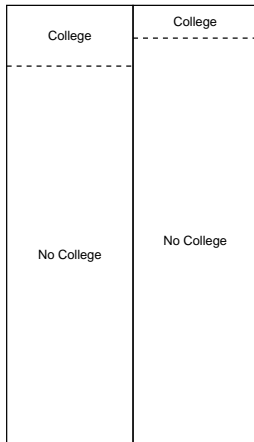
Hispanic



Female

Male

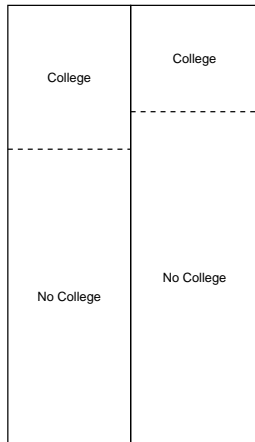
Non-Hispanic Black



Female

Male

Non-Hispanic Non-Black

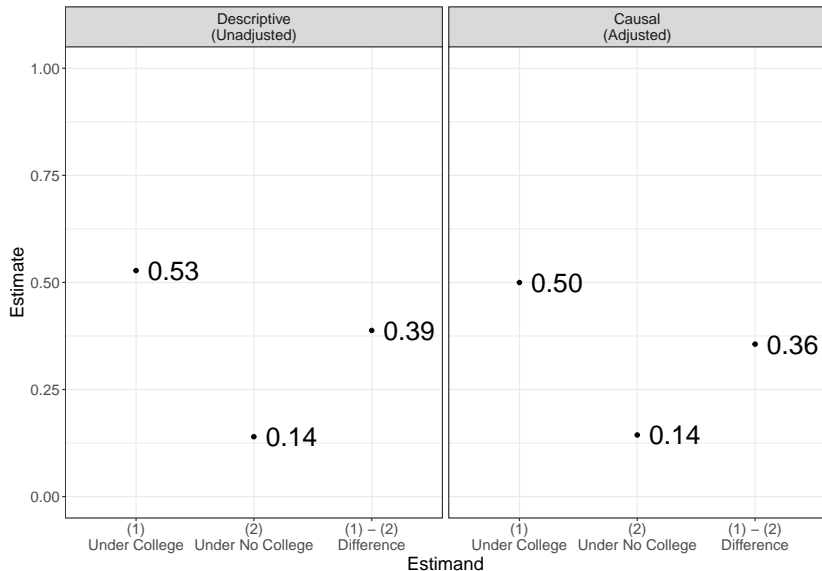


Female

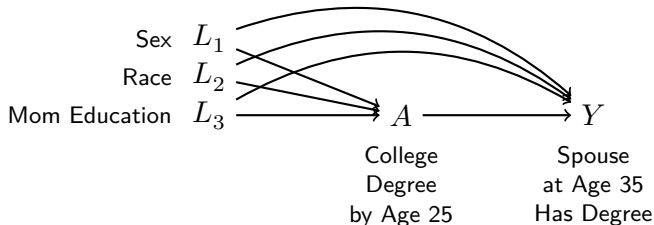
Male



# Adjust for sex and race

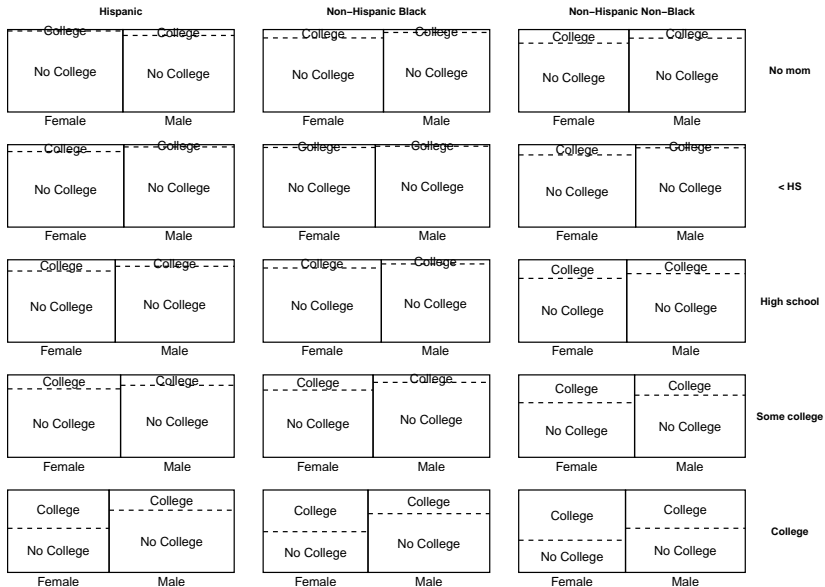


## Adjust for sex, race, mom education

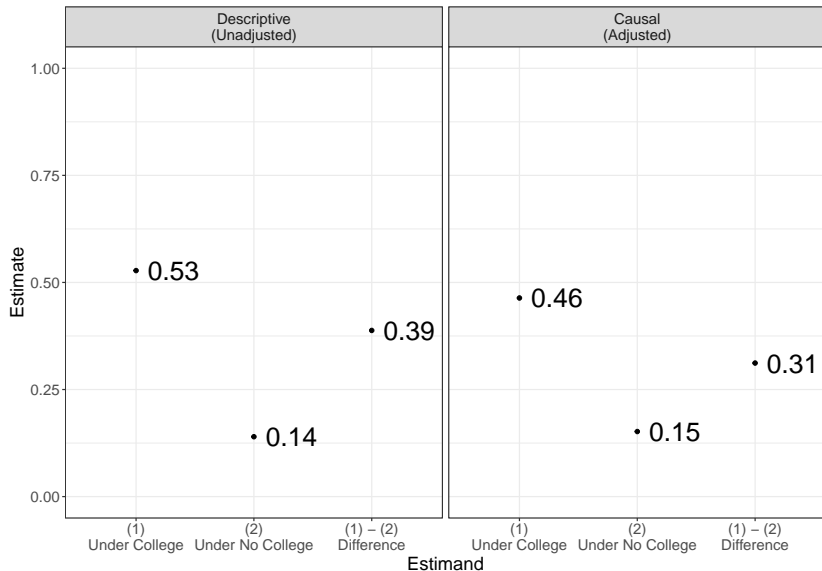


- 1) Estimate effects within subgroups defined by {race, sex, mom education}
- 2) Aggregate over subgroups

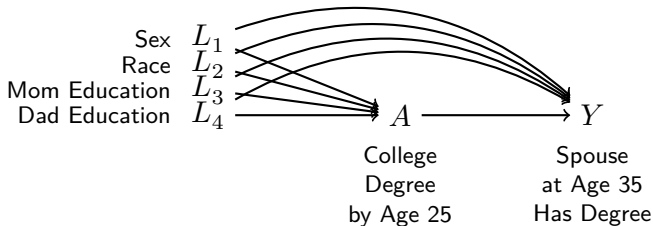
# Adjust for sex, race, mom education



# Adjust for sex, race, mom education

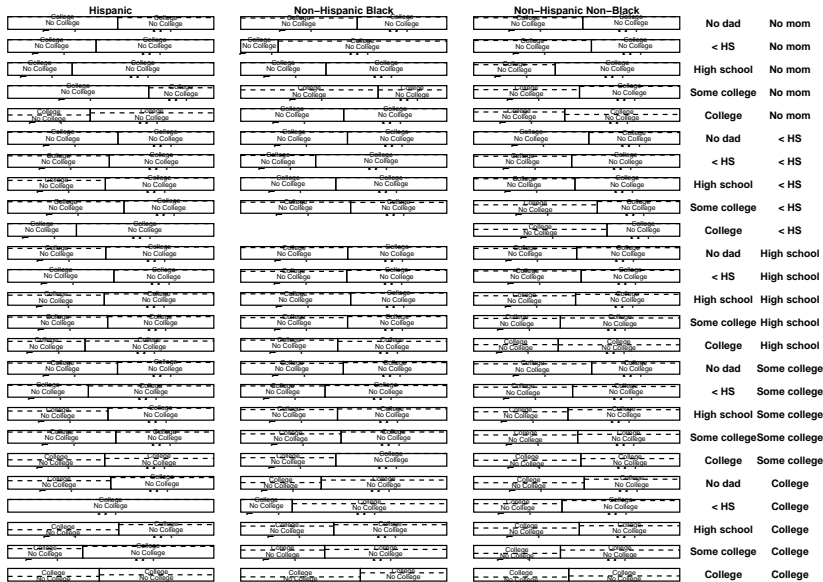


Adjust for sex, race, mom education, dad education

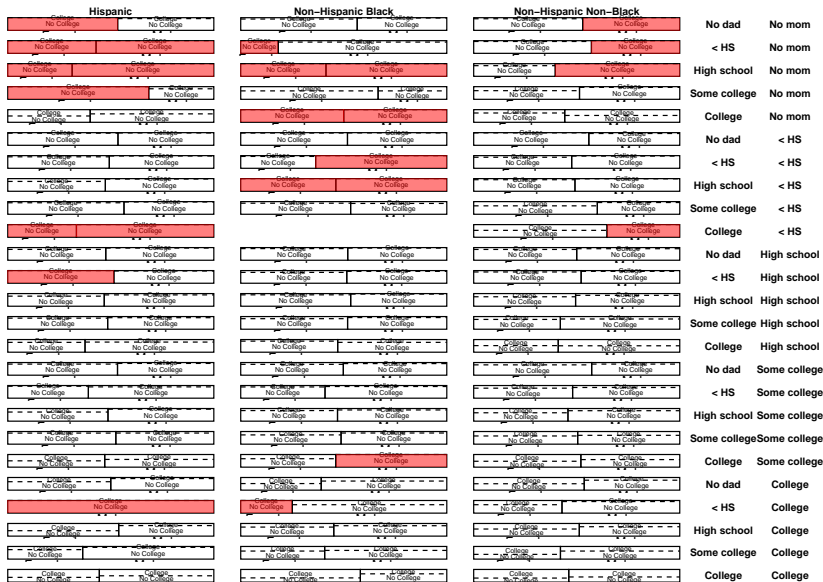


- 1) Estimate effects within subgroups defined by {race,sex, mom education, dad education}
- 2) Aggregate over subgroups

# Adjust for sex, race, mom education, dad education



# Adjust for sex, race, mom education, dad education

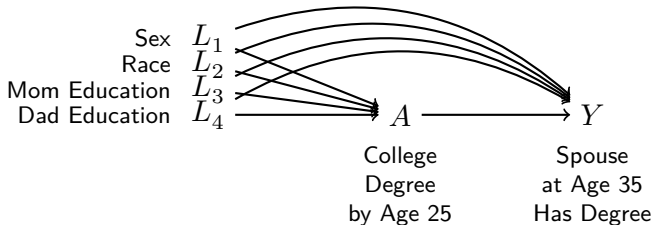


# Curse of dimensionality: Unpopulated cells

```
# 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
```



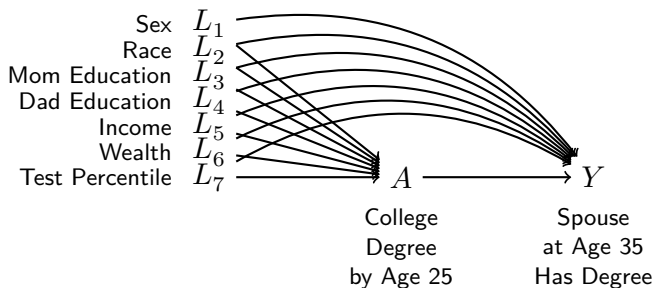
# Curse of dimensionality



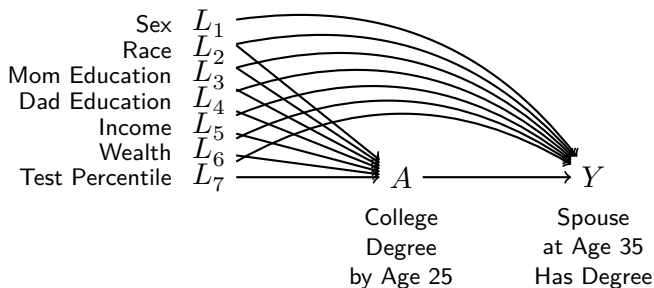
**4.2%** of the sample

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

# 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!