Randomization and matching

Session 7

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

The magic of randomization

How to analyze RCTs

The "gold" standard

Adjustment with matching

The magic of randomization

Why randomize?

Fundamental problem of causal inference

$$\delta_i = Y_i^1 - Y_i^0 \quad \text{in real life is} \quad \delta_i = Y_i^1 - ???$$

Individual-level effects are impossible to observe!

There are no individual counterfactuals!

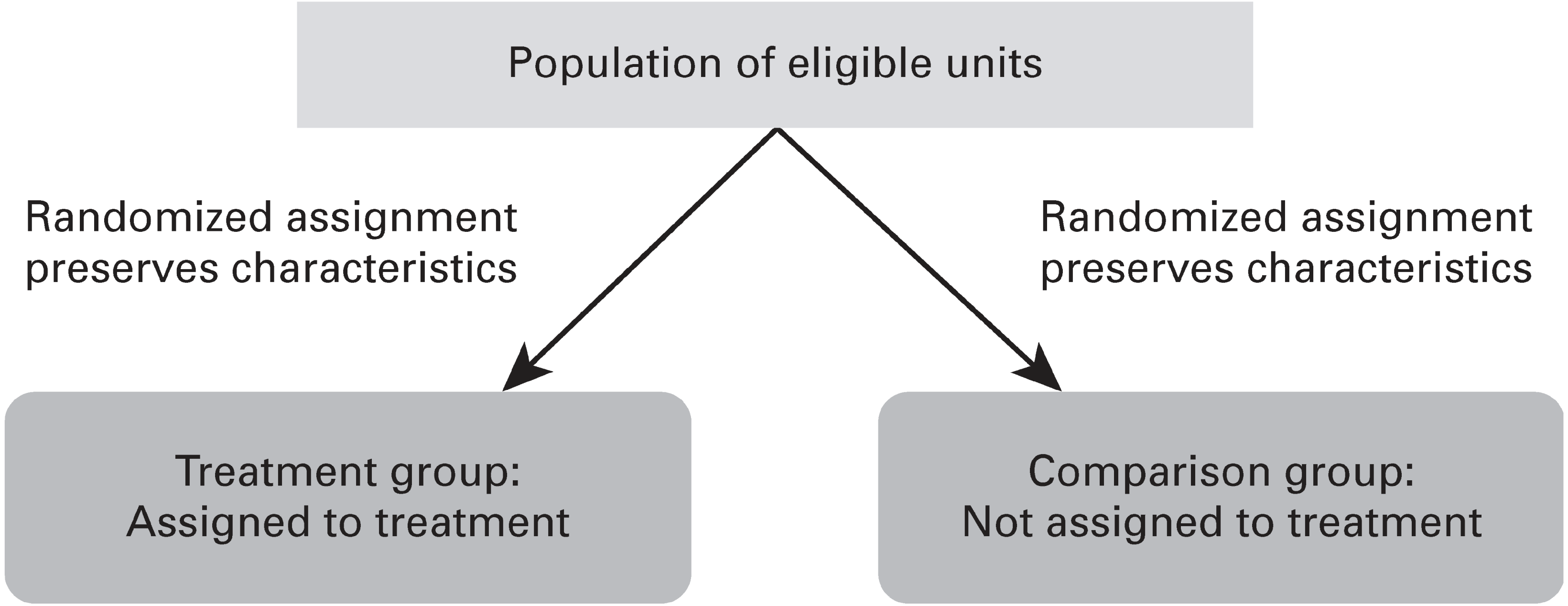
Why randomize?

$$\delta = (\bar{Y} \mid P = 1) - (\bar{Y} \mid P = 0)$$

Comparing average outcomes only works if groups that received/didn't receive treatment look the same

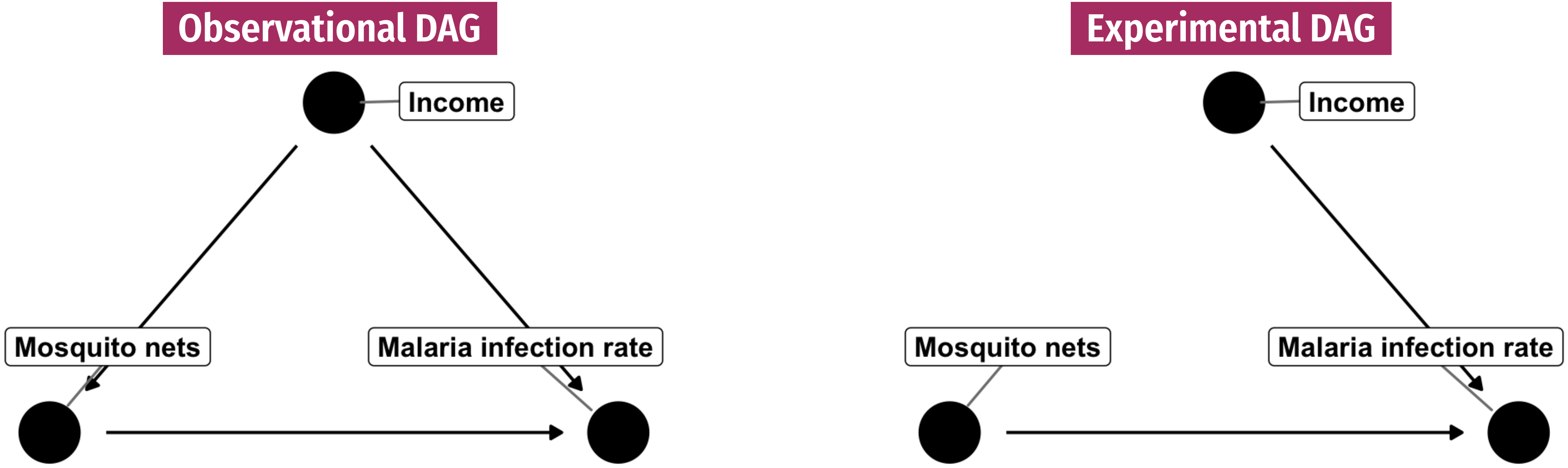
Why randomize?

With big enough samples, the magic of randomization helps make comparison groups comparable



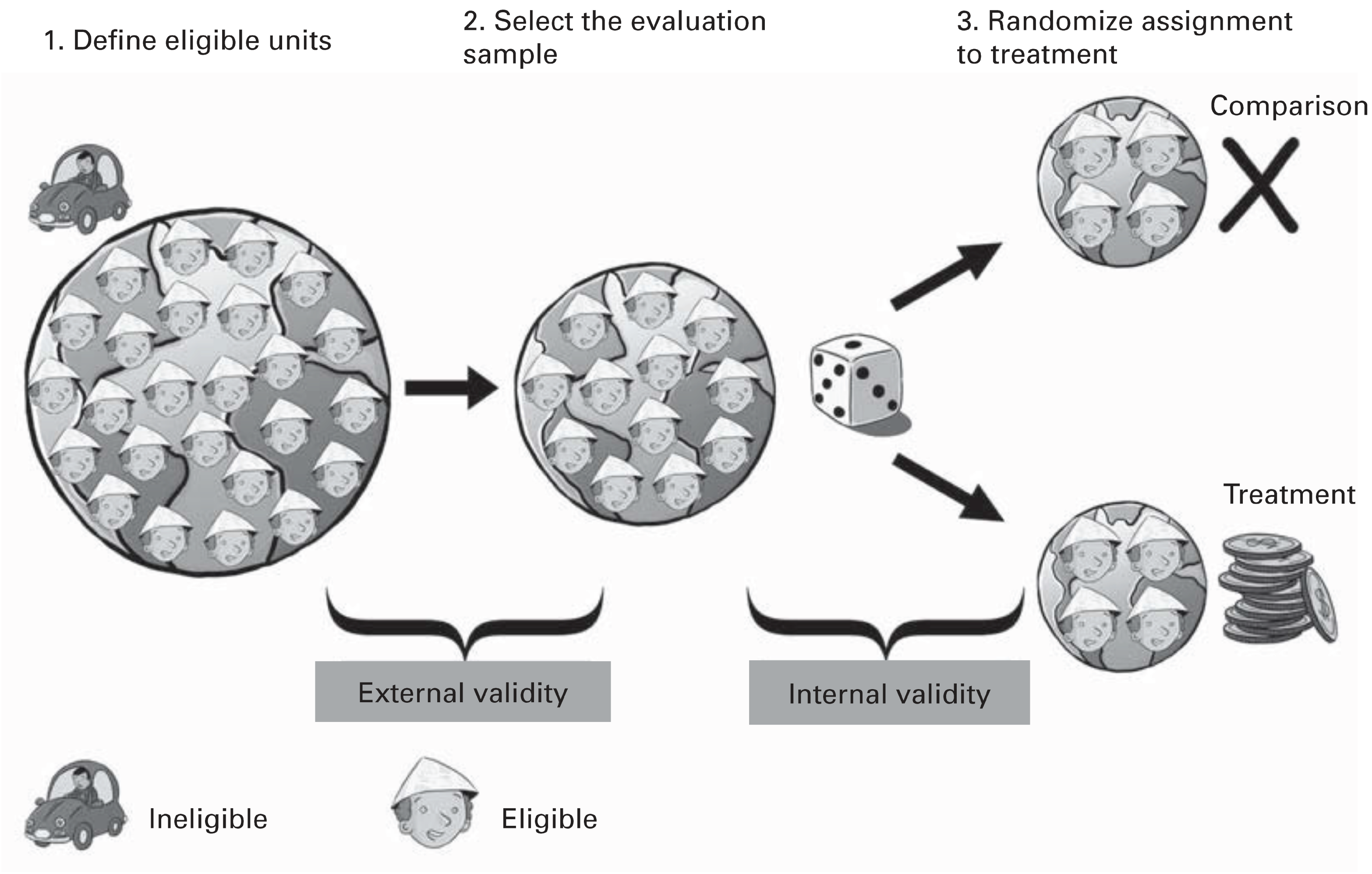
RCTs and DAGs

$$E[\text{Malaria infection rate} \mid do(\text{Mosquito net})]$$



When you *do*() X, delete all arrows into X; confounders don't influence treatment!

How to randomize?



Random assignment

Coins

Dice

Unbiased lottery

Random numbers + threshold

Atmospheric noise

random.org

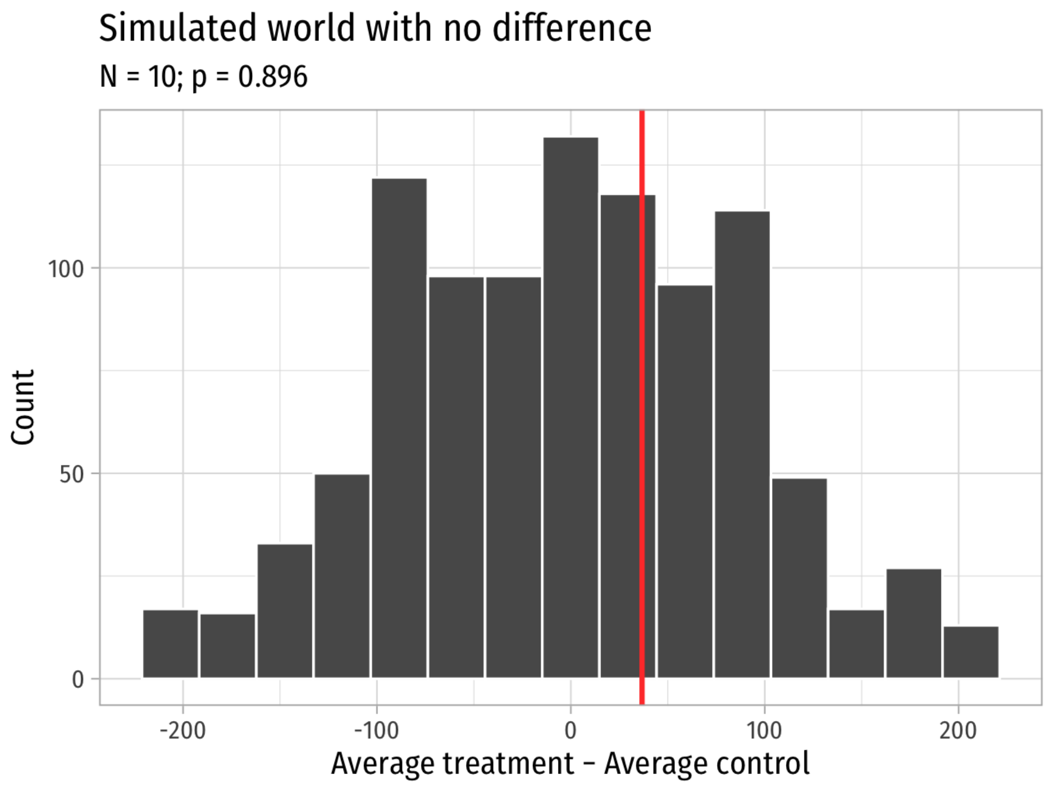
How big of a sample?

A training program causes incomes to rise by $40

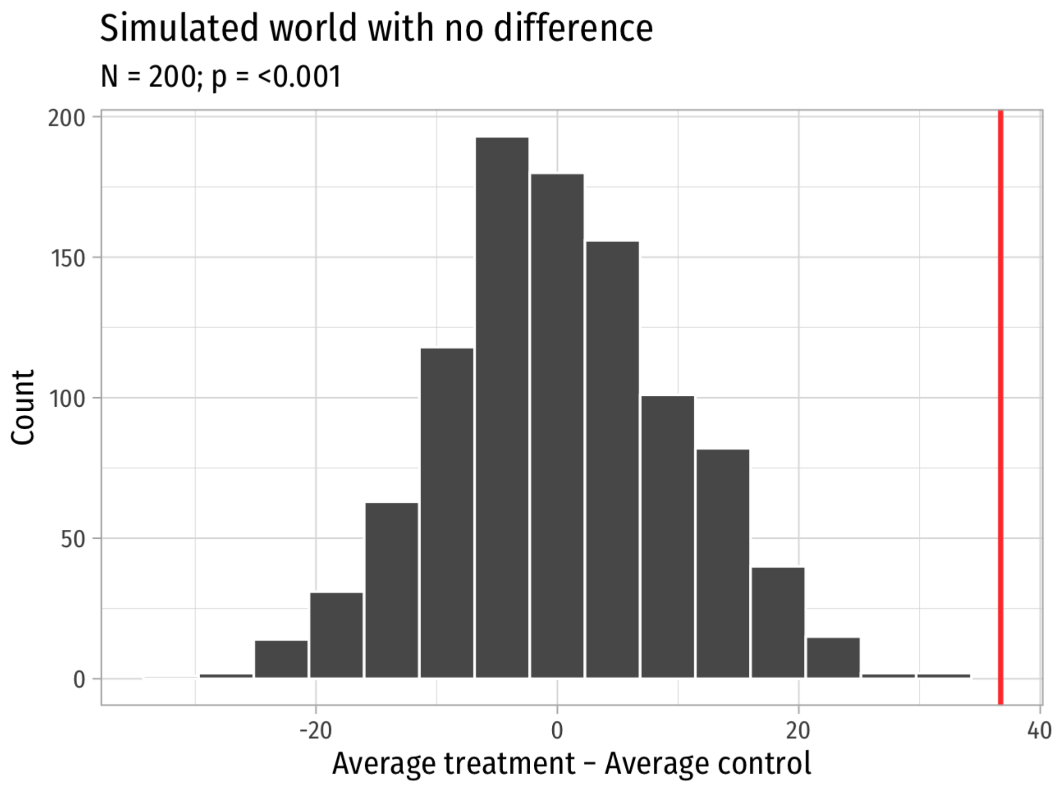
Person	Group	Before	After	Difference
295	Control	122.09	229.04	106.95
126	Treatment	205.60	199.84	-5.76
400	Control	133.25	130.40	-2.85
94	Treatment	270.11	206.56	-63.54
250	Control	344.37	222.89	-121.49
59	Treatment	312.41	268.06	-44.35

Power

Enroll 10 participants

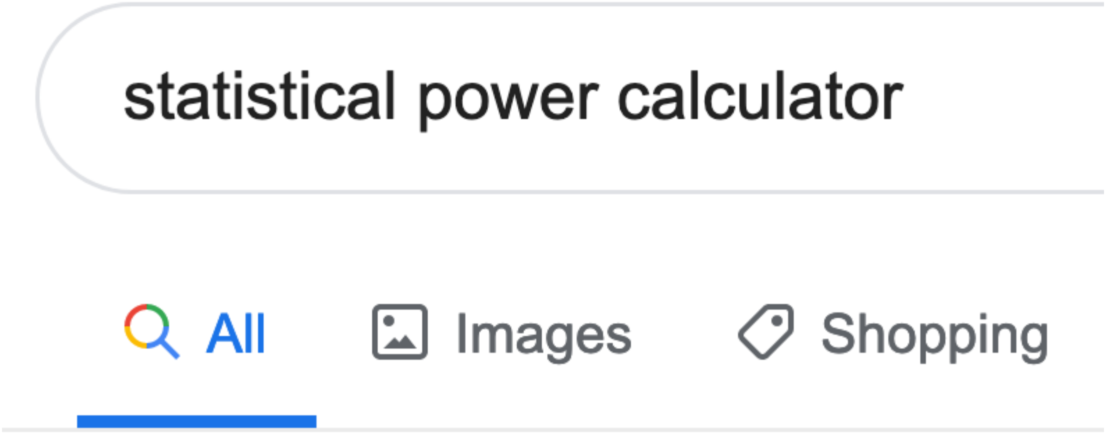


Enroll 200 participants



What's the right sample size?

Use a statistical power calculator to make sure you can potentially detect an effect



How to analyze RCTs

How to analyze RCTs

Surprisingly easy, statistically!

Step 1: Check that key demographics and other confounders are balanced

Step 2: Find difference in average outcome in treatment and control groups

Example RCT

```
imaginary_program
```

```
## # A tibble: 800 x 6
##    person treatment   age sex    income_after male_num
##     <int> <chr>     <dbl> <chr>         <dbl>    <dbl>
##  1    498 Control      45 Female         179.        0
##  2    308 Treatment    37 Male           247.        1
##  3    677 Control      35 Female         369.        0
##  4     31 Treatment    39 Female         203.        0
##  5    543 Control      36 Female         190.        0
##  6    434 Control      30 Female         278.        0
##  7    234 Treatment    28 Male           356.        1
##  8    272 Treatment    45 Male           260.        1
##  9    523 Control      49 Female         174.        0
## 10    649 Control      49 Male           224.        1
## # … with 790 more rows
```

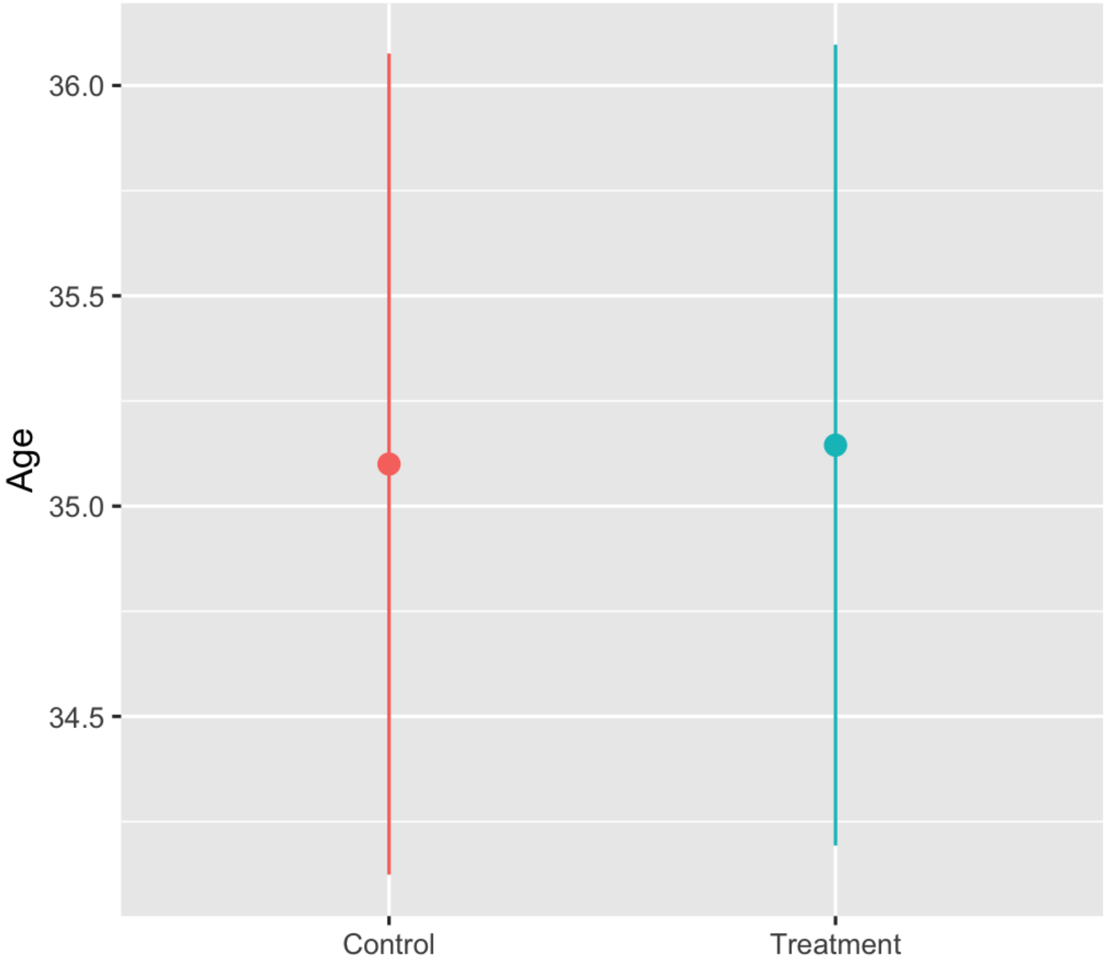
1. Check balance

```r
imaginary_program %>%
  group_by(treatment) %>%
  summarize(avg_age = mean(age),
            prop_male = mean(sex == "Male"))
```

```
## # A tibble: 2 x 3
##   treatment avg_age prop_male
##   <chr>       <dbl>     <dbl>
## 1 Control      35.1     0.562
## 2 Treatment    35.1     0.512
```

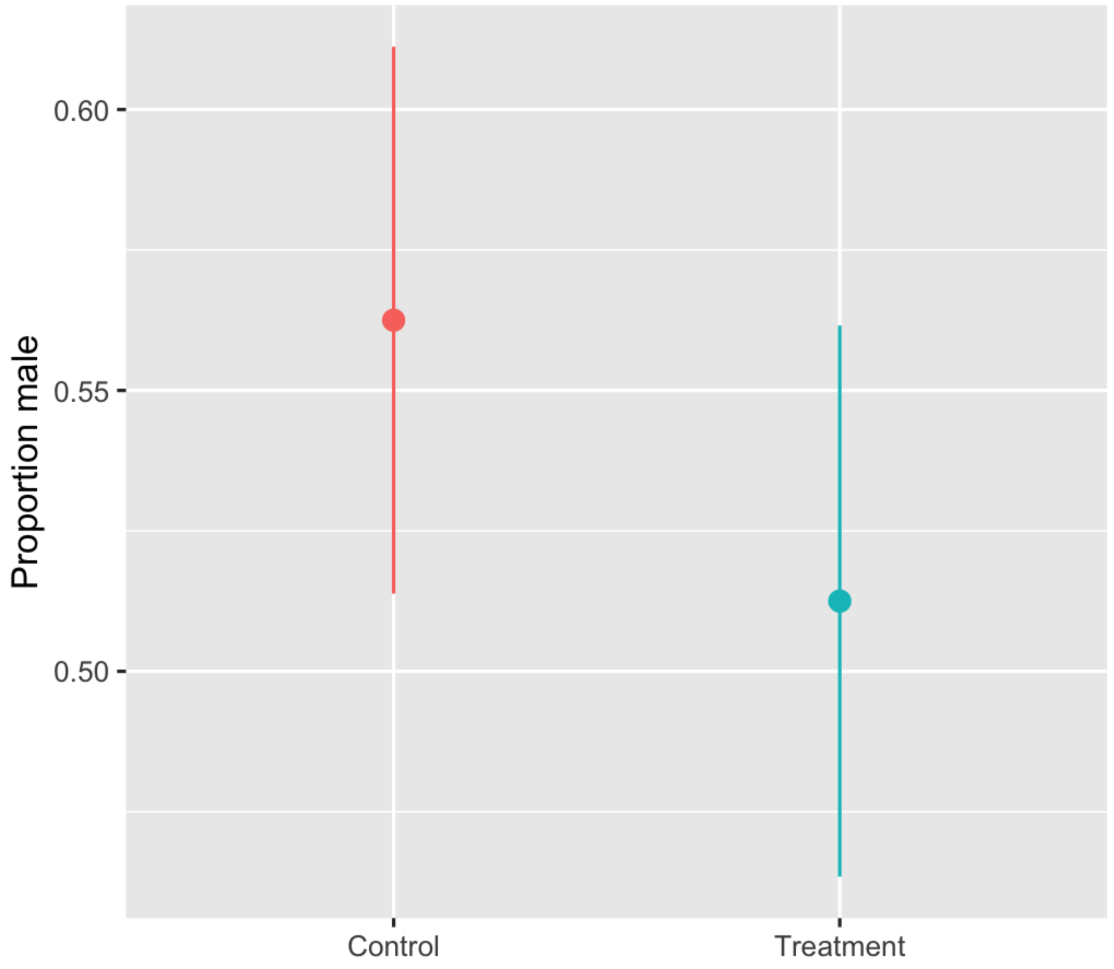
1. Check balance

```
ggplot(imaginary_program,
       aes(x = treatment, y = age,
           color = treatment)) +
  stat_summary(geom = "pointrange",
               fun.data = "mean_se",
               fun.args = list(mult=1.96)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Age")
```



1. Check balance

```r
ggplot(imaginary_program,
       aes(x = treatment, y = male_num,
           color = treatment)) +
  stat_summary(geom = "pointrange",
               fun.data = "mean_se",
               fun.args = list(mult=1.96)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Proportion male")
```



2. Calculate difference

Group means

```
imaginary_program %>%
  group_by(treatment) %>%
  summarize(avg_outcome = mean(income_after))
```

```
## # A tibble: 2 x 2
##   treatment avg_outcome
##   <chr>           <dbl>
## 1 Control          205.
## 2 Treatment        251.
```

```
251 - 205
```

```
## [1] 46
```

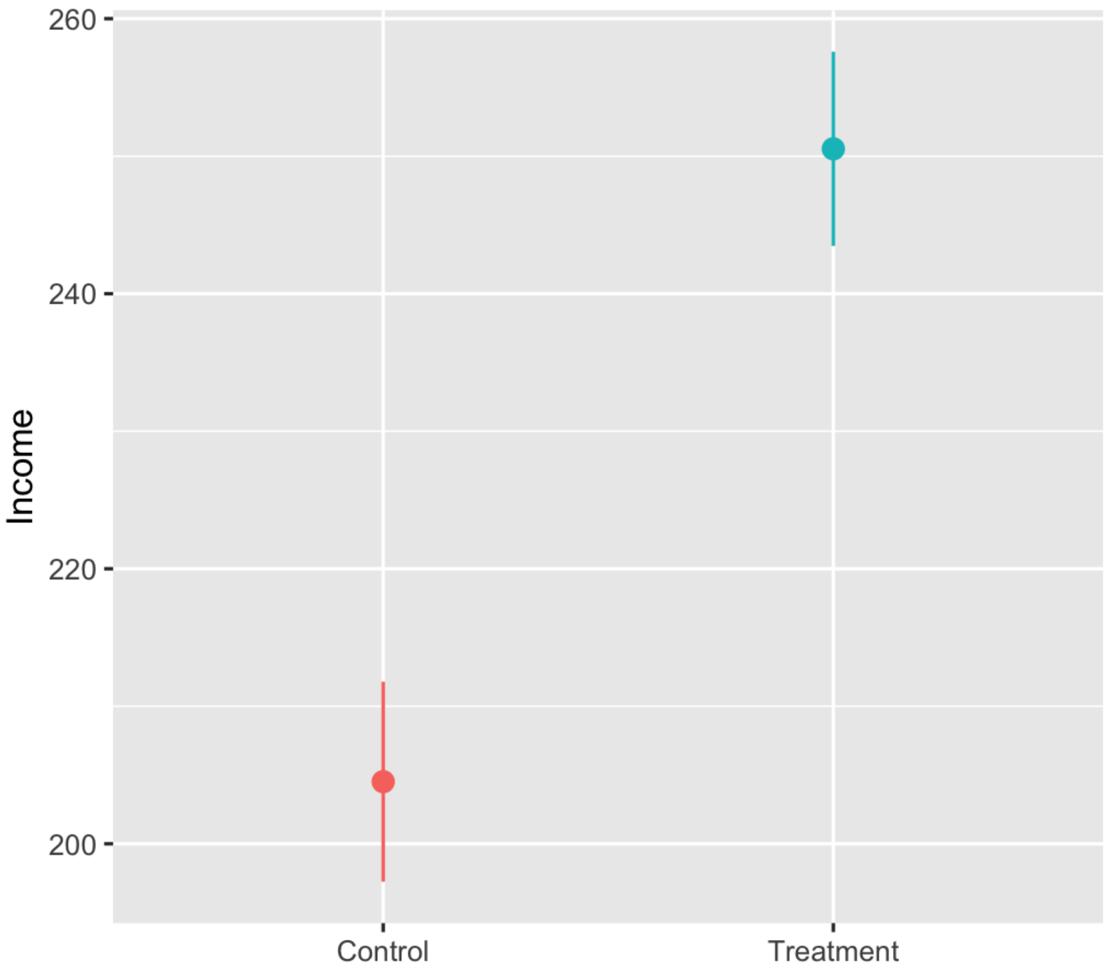
Regression

```
rct_model <- lm(income_after ~ treatment,
                data = imaginary_program)
tidy(rct_model)
```

```
## # A tibble: 2 x 3
##   term               estimate std.error
##   <chr>                 <dbl>     <dbl>
## 1 (Intercept)           205.       3.66
## 2 treatmentTreatment     46.0      5.17
```

2a. Show difference

```
ggplot(imaginary_program,
       aes(x = treatment,
           y = income_after,
           color = treatment)) +
  stat_summary(geom = "pointrange",
               fun.data = "mean_se",
               fun.args = list(mult=1.96)) +
  guides(color = FALSE) +
  labs(x = NULL, y = "Income")
```



Should you control for stuff?

No!

All arrows into the treatment node are removed; there's theoretically no confounding!

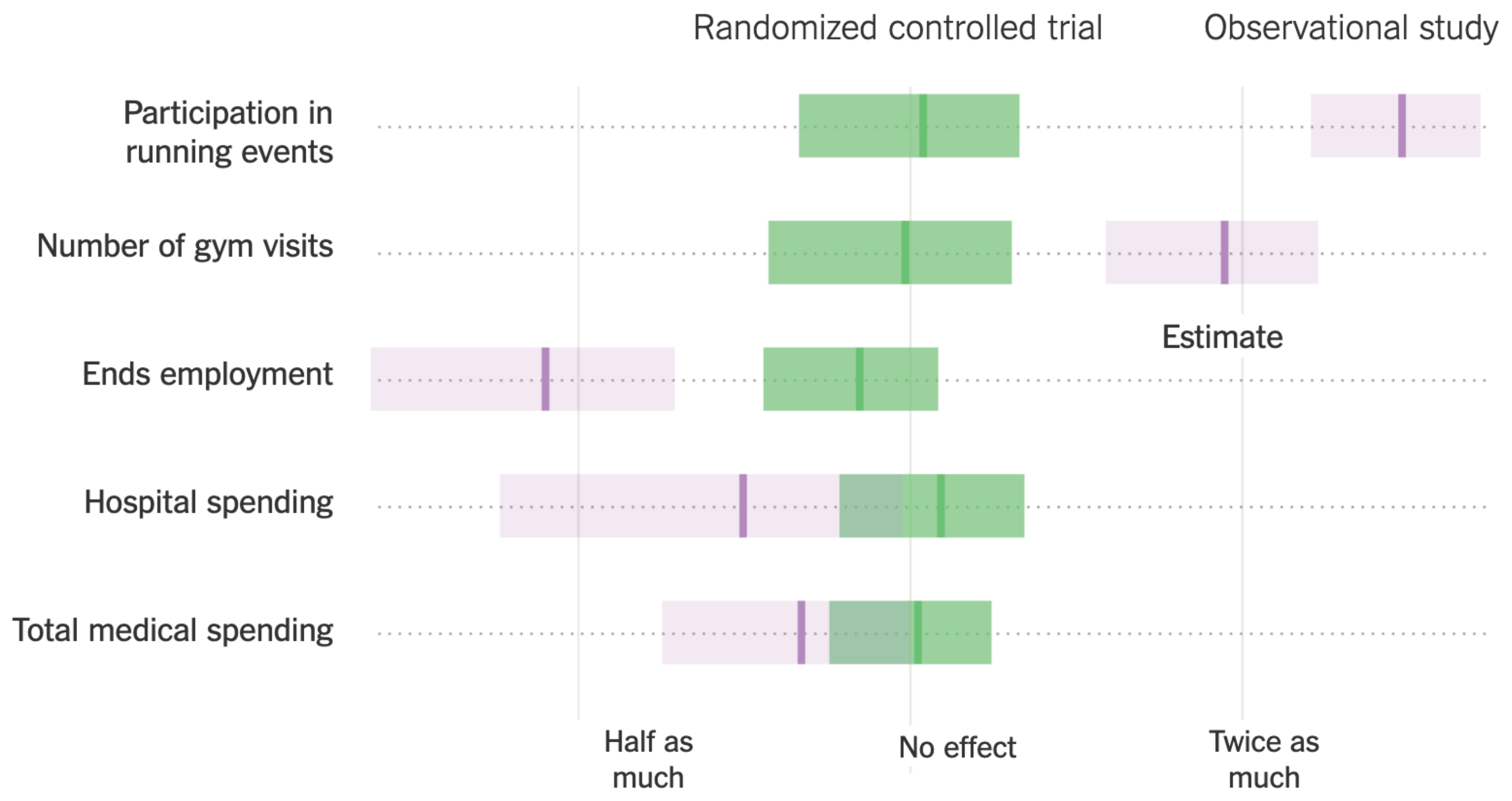
The "gold" standard

Types of research

Experimental studies vs. observational studies

Which is better?

How the Illinois Wellness Program Affected . . .



Source: What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study

Google

rct "gold standard"

All Shopping News V

About 636,000 results (0.67 seconds)

BJOG. Author manuscript; available in PMC 2018 Dec 1.
Published in final edited form as:
BJOG. 2018 Dec; 125(13): 1716.
Published online 2018 Jun 19. doi: 10.1111/1471-0528.15199

PMCID: PMC6235704
NIHMSID: NIHMS966617
PMID: 29916205

Author Manuscript

Randomised controlled trials—the gold standard for effectiveness research

Eduardo Hariton, MD, MBA[1] and Joseph J. Locascio, PhD[2]

▸ Author information ▸ Copyright and License information Disclaimer

The publisher's final edited version of this article is available at BJOG
See other articles in PMC that cite the published article.

Randomized Assignment of Treatment

When a program is assigned at random—that is, using a lottery—over a large eligible population, we can generate a robust estimate of the counterfactual. *Randomized assignment* of treatment is considered the gold standard of impact evaluation. It uses a random process, or chance, to decide who is granted access to the program and who is not.[1] Under randomized assignment, every eligible unit (for example, an individual, household, business, The Washington Post

Democracy Dies in Darkness

Business

3 share Nobel Prize in economics for 'experimental approach' to solving poverty

Esther Duflo, who at 46 is the award's youngest winner, shares the hon
fellow MIT economist Abhijit Banerjee and Harvard's Michael Kremer

J-PAL

ABDUL LATIF JAMEEL POVERTY ACTION LAB

Massachusetts Institute of Technology (MIT) @MIT · 5h

Professors Esther Duflo and Abhijit Banerjee, co-directors of MIT's @JPAL , receive congratulations on the big news this morning. They share in the #NobelPrize in economic sciences "for their experimental approach to alleviating global poverty."

Photo: Bryce Vickmark

12 112 510

RCTs are great!

Super impractical to do all the time though!

"Gold standard"

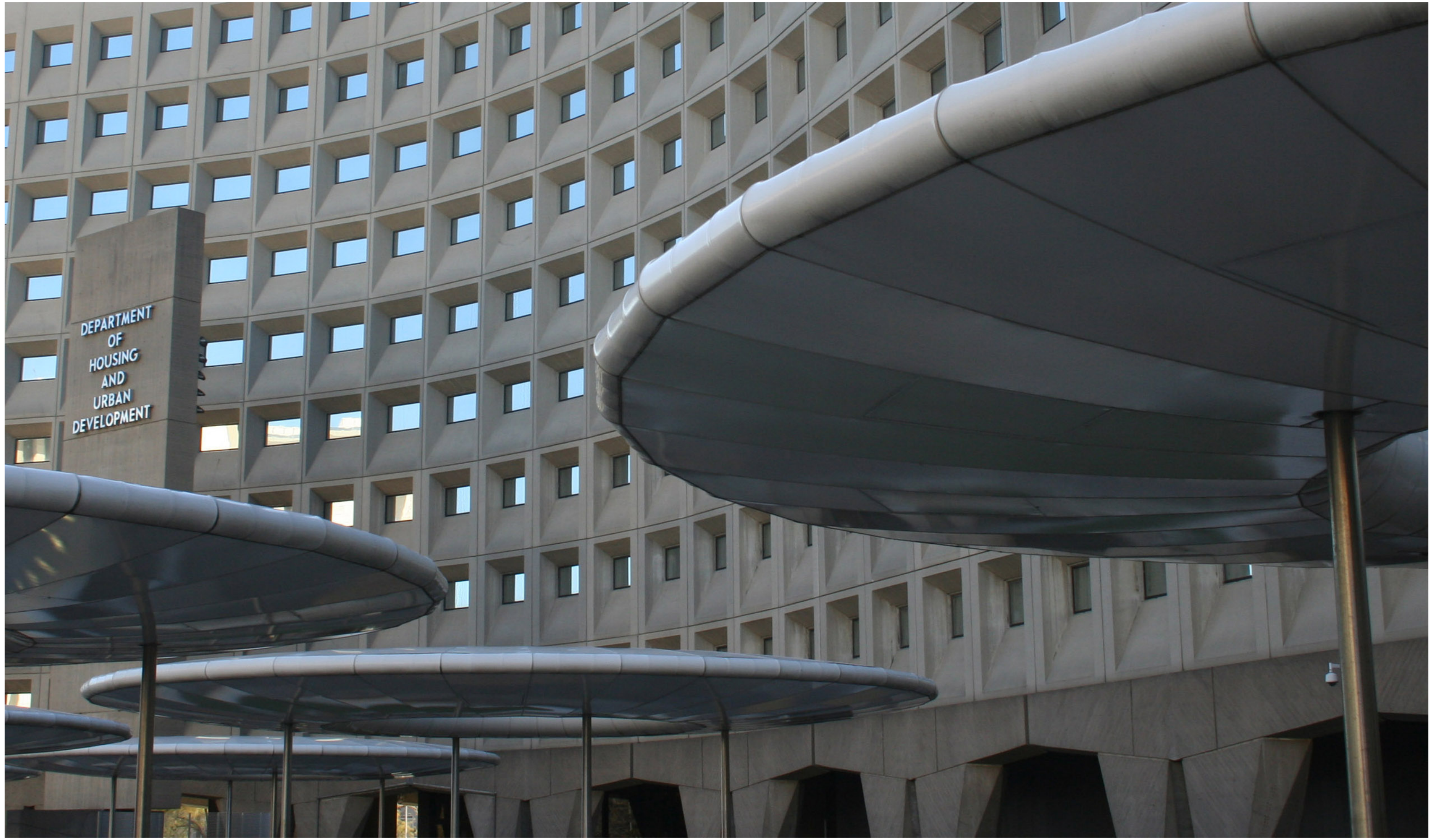
"Gold standard" implies that all causal inferences will be valid it you do the experiment right

We don't care if studies are experimental or not

We care if our causal inferences are valid

RCTs are a helpful baseline/rubric for other methods

Moving to Opportunity



RCTs and validity

Randomization fixes a ton of internal validity issues

Selection
Treatment and control groups are comparable; people don't self-select

Trends
Maturation, secular trends, seasonality, regression to the mean all generally average out

RCTs and validity

RCTs don't fix attrition!

Worst threat to internal validity for RCTs

If attrition is correlated with treatment, that's bad

People might drop out because of the treatment, or because they got/didn't get into the control group

Addressing attrition

Recruit as effectively as possible

You don't just want weird/WEIRD participants

Get people on board

Get participants invested in the experiment

Collect as much baseline information as possible

Check for randomization of attrition

RCTs and validity

Randomization failures

Check baseline pre-data

Noncompliance

Some people assigned to treatment won't take it; some people assigned to control will take it

Intent-to-treat (ITT) vs. Treatment-on-the-treated (TTE)

Other limitations

RCTs don't magically fix construct validity or statistical conclusion validity

RCTs *definitely* don't magically fix external validity

The Nobel Prize in economics goes to three groundbreaking antipoverty researchers

In the last 20 years, development economics has been transformed.
These researchers are the reason why.

By Kelsey Piper | Oct 14, 2019, 3:30pm EDT

Empiricism and development economics

The transformation of development economics into an intensely empirical field that leans heavily on randomized controlled trials hasn't been uncontroversial, and many of **the responses** to the Nobel Prize announcement acknowledge that controversy.

Critics have **complained that** randomization feels much more scientific than other approaches but doesn't necessarily answer our questions any more definitively. **Others worry** that the focus on small-scale questions — Do wristbands increase vaccination rates? Do textbooks improve school performance? — might distract us from addressing larger, structural contributors to poverty.

When to randomly assign

Demand for treatment exceeds supply

Treatment will be phased in over time

Treatment is in equipoise (genuine uncertainty)

Local culture open to randomization

When you're a nondemocratic monopolist

When people won't know (and it's ethical!)

When lotteries are going to happen anyway

When to not randomly assign

When you need immediate results

When it's unethical or illegal

When it's something that happened in the past

When it involves universal ongoing phenomena

Adjustment with matching

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

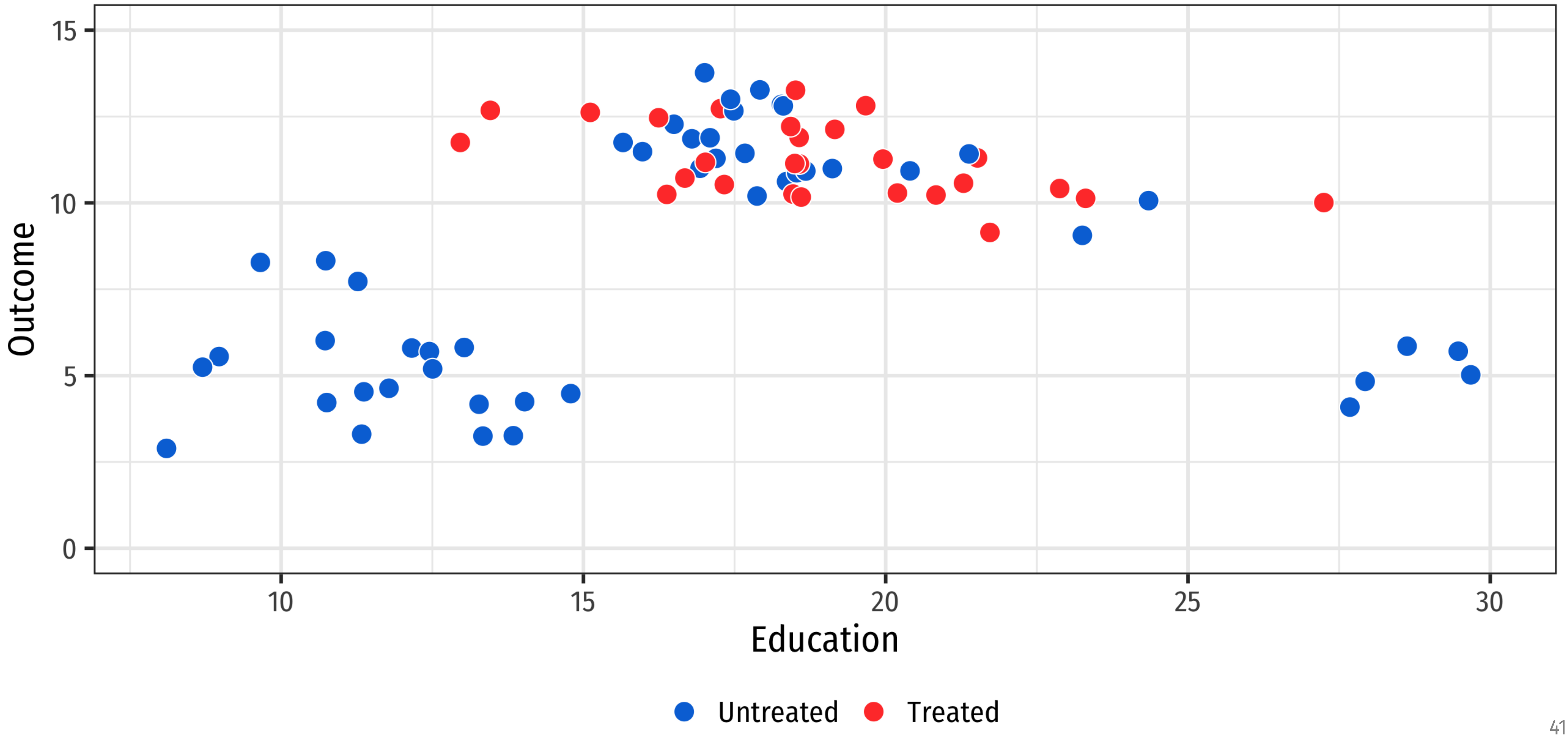
Why match?

Reduce model dependence

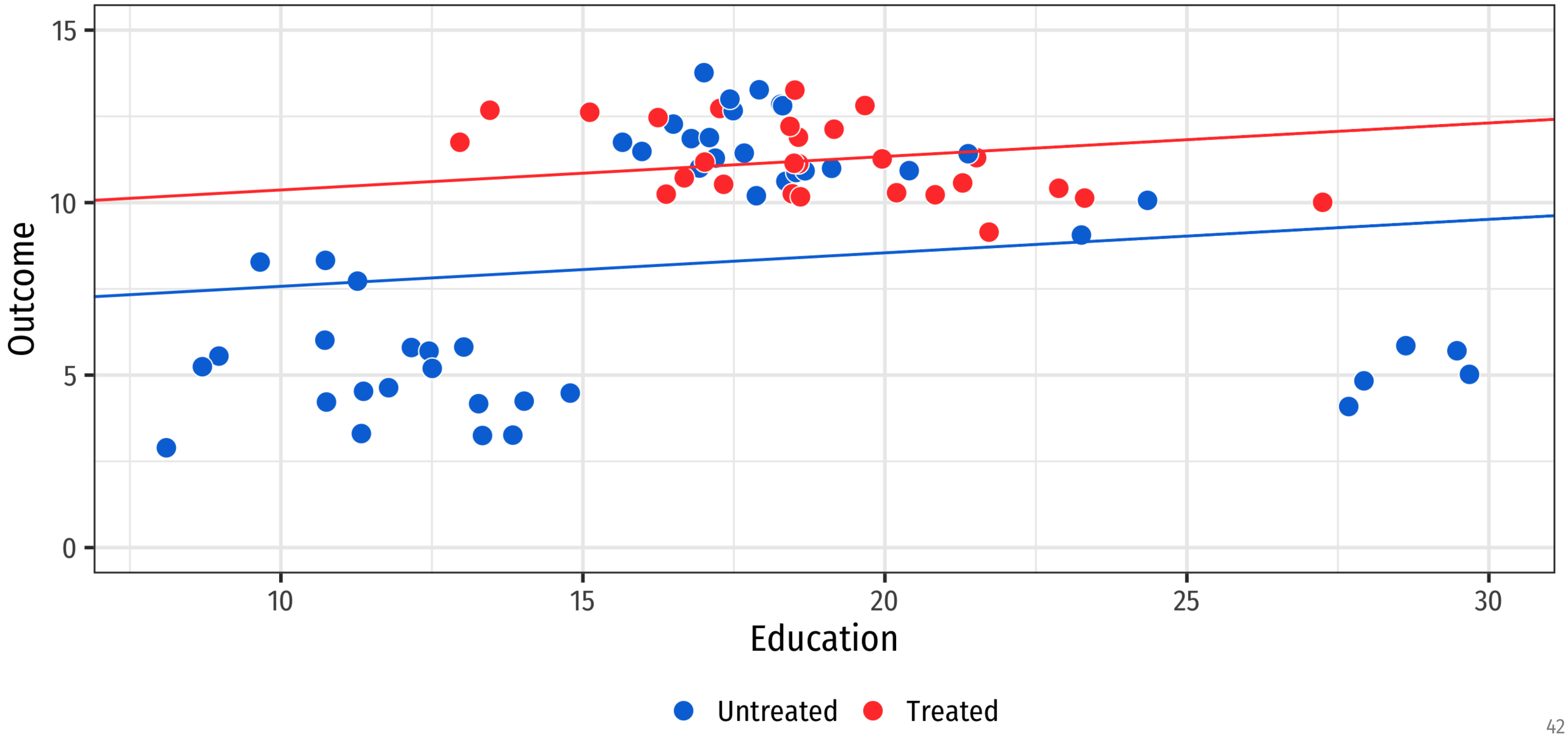
Imbalance → model dependence → researcher discretion → bias

Compare apples to apples

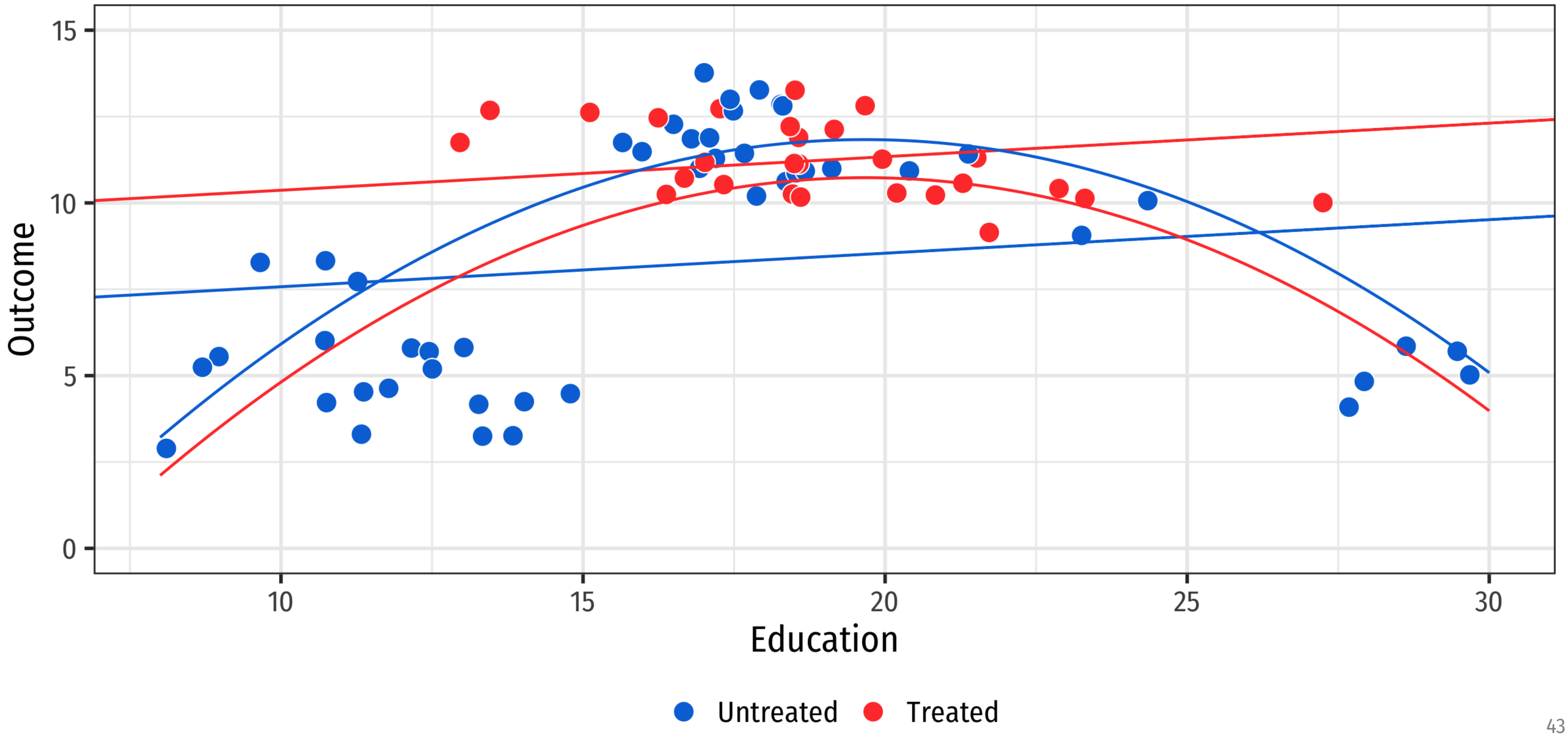
It's a way to adjust for backdoors!

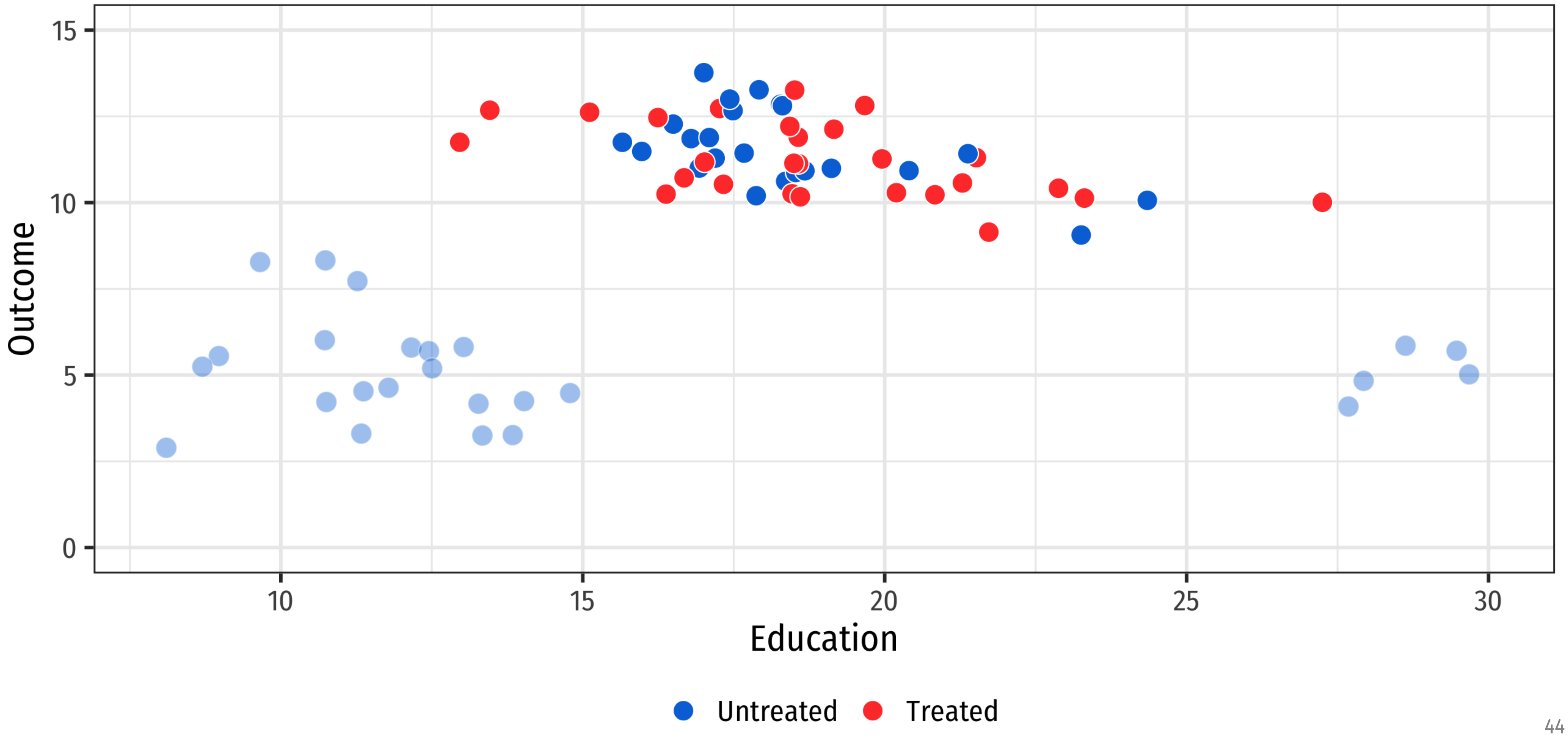

15
10
5
0
Outcome
10
15
20
25
30
Education
Untreated
Treated

$$\text{Outcome} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Treatment}$$

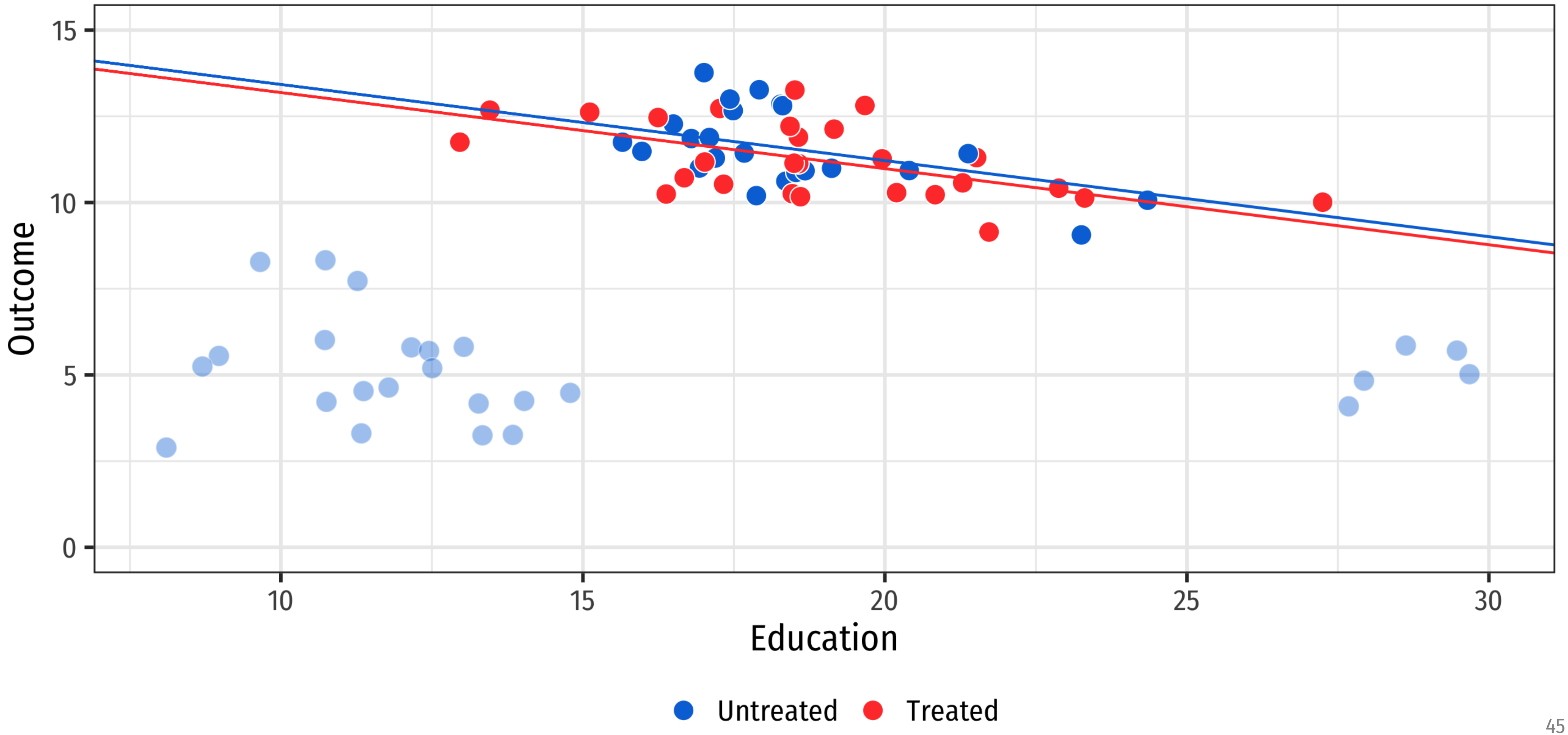


$$\text{Outcome} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Education}^2 + \beta_3\text{Treatment}$$

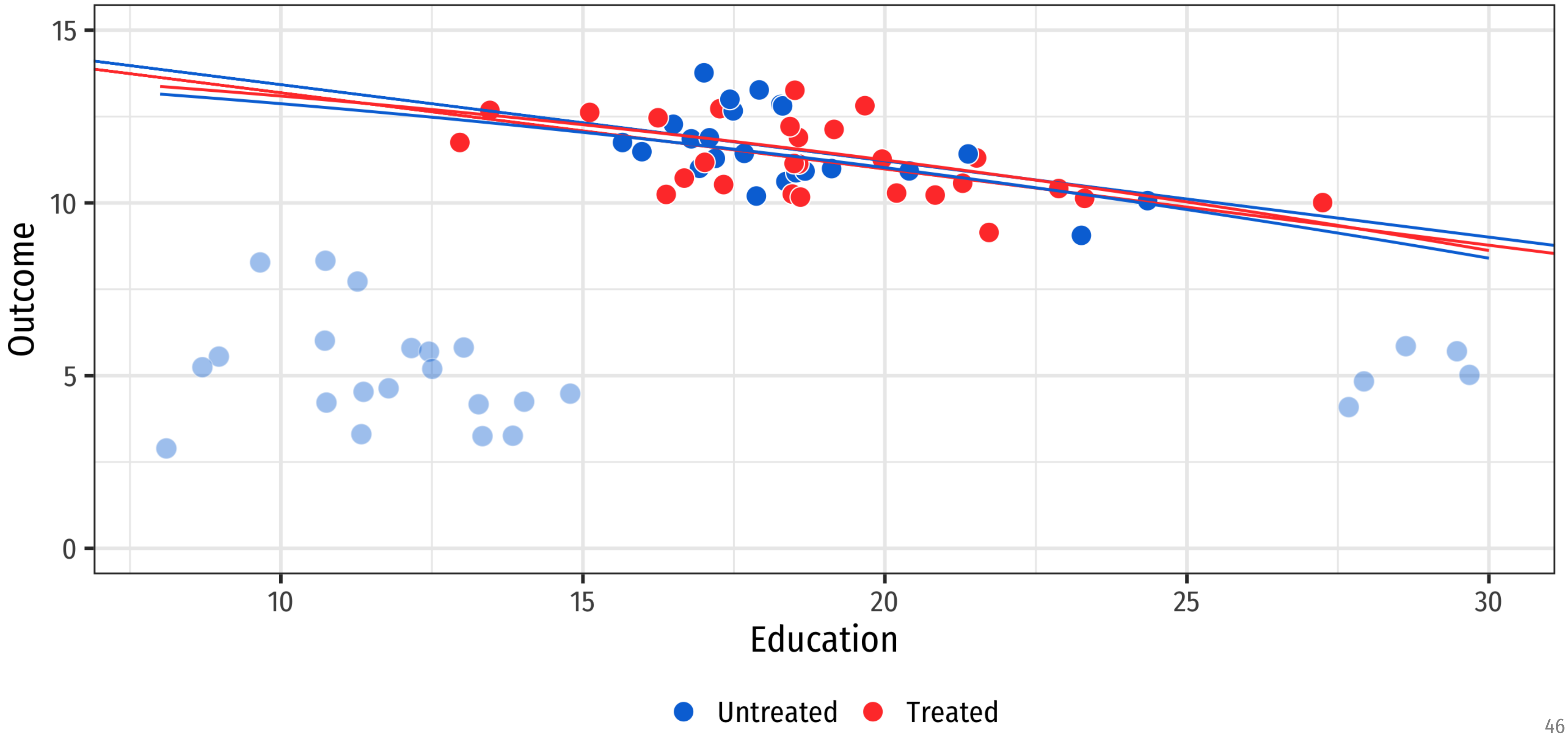



15
10
5
0
Outcome
10
15
20
25
30
Education
Untreated
Treated

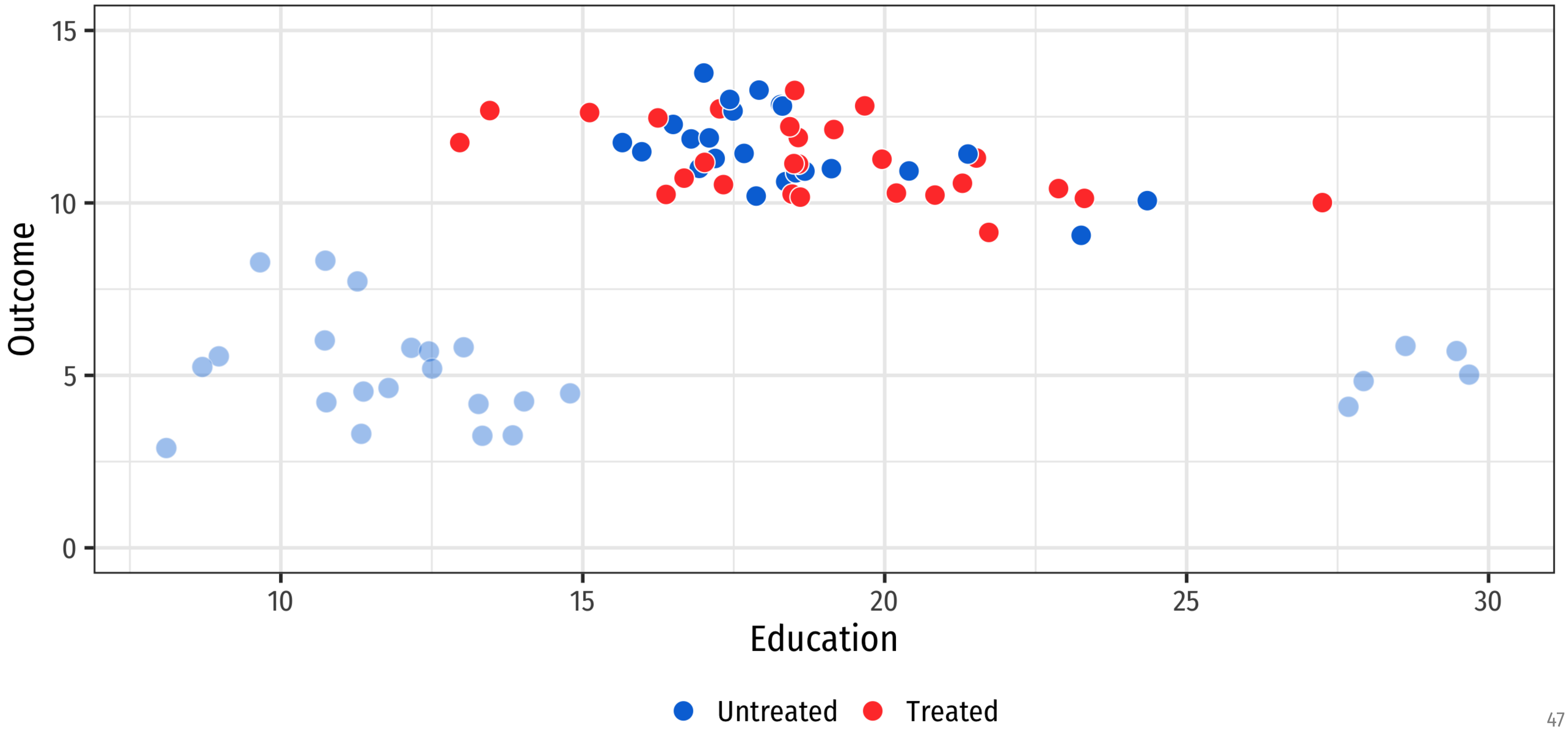
$$\text{Outcome} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Treatment}$$



$$\text{Outcome} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Education}^2 + \beta_3 \text{Treatment}$$



How do we know that we can remove these points?



General process for matching

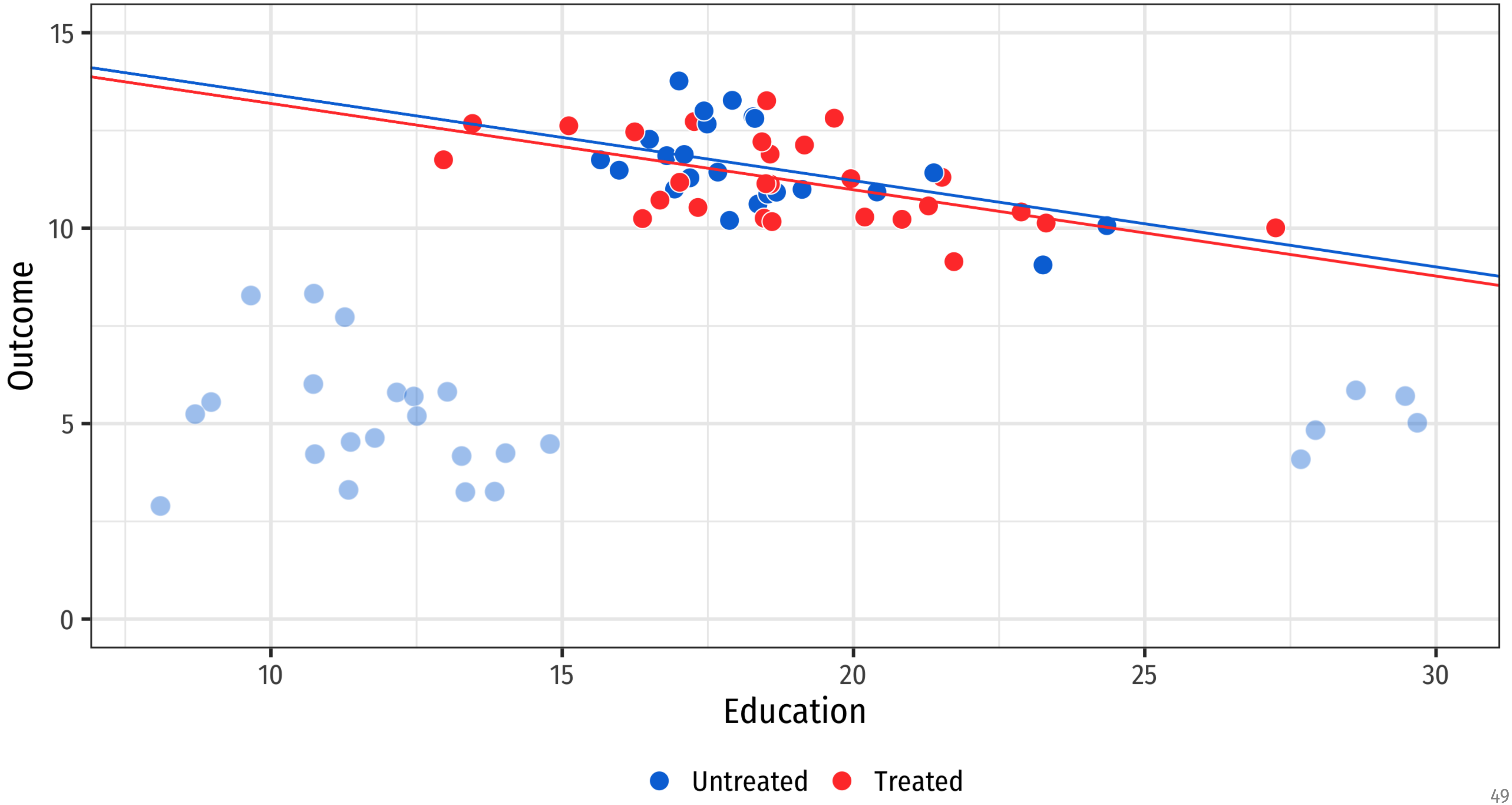
Step 1. Preprocessing

Do something to guess or model the assignment to treatment

Use what you know about the DAG to inform this guessing!

Step 2. Estimation

Use the new trimmed/preprocessed data to build a model, calculate difference in means, etc.


15
10
5
0
Outcome
10
15
20
25
30
Education
Untreated
Treated

Different methods

Nearest neighbor matching (NN)

Mahalanobis distance / Euclidean distance

~~**Propensity score matching (PSM)**~~

Inverse probability weighting (IPW)

(and lots of other methods we're not covering!)

Nearest neighbor matching

Find untreated observations that are very close/similar to treated observations based on confounders

Lots of mathy ways to measure distance

Mahalanobis and Euclidean distance are fairly common

US MARKETS

There’s a 70% chance of recession in the next six months, new study from MIT and State Street finds

PUBLISHED WED, FEB 5 2020•12:20 PM EST | UPDATED WED, FEB 5 2020•4:13 PM EST

Pippa Stevens
@PIPPASTEVENS13

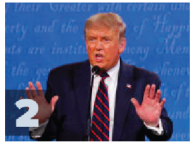
SHARE

KEY POINTS

- A new study from the MIT Sloan School of Management and State Street Associate says there’s a 70% chance that a recession will occur in the next six months.
- The researches used a scientific approach initially developed to measure human skulls to determine how the relationship of four factors compares to prior recessions.
- The index currently stands at 76%. Looking at data back to 1916, the researchers found that once the index topped 70%, the likelihood of a recession rose to 70%.

TRENDING NOW

House passes $2.2 trillion Democratic coronavirus stimulus bill



Trump suggests he won’t ‘allow’ rule changes for next debates with Biden

Top Trump aide Hicks tests positive for coronavirus after traveling with president

US MARKETS

There's a 70% chance of recession in the next six months, new study from MIT and State Street finds

PUBLISHED WED, FEB 5 2020•12:20 PM EST | UPDATED WED, FEB 5 2020•4:13 PM EST

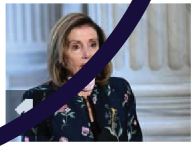
Pippa Stevens
@PIPPASTEVENS13

SHARE

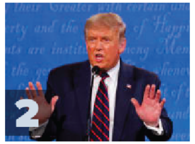
That's just Mahalanobis matching!

KEY POINTS

- A new study from the MIT Sloan School of Management and State Street Associate says there's a 70% chance that a recession will occur in the next six months.
- The researches used a scientific approach initially developed to measure human skulls to determine how the relationship of four factors compares to prior recessions.
- The index currently stands at 76%. Looking at data back to 1916, the researchers found that once the index topped 70%, the likelihood of a recession rose to 70%.



House passes $2.2 trillion Democratic coronavirus stimulus bill



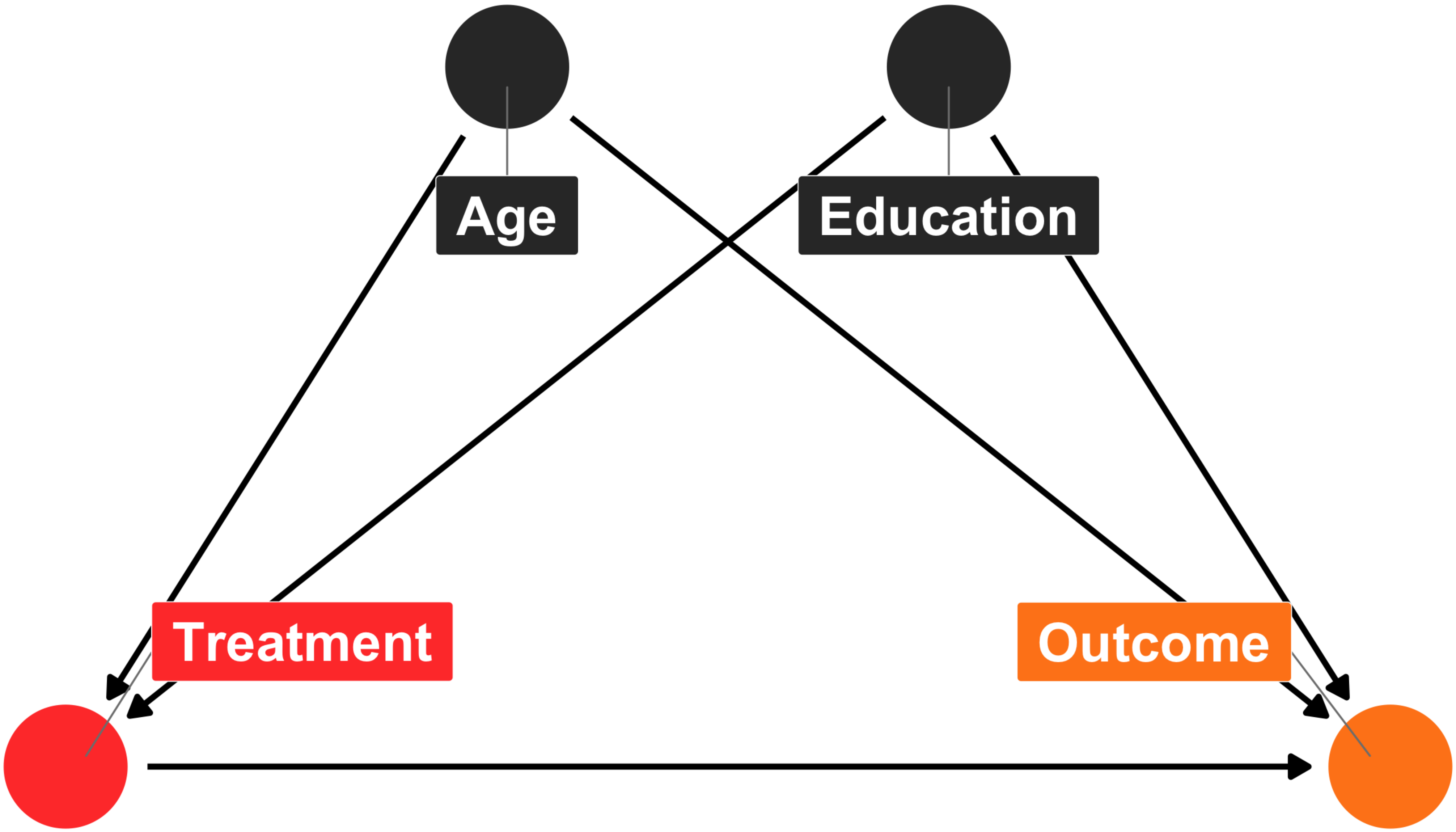
Trump suggests he won't 'allow' rule changes for next debates with Biden

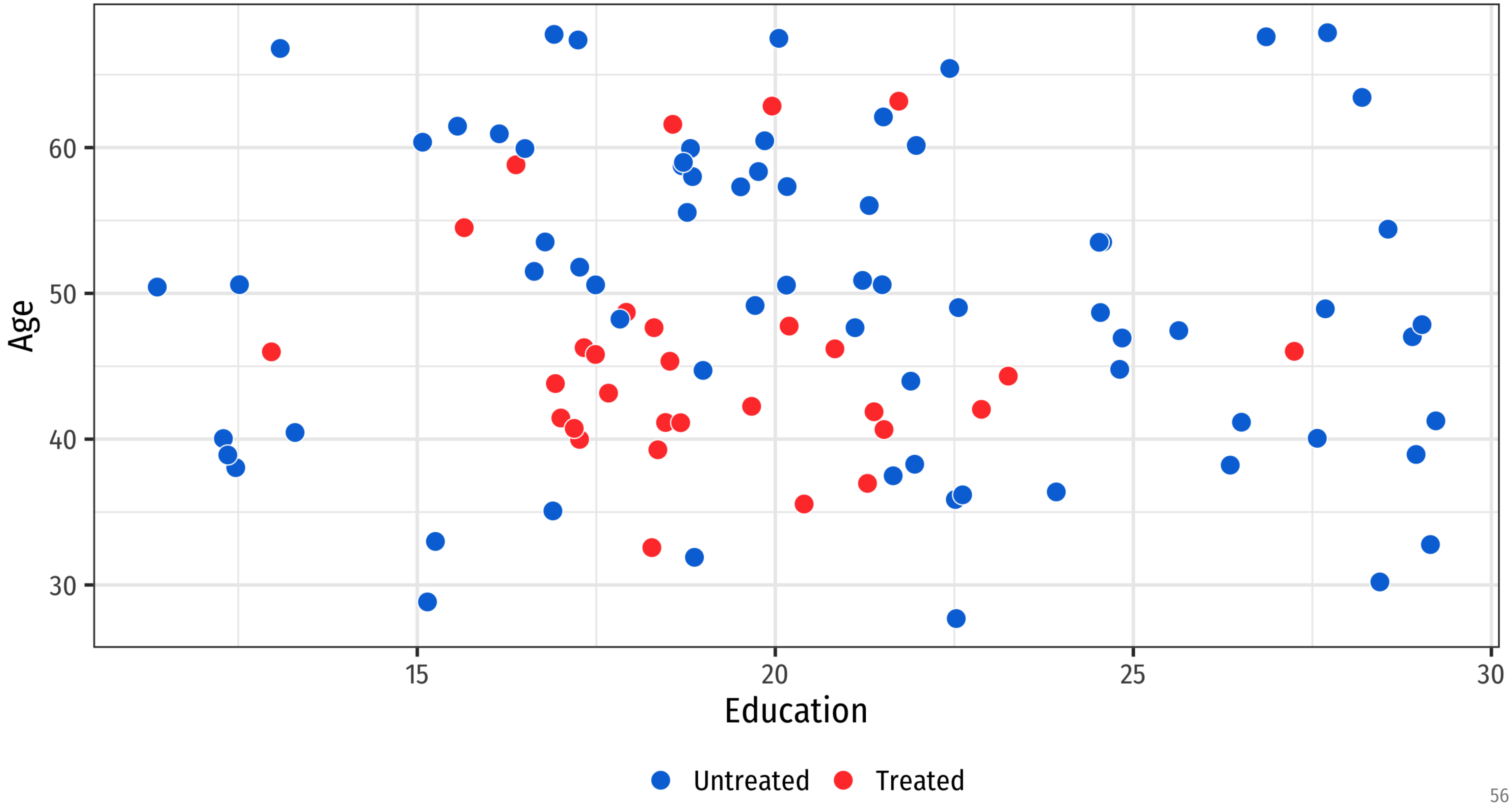
Top Trump aide Hicks tests positive for coronavirus after traveling with president

Matching and eugenics

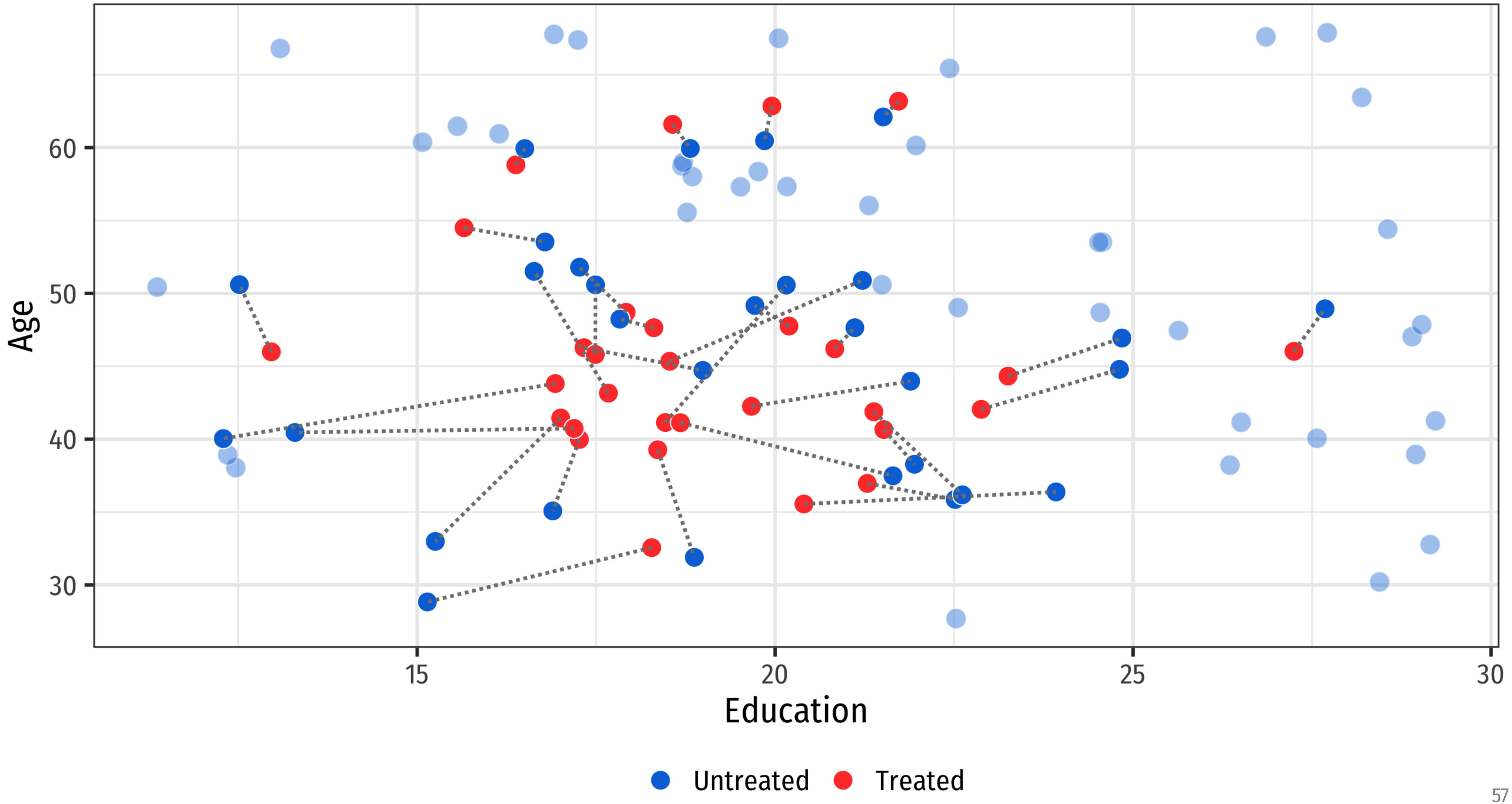
Prasanta Chandra Mahalanobis

Tried to prove brain size differences between castes; low-key eugenicist

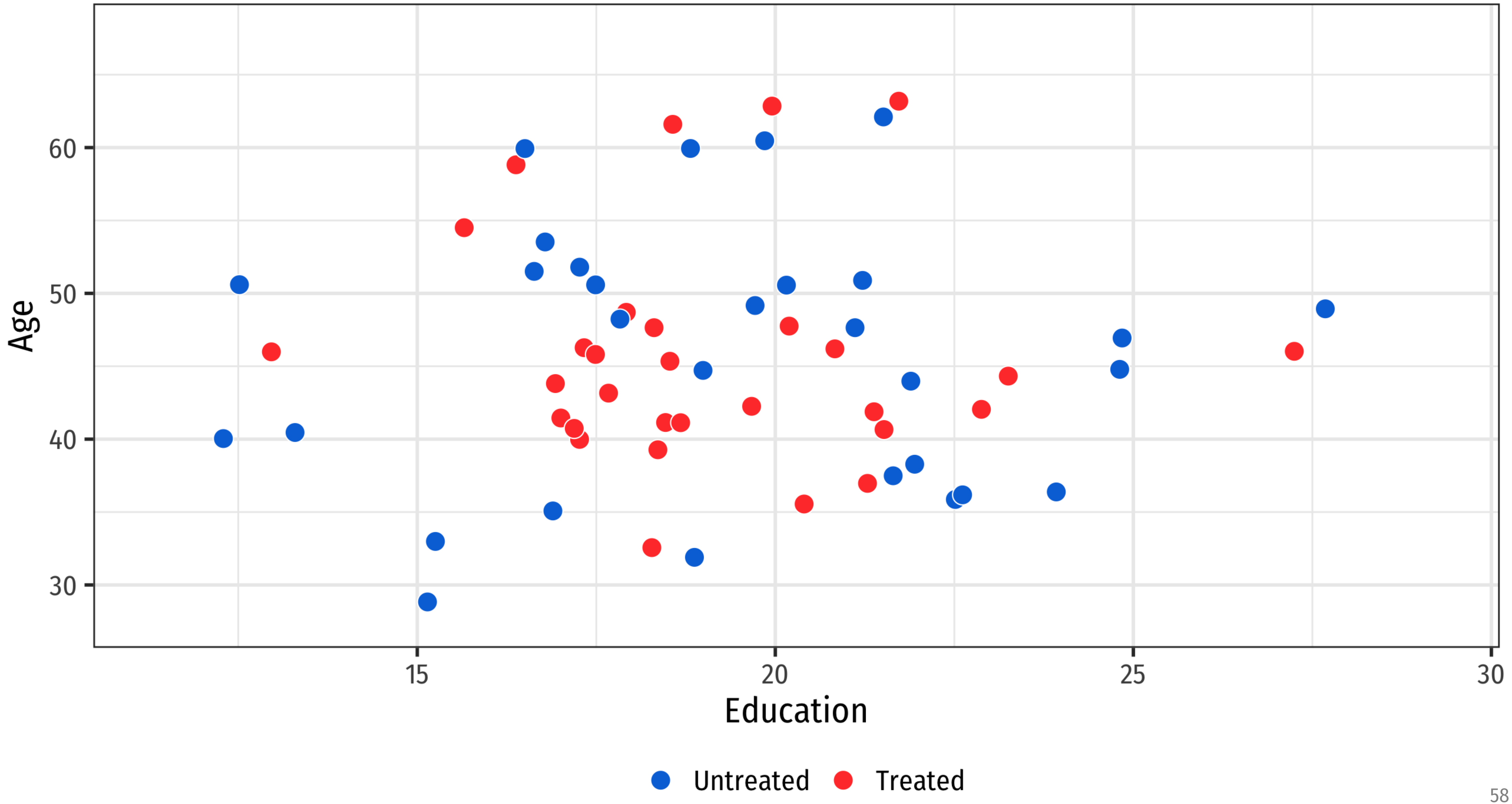

Age
Education
Treatment
Outcome



Age
Education
60
50
40
30
15
20
25
30
Untreated
Treated

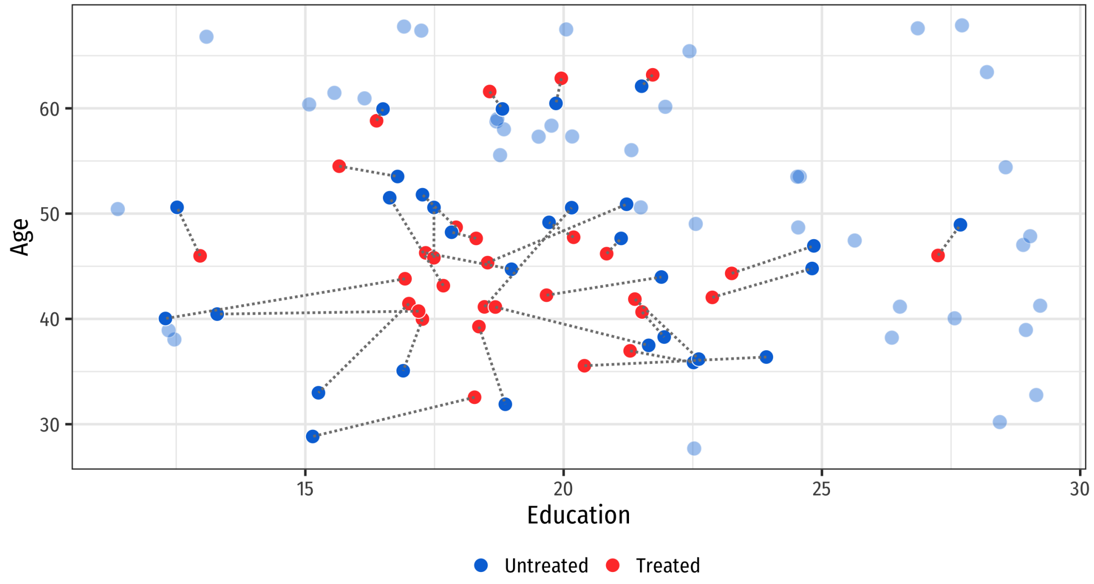


Age
60
50
40
30
15
20
25
30
Education
Untreated
Treated


Age
60
50
40
30
15
20
25
30
Education
Untreated
Treated

Potential problems with matching

Nearest neighbor matching can be greedy!



Solution: Don't throw everything away!

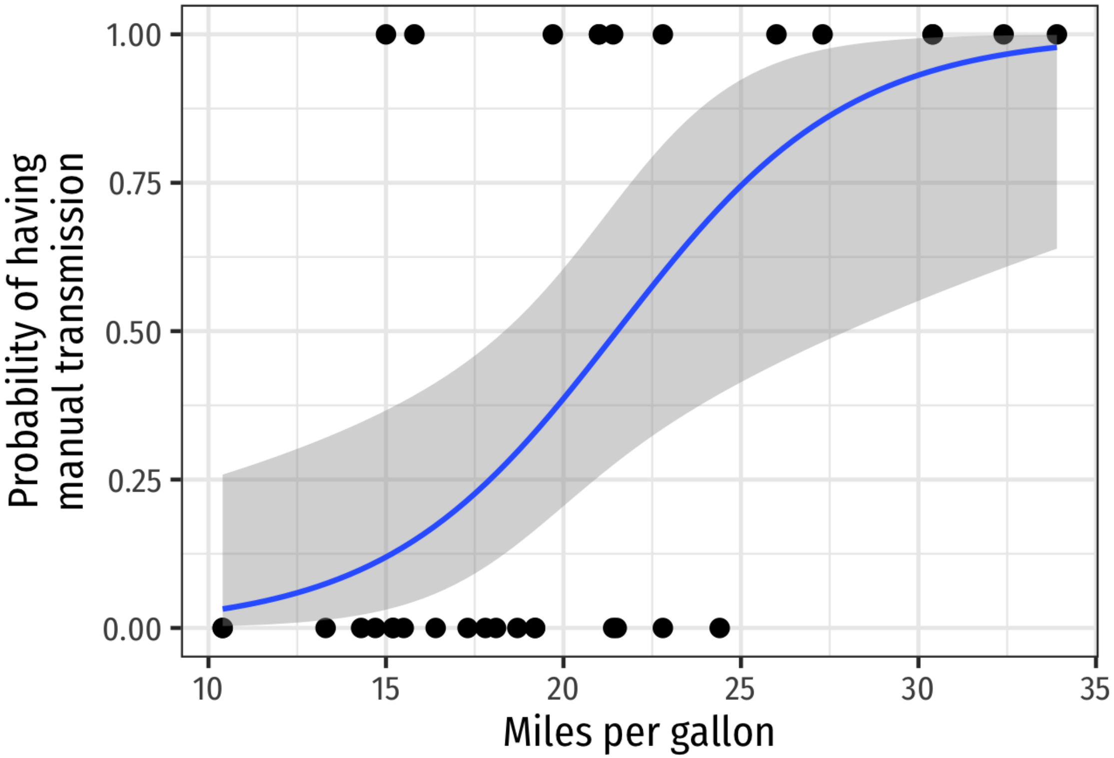
Propensity scores

Predict the probability of assignment to treatment using a model

Logistic regression, probit regression, machine learning, etc.

Here's logistic regression:

$$\log \frac{p_{\text{Treated}}}{1 - p_{\text{Treated}}} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Age}$$



$$\log \frac{p_{\text{Manual}}}{1 - p_{\text{Manual}}} = \beta_0 + \beta_1 \text{MPG}$$

```
model_transmission <- glm(am ~ mpg, data = mtcars, family = binomial(link = "logit"))
```

Log odds (default coefficient unit of measurement; fairly uninterpretable)

Odds ratios (e^{β}; centered around 1: 1.5 means 50% more likely; 0.75 means 25% less likely)

```r
tidy(model_transmission)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -6.60      2.35      -2.81 0.00498
## 2 mpg            0.307     0.115      2.67 0.00751
```

```r
tidy(model_transmission,
     exponentiate = TRUE)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  0.00136     2.35      -2.81 0.00498
## 2 mpg          1.36        0.115      2.67 0.00751
```

Plug all the values of MPG into the model and find the predicted probability of manual transmission

```r
augment(model_transmission, data = mtcars, type.predict = "response")
```

```
## # A tibble: 32 x 3
##       mpg    am .fitted
##     <dbl> <dbl>   <dbl>
##  1  21        1  0.461
##  2  21        1  0.461
##  3  22.8      1  0.598
##  4  21.4      0  0.492
##  5  18.7      0  0.297
##  6  18.1      0  0.260
##  7  14.3      0  0.0986
##  8  24.4      0  0.708
##  9  22.8      0  0.598
## 10  19.2      0  0.330
## # … with 22 more rows
```

Row 7 is highly unlikely to be manual (1)

Row 8 is highly likely to be manual

Propensity score matching

Super popular method

There are mathy reasons why it's not great for matching *for identification purposes*

**Propensity scores are fine!
Using them for matching isn't!**

Why Propensity Scores Should Not Be Used for Matching

Gary King[1] and Richard Nielsen[2]

[1] Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: king@harvard.edu, URL: http://GaryKing.org
[2] Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. Email: rnielsen@mit.edu, URL: http://www.mit.edu/~rnielsen

Abstract

We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias. The weakness of PSM comes from its attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment. PSM is thus uniquely blind to the often large portion of imbalance that can be eliminated by approximating full blocking with other matching methods. Moreover, in data balanced enough to approximate complete randomization, either to begin with or after pruning some observations, PSM approximates random matching which, we show, increases imbalance even relative to the original data. Although these results suggest researchers replace PSM with one of the other available matching methods, propensity scores have other productive uses.

Keywords: matching, propensity score matching, coarsened exact matching, Mahalanobis distance matching, model dependence

Weighting

Make some observations more important than others

	Young	Middle	Old
Population	30%	40%	30%
Sample	60%	30%	10%

Weighting

Make some observations more important than others

	Young	**Middle**	**Old**
Population	30%	40%	30%
Sample	60%	30%	10%
Weight	30 / 60 **0.5**	40 / 30 **1.333**	30 / 10 **3**

Multiply weights by average values
(or us in regression) to adjust for importance

Inverse probability weighting

Use propensity scores to weight observations by how "weird" they are

Observations with high probability of treatment who don't get it (and vice versa) have higher weight

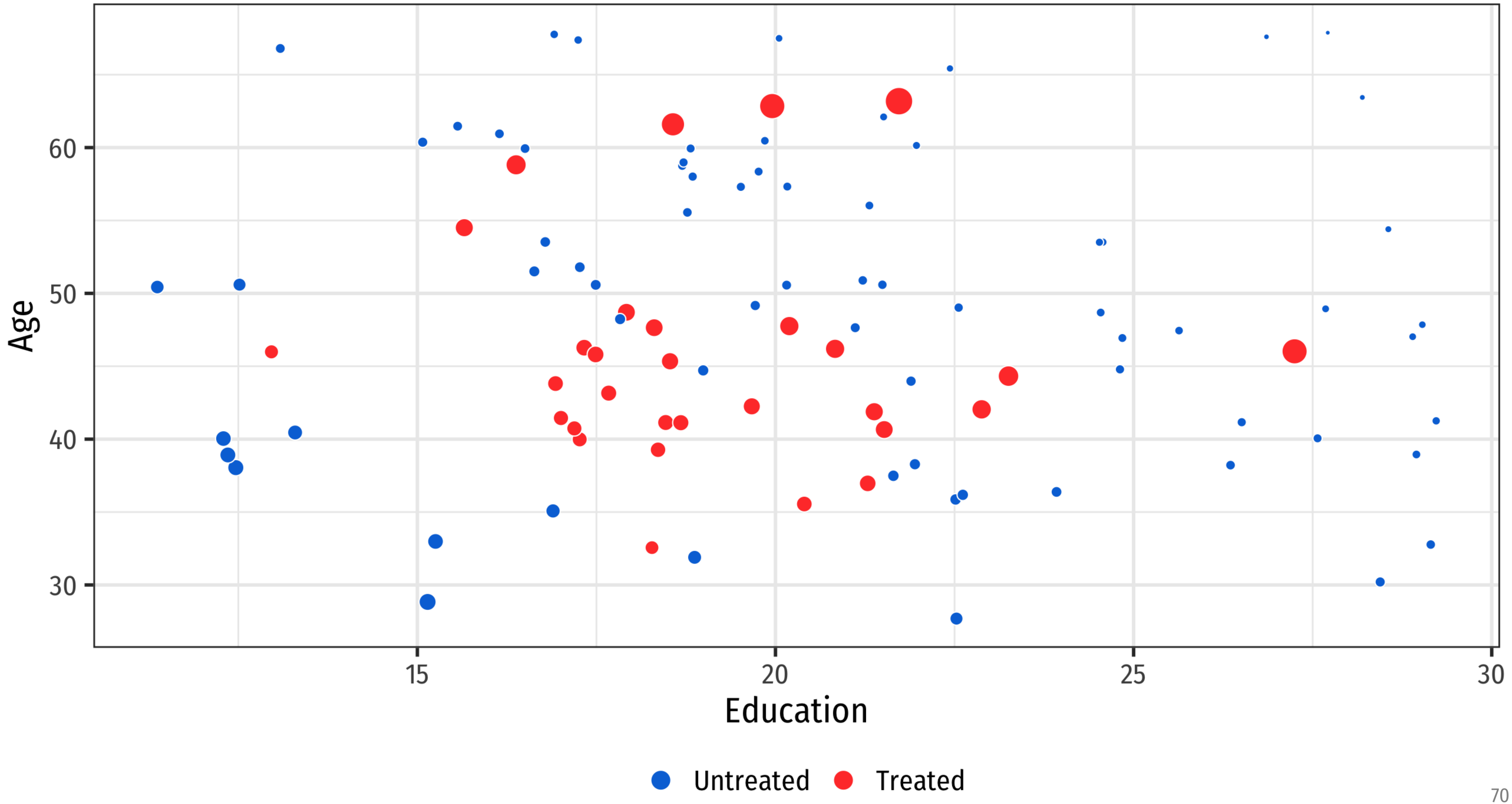
$$\frac{\text{Treatment}}{\text{Propensity}} + \frac{1 - \text{Treatment}}{1 - \text{Propensity}}$$

```
augment(model_transmission, data = mtcars, type.predict = "response") %>%
  select(mpg, am, propensity = .fitted) %>%
   mutate(ip_weight = (am / propensity) + ((1 - am) / (1 - propensity)))
```

```
## # A tibble: 32 x 4
##      mpg    am propensity ip_weight
##    <dbl> <dbl>      <dbl>     <dbl>
##  1  21       1     0.461       2.17
##  2  21       1     0.461       2.17
##  3  22.8     1     0.598       1.67
##  4  21.4     0     0.492       1.97
##  5  18.7     0     0.297       1.42
##  6  18.1     0     0.260       1.35
##  7  14.3     0     0.0986      1.11
##  8  24.4     0     0.708       3.43
##  9  22.8     0     0.598       2.49
## 10  19.2     0     0.330       1.49
### … with 22 more rows
```

Row 7 is highly unlikely to be manual and isn't.
Boring! Low IPW.

Row 8 is highly likely to be manual, but isn't.
That's weird! High IPW.



60
50
40
30
Age
15
20
25
30
Education
Untreated
Treated

Examples!