STRATIFIED ANALYSIS: PROPENSITY SCORES

Barbra Dickerman, Joy Shi, Miguel Hernán
DEPARTMENTS OF EPIDEMIOLOGY AND BIOSTATISTICS

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Learning objectives
At the end of this lecture you will be able to

- Define and estimate propensity scores
- Use estimated propensity scores to estimate conditional effects using stratification and regression
- Understand the relative advantages and disadvantages of propensity score matching

☐ Key concepts

- Propensity score
- Stratification, regression, matching
- Bias-variance tradeoff for propensity score estimation

Study population

- 1629 cigarette smokers
- Aged 25-74 years when interviewed in 1971-75 (baseline)
- Interviewed again in 1982
- Known sex, age, race, weight, height, education, alcohol use, and smoking intensity at both baseline and follow-up visits, and who answered the general medical history questionnaire at baseline

Key variables

Treatment A	Quit smoking between baseline and 1982 1: yes, 0: no
Continuous outcome Y	Weight gain, kg Weight in 1982 minus baseline weight Available for 1566 individuals
Dichotomous outcome D	Death by 1992 1: yes, 0: no
Baseline (pre-treatment) covariates	Age, sex, race, alcohol use, intensity of smoking, weight…

The causal questions of interest (informal version)

What is the effect of smoking cessation on

1. weight gain?
2. death?

Last time we answered causal question #1 using

- Outcome regression
  - A stratification-based method

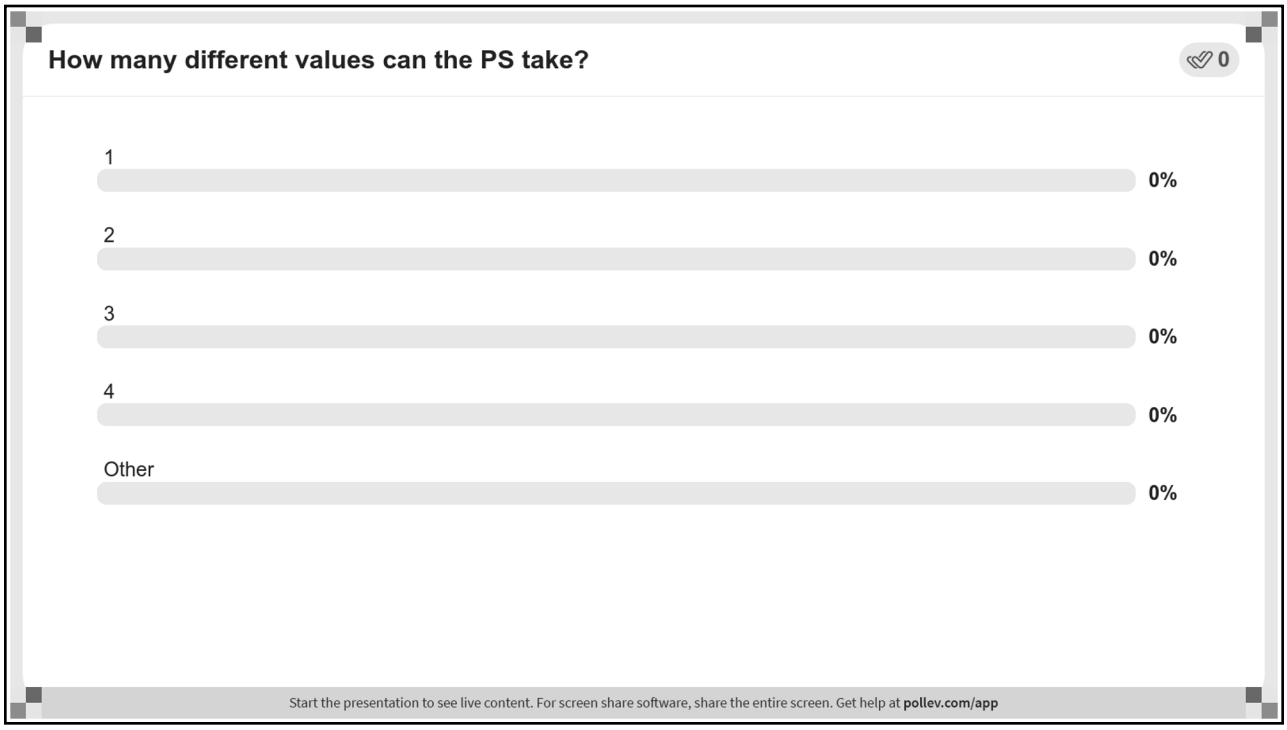
- Today we describe other stratification-based methods based on the "propensity score"

Propensity score (PS)

- Probability of receiving treatment $A=1$ conditional on the confounders L
  - $\text{PS} = \Pr[A=1|L]$
- In our example, an individual's propensity score is the probability that they quit smoking given their confounder values
  - This probability needs to be estimated
- First we describe *how* to estimate the PS
- Second we discuss *why* the PS is helpful

1. Estimation of the PS

- Suppose L includes only two dichotomous variables
  - L_1 – Sex: Women (1), Men (0)
  - L_2 – Age: older than 50 (1), 50 or less (0)
- There are 4 strata:
  - Younger men ($L_1=0, L_2=0$)
  - Older men ($L_1=0, L_2=1$)
  - Younger women ($L_1=1, L_2=0$)
  - Older women ($L_1=1, L_2=1$)



Nonparametric estimation of PS

Sample proportion of quitters by age, sex

- A consistent estimator of the PS is the proportion of quitters in each stratum
- Younger men (L_1=0, L_2=0)
  - 132/515 = 0.256
- Older men (L_1=0, L_2=1)
  - 88/247 = 0.356
- Younger women (L_1=1, L_2=0)
  - 115/583 = 0.197
- Older women (L_1=1, L_2=1)
  - 68/221 = 0.308

See 2.2_propensity.R, lines 10-14

Nonparametric estimation of PS
Saturated logistic regression model

$$\text{logit}\ \Pr[A=1|\, L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_1 L_2$$

- Interpretation of parameters
  - PS in younger men
    - $\Pr[A=1|\, L_1=0, L_2=0] = \text{expit}\ (\alpha_0)$
  - PS in older men
    - $\Pr[A=1|\, L_1=0, L_2=1] = \text{expit}\ (\alpha_0 + \alpha_2)$

Reminder

Let $\Pr[A=1|L] = p$

- Logit transformation
  - $\text{logit}\ (p) = \log [\, p / (1-p)\,] = x$
- Expit transformation
  - inverse logit transformation
  - $p = \text{expit}\ (x) = \exp(x) / [1+ \exp(x)]$

PS in younger women, $\Pr[A = 1|L_1 = 1, L_2 = 0]$**?**

0

Model: $\text{logit}\ \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_1 L_2$

$\text{expit}(\alpha_0)$ 0%

$\text{expit}(\alpha_0 + \alpha_1)$ 0%

$\text{expit}(\alpha_0 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3)$ 0%

PS in older women, $\Pr[A = 1|L_1 = 1, L_2 = 1]$**?**

0

Model: $\text{logit}\ \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_1 L_2$

$\text{expit}(\alpha_0)$ 0%

$\text{expit}(\alpha_0 + \alpha_1)$ 0%

$\text{expit}(\alpha_0 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3)$ 0%

Nonparametric estimation of PS

Saturated logistic regression model

$$\text{logit Pr}[A=1|\, L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_1 L_2$$

- Parameter estimates
  - $\hat{\alpha}_0$ = -1.07
  - $\hat{\alpha}_1$ = -0.34
  - $\hat{\alpha}_2$ = 0.47
  - $\hat{\alpha}_3$ = 0.12

See 2.2_propensity.R, lines 18-19

- After applying the expit transformation, these parameter estimates result in the same PS estimates obtained via sample proportions

Parametric estimation of PS

Nonsaturated logistic regression model

$$\text{logit Pr}[A=1|\, L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$$

- Interpretation of parameters
  - PS in younger men
    - $\Pr[A=1|\, L_1=0, L_2=0] = \text{expit}(\alpha_0)$
  - PS in older men
    - $\Pr[A=1|\, L_1=0, L_2=1] = \text{expit}(\alpha_0 + \alpha_2)$

PS in younger women, $\Pr[A = 1|L_1 = 1, L_2 = 0]$**?**

0

Model: $\text{logit}\ \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$

$\text{expit}(\alpha_0)$ 0%

$\text{expit}(\alpha_0 + \alpha_1)$ 0%

$\text{expit}(\alpha_0 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2)$ 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

PS in older women, $\Pr[A = 1|L_1 = 1, L_2 = 1]$**?**

0

Model: $\text{logit}\ \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$

$\text{expit}(\alpha_0)$ 0%

$\text{expit}(\alpha_0 + \alpha_1)$ 0%

$\text{expit}(\alpha_0 + \alpha_2)$ 0%

$\text{expit}(\alpha_0 + \alpha_1 + \alpha_2)$ 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Parametric estimation of PS

Nonsaturated logistic regression model

$$\text{logit}\ \Pr[A=1 \mid L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$$

- Parameter estimates
  - $\hat{\alpha}_0$ = -1.09
  - $\hat{\alpha}_1$ = -0.30
  - $\hat{\alpha}_2$ = 0.53

See 2.2_propensity.R, lines 26-27

- These estimates result in slightly different PS estimates
  - Younger men $(L_1=0, L_2=0)$: 0.252
  - Older men $(L_1=0, L_2=1)$: 0.364
  - Younger women $(L_1=1, L_2=0)$: 0.201
  - Older women $(L_1=1, L_2=1)$: 0.299

Parametric estimation of PS

Nonsaturated logistic regression model

$$\text{logit}\ \Pr[A=1 \mid L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$$

- Same model as before except that it has no parameter for the product term $L_1 \times L_2$
- This model imposes a restriction on the possible values of the PS
  - The restriction that $\alpha_3 = 0$
  - What does this restriction mean?

Restricting the model such that $\alpha_3 = 0$ means that the men-women difference in logit PS is equal in old and young people

$\text{logit } \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$

0

True 0%

False 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Restricting the model such that $\alpha_3 = 0$ means that the odds ratio of smoking cessation for men vs. women is equal in old and young people

$\text{logit } \Pr[A = 1|L_1, L_2] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$

0

True 0%

False 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

What if higher dimensional setting?

- In our smoking cessation example
  - age as a continuous variable (L_2 measured in years),
  - additional confounders besides age and sex (L_3, L_4,…)
- Then nonparametric estimation may become impossible
  - Not enough data
- Need to make modeling assumptions
  - e.g., $\text{logit Pr}[A=1| L_1, L_2, L_3] = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_3$

In our smoking cessation example

- There are many potential confounders and some of them are continuous variables
- We can fit a PS model with
  - one parameter per indicator for categorical variables
  - linear and quadratic terms for continuous variables
  - few or no product terms between variables
- See computer code

What if PS model is misspecified?

- Examples
  - the model includes a linear and quadratic term for age but it should be a cubic polynomial
  - the model does not include product terms between age and sex but it should
- Then the estimate of PS will be wrong
  - The analysis described in the following slides will be wrong

2. Why should we care about the PS?

- If there is exchangeability conditional on the confounders L, there is exchangeability conditional on the PS
  - Rosenbaum and Rubin (1983)
- That is, if conditioning on the confounders is sufficient to block all backdoor paths between treatment and outcome, then conditioning on the PS is sufficient too
- We can replace the confounders *L* by the PS

The PS can be used in various ways

- Stratification on PS categories
- Outcome regression on PS
- Matching on PS
  - We will discuss these 3 today

- Inverse probability weighting
  - Coming soon
- G-estimation
  - Coming later

Stratification on PS categories

1. Estimate the PS for each individual
   - A continuous variable
2. Create a categorical PSc variable using the continuous PS
   - e.g., 10 levels, one per decile of the PS
3. Stratify individuals by decile of the PS
4. Estimate the effect in each decile
   - $\mathrm{E}[Y|A=1, PSc] - \mathrm{E}[Y|A=0, PSc]$

Stratification on PS categories
The estimand

- The average causal effects are now conditional on the categories of PS
- In our example, the differences in mean weight gain are now defined within deciles of the PS rather than within levels of the confounders in L

Stratification on PS categories
The effect of smoking cessation

- Mean difference in weight gain (95% CI)
  - 1st decile: 0.2 (-5.1, 5.5)
  - 2nd decile: 4.9 (2.4, 7.7)
  - 3rd decile: 4.7 (1.6, 7.8)
  - 4th decile: 2.3 (-1.1, 5.6)
  - …
- Too many estimates? Too imprecise?
  - Let's combine them

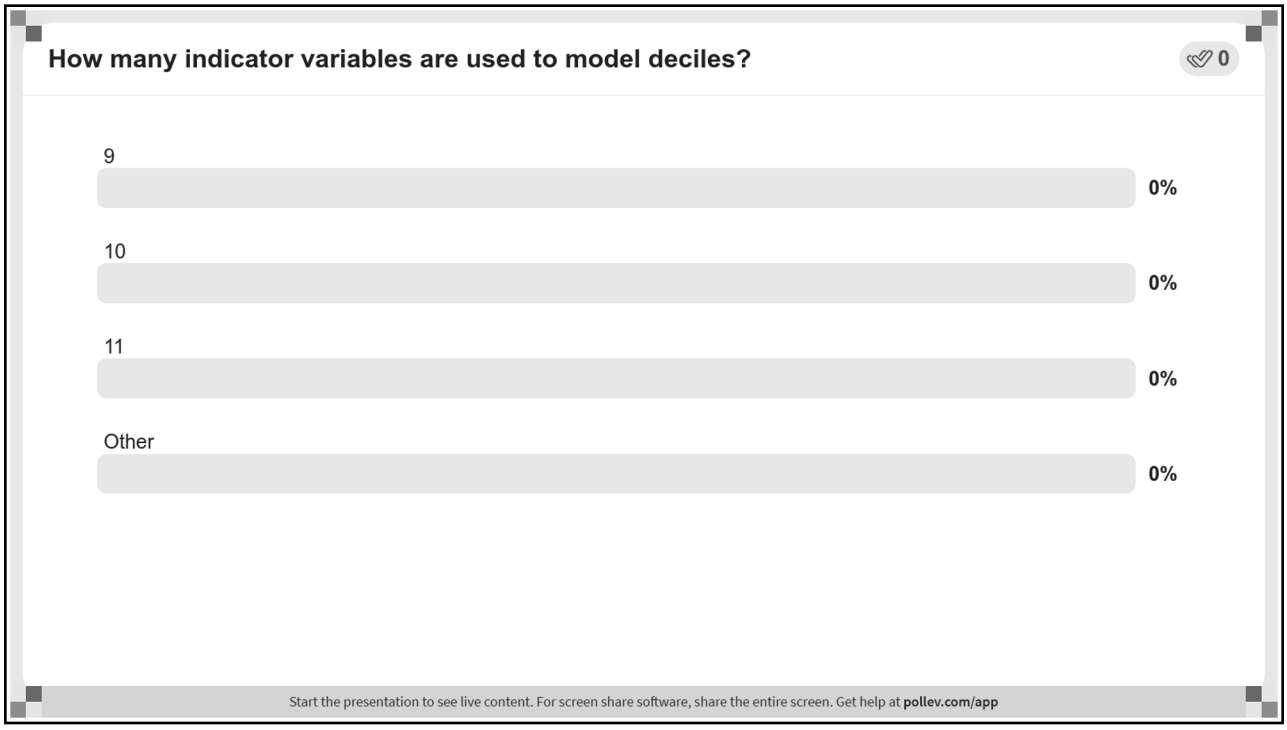
Outcome regression on PS categories

- Outcome regression with confounders *L* replaced by the categories (e.g., deciles) of the estimated PS
- To estimate the effect of smoking cessation on weight gain, we can fit the model

$$\mathrm{E}[Y|A,PSc] = \theta_0 + \theta_1 A + \theta_2 PSc1 + \theta_3 PSc2 + \ldots$$

  - where PS1 is an indicator for category 1 of PS, PS2 an indicator for category 2, etc.
- rather than the model

$$\mathrm{E}[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L_1 + \theta_3 L_2 + \ldots$$



Outcome regression on PS categories

- Linear regression model
  - $\mathrm{E}[Y|A,PSc] = \theta_0 + \theta_1 A + \theta_2 PSc1 + \theta_3 PSc2 + \theta_4 PSc3 + \ldots$
- The pooled estimate of the mean difference in weight gain conditional on the PS deciles is
  - $\hat{\theta}_1$ = 3.4
    - 95% CI: 2.5, 4.3

See 2.2_propensity.R, lines 65-66

Outcome regression on continuous PS

- Outcome regression with confounders *L* replaced by the estimated PS
- To estimate the effect of smoking cessation on weight gain, we can fit the model

  $\mathrm{E}[Y|A,PS] = \theta_0 + \theta_1 A + \theta_2 PS + \theta_3 PS^2$
- rather than the model

  $\mathrm{E}[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L_1 + \theta_3 L_2 + \theta_4 L_3 \ldots$

Outcome regression on continuous PS

- Linear regression model
  - $\mathrm{E}[Y|A,PS] = \theta_0 + \theta_1 A + \theta_2 PS + \theta_3 PS^2$

- The pooled estimate of the mean weight gain difference conditional on the continuous PS
  - $\hat{\theta}_1$ = 3.5
    - 95% CI: 2.5, 4.4

See 2.2_propensity.R, lines 69-70

Potential sources of bias for stratification/regression on PS

- Misspecification of the PS model
- Residual confounding within categories of PS
  - e.g., a decile-based classification may be too coarse
- Misspecification of outcome model
  - Same as for outcome regression
  - e.g., we include linear/quadratic terms for PS but there should be a cubic term too

An alternative: Propensity score matching

- For each treated individual, find an untreated one with same value of PS
  - That is, match on the propensity score
- Repeat this procedure for all treated individuals
  - Until you have as many of them as possible matched with untreated individuals
  - A matched cohort (matching can be 1-to-many)
- Conduct the analysis on the matched cohort
  - Discard data on unmatched individuals

Problem: Not 2 individuals with same PS?

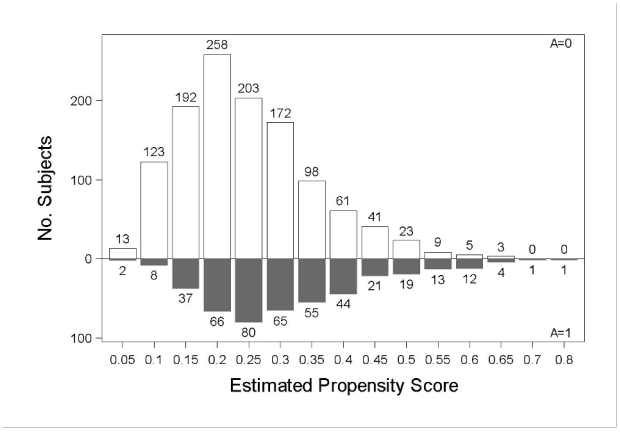
- Then need to match each treated individual with one or more untreated individuals with a "close enough" PS value
  - For example, untreated individual 233 (estimated PS = 0.0987) might be matched with treated individual 22904 (estimated PS = 0.0822)
- Many ways of defining closeness
  - e.g., PS within 0.05, or some other small difference

Defining closeness: Another bias-variance tradeoff

- If the closeness criteria are too loose
  - matched individuals have different PS values
  - exchangeability does not hold
  - residual confounding again
- If the closeness criteria are too tight
  - many individuals are excluded
  - approximate exchangeability but the effect estimate will have wider 95% confidence intervals

Which effect does PS matching estimate?

- In our example,
  - the effect of smoking cessation in individuals with PS<0.67
  - Who are these people?



Propensity score matching: The estimand

- PS matching may estimate the average causal effect in a population that is not well characterized
  - Because people don't have their PS tattooed on their forehead
- Better to characterize the target population in terms of observed variables
  - In our example, individuals with PS>0.67 were over age 50 and had smoked for less than 10 years
  - PS matching then estimates the effect in smokers under age 50 and smokers 50 and over who had smoked for at least 10 years

Conditions for the validity of these propensity score methods

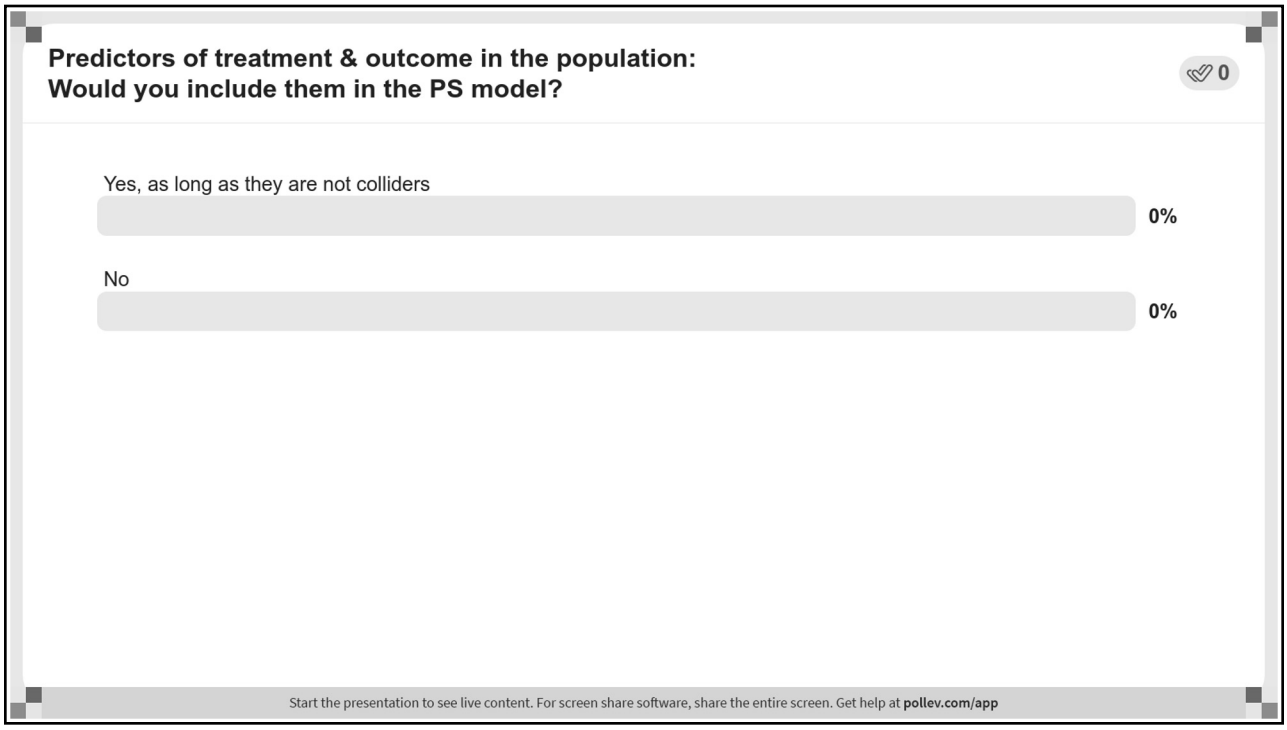
- Exchangeability, positivity, and consistency
  - Same as stratified-based methods
- No misspecification of
  - model for treatment
  - model for the outcome conditional on the PS
- Dichotomous treatment
  - PS not well defined for polytomous and continuous treatments

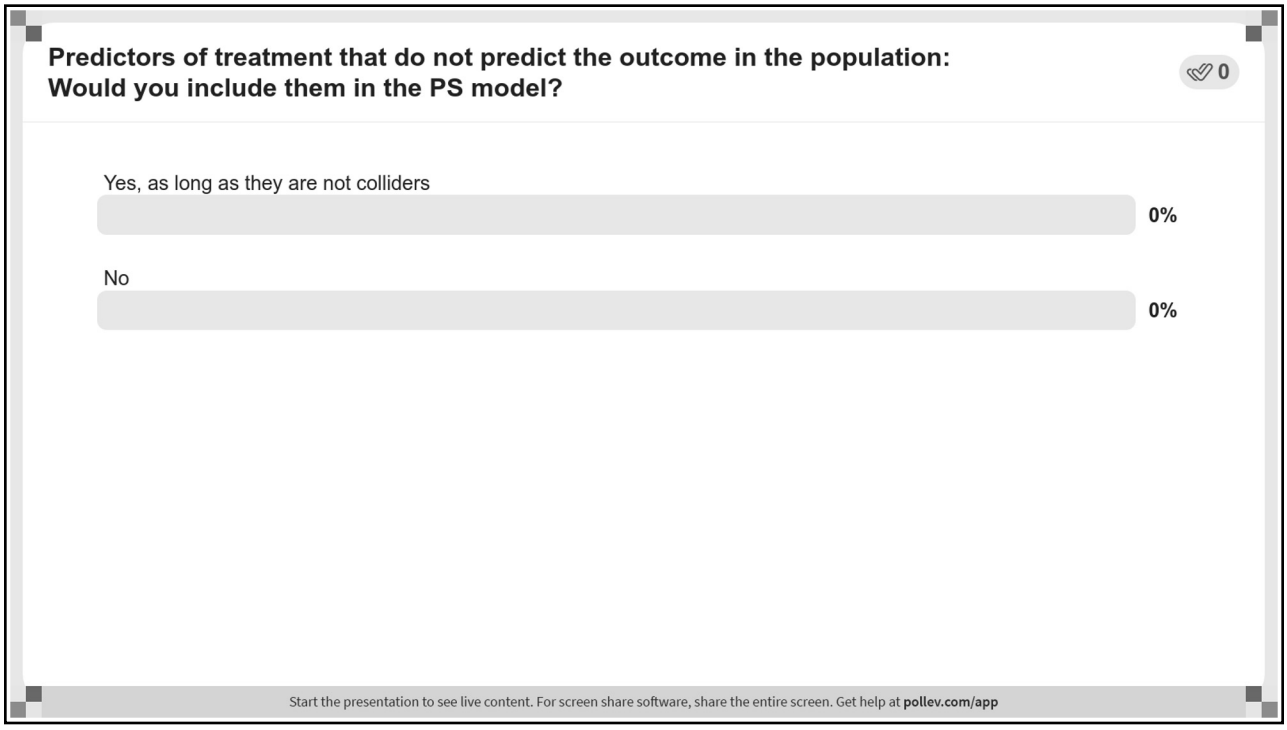
Stratification-based methods: Some caveats

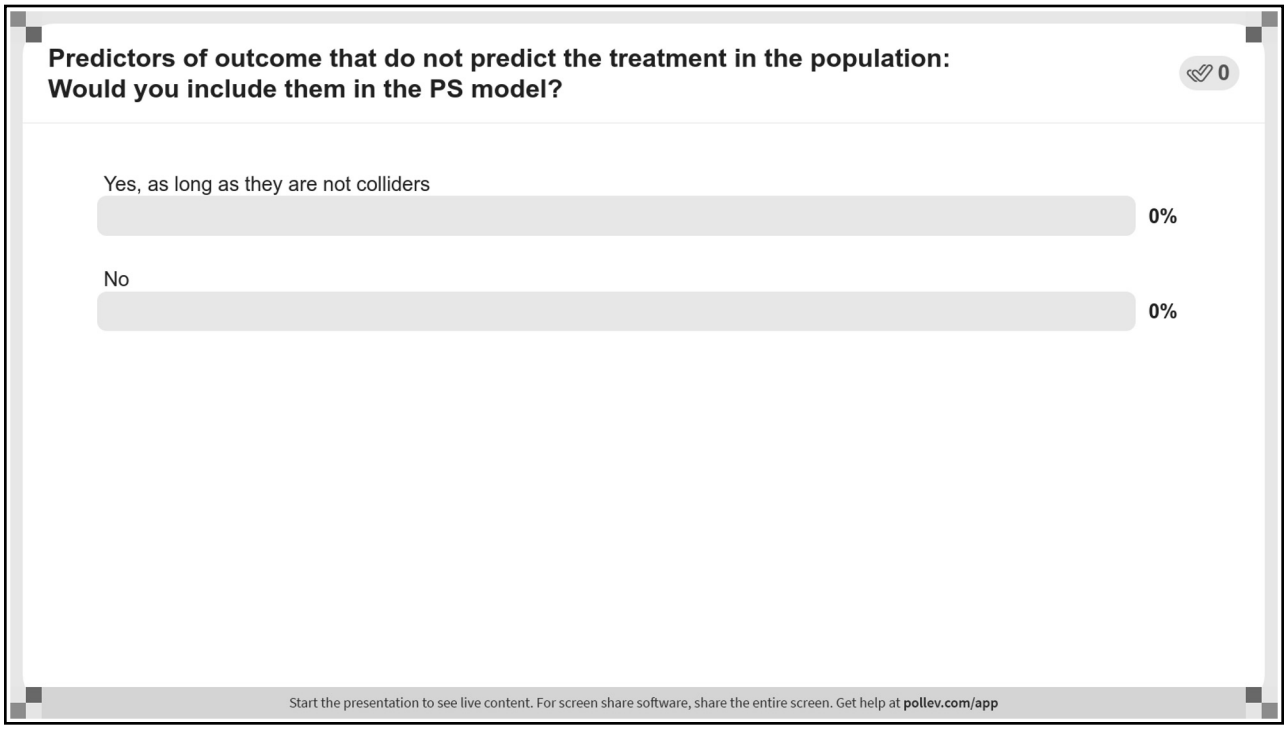
- We must worry about effect modification
  - even if we are not interested in effect modification
- Not average causal effect in the population
  - But average causal effect in strata defined by confounders or PS
- Possible bias when estimating the effect of time-varying treatments
  - More later

Causal question 2

- Causal effect of smoking cessation on death
- We can use a logistic regression model conditional on the estimated PS
  - All of the above discussion applies except that differences of means can be replaced by odds ratios
    - or other effect measures
- See Homework #2


Predictors of treatment & outcome in the population:
Would you include them in the PS model?
0
Yes, as long as they are not colliders
0%
No
0%
Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app


Predictors of treatment that do not predict the outcome in the population:
Would you include them in the PS model?
0
Yes, as long as they are not colliders
0%
No
0%
Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app



PS models

- Also used for IP weighting and g-estimation
- The goal is **not** to predict treatment perfectly
  - But to adjust for confounding
- Need to include all confounders to eliminate confounding
  - variables that help block all backdoor paths between treatment and outcome

A last note: 3 types of models

1. Predictive models
   - To predict a variable as well as possible in a particular population/setting
   - Parameters do not have a causal interpretation
2. Propensity score models
   - To predict treatment, but not as well as possible
   - Parameters do not have a causal interpretation
3. Structural models
   - To estimate the effect of treatment on outcome
   - Parameters do have a causal interpretation

Readings

- ☐ *Causal Inference: What If*. Chapter 15

Progress report

1. Introduction to modeling
2. Stratified analysis
   - outcome regression
   - propensity scores
3. Standardization