STANDARDIZATION: ESTIMATION

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



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

- Use standardization to estimate unconditional effects using parametric and nonparametric estimators
- ☐ Key concepts
 - Standardization
 - **■** Bootstrapping

□ 1629 ciga	rette smokers
□ Aged 25-7 (baseline)	74 years when interviewed in 1971-75
☐ Interviewe	ed again in 1982
alcohol us and follow	x, age, race, weight, height, education, e, and smoking intensity at both baseline y-up visits, and who answered the general istory 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

Causal questions of interest

- 1. What is the effect of smoking cessation on weight gain?
- ☐ This is an informal statement of the questions
- 2. What is the effect of smoking cessation on risk of death?

Standardization 5

A more formal version of causal question #1 First define the counterfactual means

if everybody had quit smoking

- \blacksquare $E[Y^{a=1}]$
- Y^{a=1} is an individual's outcome under a=1

if nobody had quit smoking

- \blacksquare $E[Y^{a=0}]$
- $Y^{a=0}$ is an individual's outcome under a=0

Then the formal question is:

 \square What is the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$?

Standardization

6

The average causal effect

$$E[Y^{a=1}] - E[Y^{a=0}]$$

- ☐ The effect that would be estimated in a hypothetical randomized trial of smoking cessation
- ☐ The unconditional (marginal) effect in the population, not the effect conditional on
 - the confounders
 - the propensity score

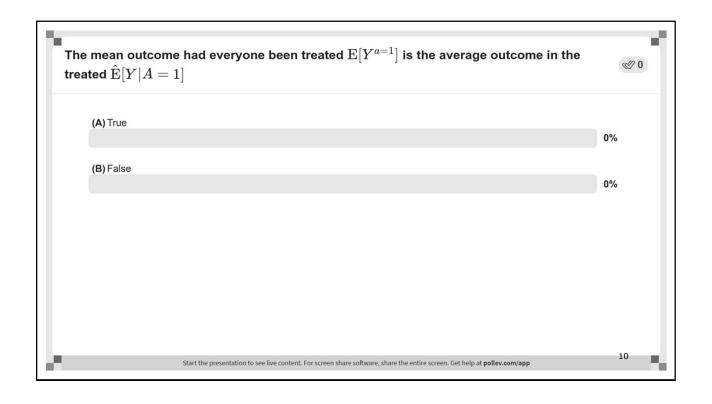
Standardization

Plan for today: Estimation of the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$

- 1. When confounding adjustment is not required
- 2. When confounding adjustment is required
 - Standardization by, say, 1-3 variables
 - Standardization by many variables
- ☐ In each case, we need to estimate 2 quantities
 - $E[Y^{a=1}]$: mean outcome had everybody been treated
 - $E[Y^{a=0}]$: mean outcome had nobody been treated

What if our study were an ideal randomized experiment...

- ☐ ... in which 403/1566 individuals had been randomly assigned to "smoking cessation"?
 - and adhered to their assignment
- ☐ Then the treated and the untreated would be exchangeable
 - there would be no confounding



Nonparametric estimation

Sample average

- ☐ Average weight gain in the treated
 - $\hat{E}[Y|A=1] = 4.5 \text{ kg}$
- ☐ Average weight gain in the untreated
 - $\hat{E}[Y|A=0] = 2.0 \text{ kg}$
- \square Difference $\hat{E}[Y|A=1] \hat{E}[Y|A=0]$
 - 2.5 kg

□ 95% CI: 1.7, 3.4

See 3_standardization.R, lines 15-18

■ A valid estimate of the average causal effect if the study were a randomized experiment

Standardization 11

Nonparametric estimation

Saturated linear model $E[Y|A] = \theta_0 + \theta_1 A$

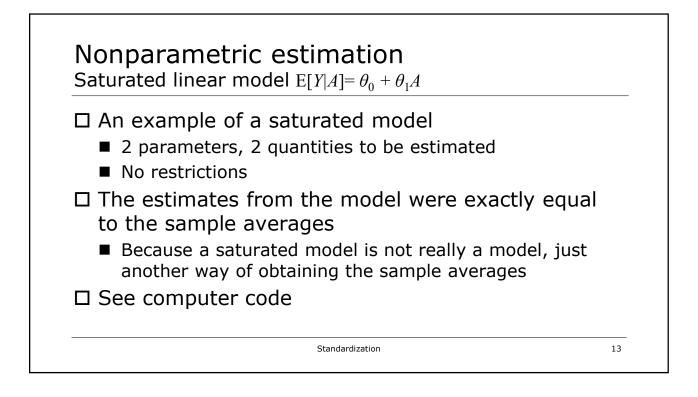
- ☐ Interpretation of parameters
 - Mean weight gain in untreated $E[Y|A=0] = \theta_0$
 - Mean weight gain in treated $E[Y|A=1] = \theta_0 + \theta_1$
 - Difference $E[Y|A=1] E[Y|A=0] = \theta_1$
- □ Parameter estimates
 - \blacksquare $\hat{\theta}_0 = 2.0$
 - $\hat{\theta}_1 = 2.5$

□ 95% CI: 1.7, 3.4

See 3 standardization.R, lines 21-22

Standardization

12



But our study is not a marginally randomized experiment	
□ in which individuals in the study population were randomly assigned to smoking cessation	
Standardization	14

Plan for today: Estimation of the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$

- 1. When confounding adjustment is not required
 - Marginally randomized experiments
- 2. When confounding adjustment is required
 - Conditionally randomized experiments, observational studies
 - Standardization by, say, 1-3 variables
 - Standardization by many variables

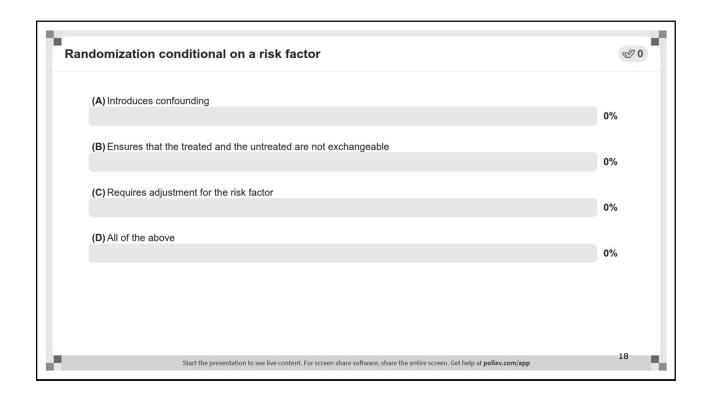
Standardization 15

What if this were an ideal randomized experiment...

- ☐ ... in which individuals had been randomly assigned to "smoking cessation" with a probability that depends on their age group?
 - Randomization is conditional on age group, rather than unconditional (or marginal)
- ☐ Probability of being assigned to smoking cessation is
 - 33.3% if age >50 years (L=1)
 - 22.5% if age \leq 50 years (L=0)

By design, the % of older people is greater in the smoking cessation group

- □ Older people gain less weight on average
 - Weight gain (kg) by smoking cessation status is
 - ☐ Quitters: 2.1 in older vs. 6.1 in younger
 - □ Non-quitters: -0.8 in older vs 3.0 in younger
- ☐ Is this a problem?



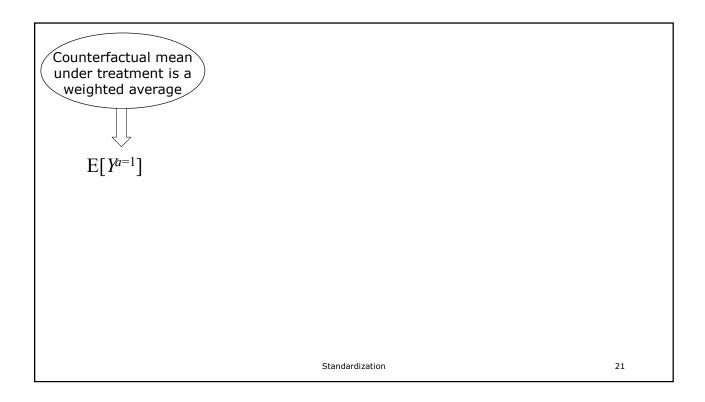
The unadjusted difference $\hat{E}[Y|A=1] - \hat{E}[Y|A=0] = 2.5$

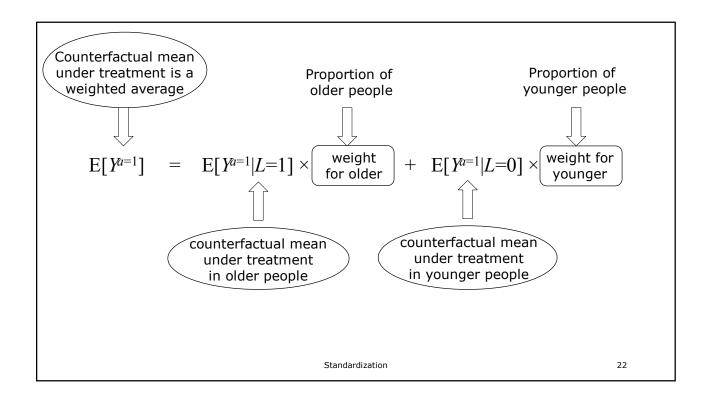
- □ will make smoking cessation look better because
 - The average outcome in the treated $\hat{E}[Y|A=1]$ is less than the mean outcome had everyone been treated $E[Y^{a=1}]$
 - The average outcome in the untreated $\hat{E}[Y|A=0]$ is greater than the mean outcome had everyone been untreated $E[Y^{a=0}]$
- ☐ Let's describe how to adjust for confounding by age group via standardization

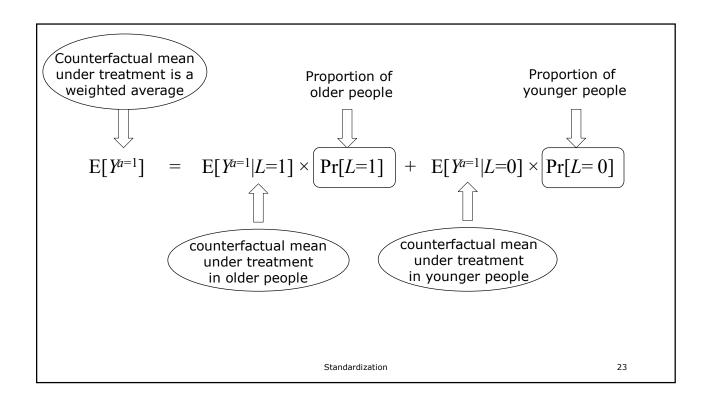
Standardization 19

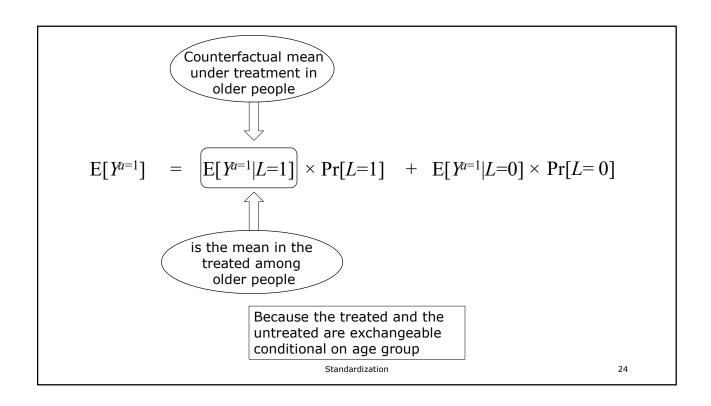
We need to estimate $E[Y^{a=1}]$ and $E[Y^{a=0}]$

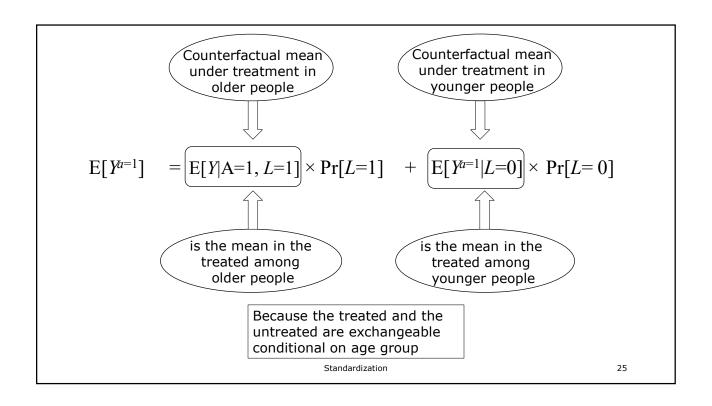
- \square First let's focus on $E[Y^{a=1}]$
 - the mean outcome had everyone been treated
- \square E[$Y^{a=1}$] is a weighted average of the corresponding means in
 - Older individuals $E[Y^{a=1}|L=1]$
 - Younger individuals $E[Y^{a=1}|L=0]$
- □ with weights equal to the proportions of older and younger individuals
 - \blacksquare Pr[L=1], Pr[L=0]

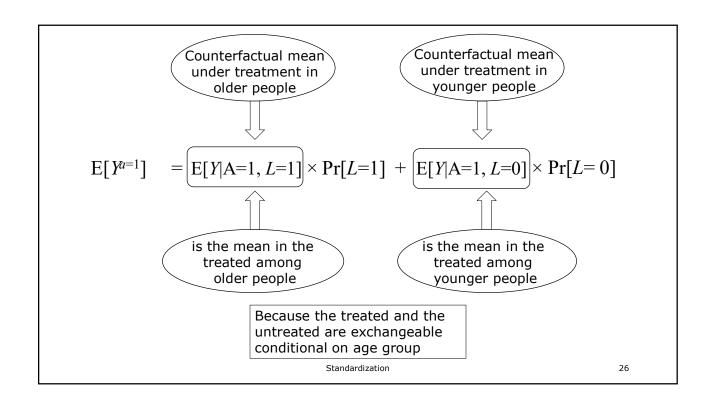


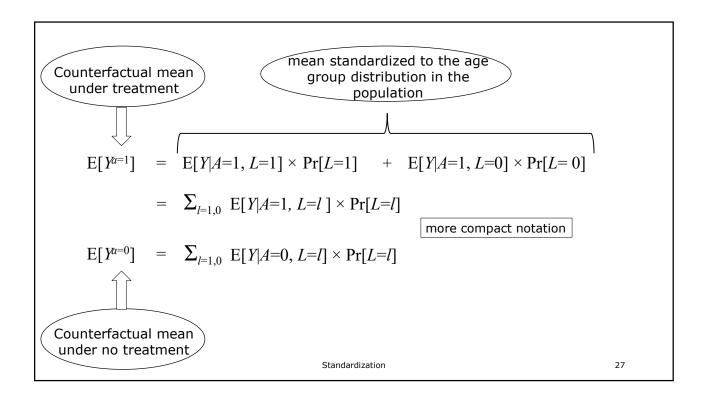












Nonparametric estimation

Sample averages and proportions

☐ Standardized average in the treated

 $E[Y|A=1, L=1] \times Pr[L=1] + E[Y|A=1, L=0] \times Pr[L=0]$

- $2.10 \times 0.2989 + 6.06 \times 0.7011 = 4.87 \text{ kg}$
- ☐ Standardized average in the untreated

 $E[Y|A=0, L=1] \times Pr[L=1] + E[Y|A=0, L=0] \times Pr[L=0]$

■ $(-0.76) \times 0.2989 + 2.99 \times 0.7011 = 1.87 \text{ kg}$

See 3_standardization.R, lines 29-35

- ☐ Difference: 3.00 kg
 - causal interpretation as $E[Y^{a=1}] E[Y^{a=0}]$ if treatment had been randomized conditional on age group

Unconditional (marginal) versus conditional effects

- ☐ We are now concerned with the average causal effect in the entire population
 - Marginal effect: $E[Y^{a=1}] E[Y^{a=0}]$
- □ Not with the average causal effect within levels of the covariates
 - Conditional effect in the younger: $E[Y^{a=1}|L=0] E[Y^{a=0}|L=0]$
 - Conditional effect in the older: $E[Y^{a=1}|L=1] E[Y^{a=0}|L=1]$

Standardization 29

Unconditional (marginal) vs. conditional effects

If age group were the only confounder

- The standardized mean difference 3.00 kg estimated here is a valid estimator of the marginal effect
- The mean difference in each age group is a valid estimator of the conditional effect estimates
 - □ Younger: E[Y|A=1, L=0] E[Y|A=0, L=0] estimate is 3.06 kg
 - □ Older: E[Y|A=1, L=1] E[Y|A=0, L=1] estimate is 2.86 kg
 - $\ \square$ Little evidence of effect modification by age group

Estimation of the average causal effect

$$E[Y^{a=1}] - E[Y^{a=0}]$$

- \square E[$Y^{a=1}$] is the standardized mean in the treated
- \square $E[Y^{a=0}]$ is the standardized mean in the untreated

☐ We can estimate the standardized means

- without models (what we have just done)
 - \square Sample averages for outcome means E[Y|A,L]
 - \square Sample proportions for confounder prevalence Pr[L=l]
- with models

Standardization 31

Nonparametric estimation

Saturated linear model $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L + \theta_3 AL$

☐ Interpretation of parameters

- $E[Y|A=0, L=0] = \theta_0$
- $E[Y|A=0, L=1] = \theta_0 + \theta_2$
- $E[Y|A=1, L=0] = \theta_0 + \theta_1$
- $E[Y|A=1, L=1] = \theta_0 + \theta_1 + \theta_2 + \theta_3$

☐ Use parameter estimates to calculate

- $\hat{E}[Y|A=0, L=0] = 2.99$
- $\hat{\mathbf{E}}[Y|A=0, L=1] = -0.76$
- $\hat{\mathbf{E}}[Y|A=1, L=0] = 6.06$
- $\hat{\mathbf{E}}[Y|A=1, L=1] = 2.10$

See 3_standardization.R, lines 38-44

Standardization

32

Nonparametric estimation

Saturated linear model $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L + \theta_3 AL$

- ☐ Saturated because
 - 4 parameters, 4 quantities to be estimated
 - No restrictions
- ☐ The estimates from the model were exactly equal to the nonparametric estimates we obtained before
 - Same standardized means
- ☐ Let us now consider a nonsaturated model

Standardization 33

Parametric estimation

Nonsaturated linear model $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L$

- ☐ Interpretation of parameters
 - $E[Y|A=0, L=0] = \theta_0$
 - $E[Y|A=0, L=1] = \theta_0 + \theta_2$
 - $E[Y|A=1, L=0] = \theta_0 + \theta_1$
 - $E[Y|A=1, L=1] = \theta_0 + \theta_1 + \theta_2$
- ☐ Use parameter estimates to calculate
 - $\hat{\mathbf{E}}[Y|A=0, L=0] = 3.01$
 - $\hat{\mathbf{E}}[Y|A=0, L=1] = -0.81$
 - $\hat{\mathbf{E}}[Y|A=1, L=0] = 6.00$
 - $\hat{\mathbf{E}}[Y|A=1, L=1] = 2.19$

See 3_standardization.R, lines 46-52

Standardization

34

Parametric estimation

Nonsaturated linear model $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L$

- ☐ Standardized average in the treated
 - \blacksquare 2.19 × 0.2989 + 6.01 × 0.7011 = 4.86
- ☐ Standardized average in the untreated
 - \blacksquare -0.81 × 0.2989 + 3.01 × 0.7011 = 1.87
- □ Difference: 2.99 kg
 - causal interpretation if
 - ☐ the treatment were randomized conditional on age group
 - ☐ the outcome model is correctly specified

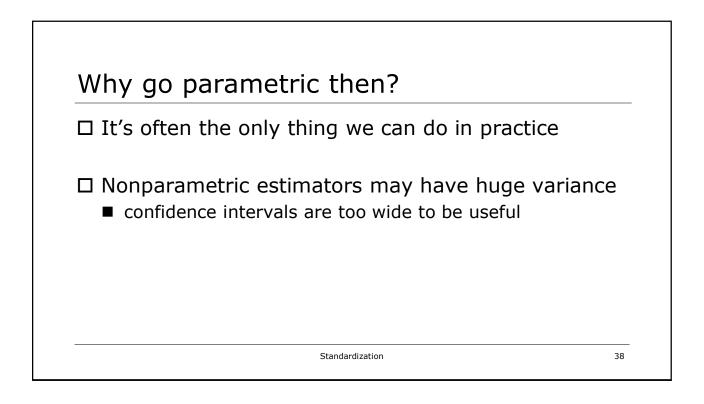
Standardization 35

Parametric estimation

Nonsaturated linear model $E[Y|A, L] = \theta_0 + \theta_1 A + \theta_2 L$

- ☐ The effect estimate was 2.99 kg
 - Very similar to nonparametric estimate 3.00 kg
- ☐ We made a modeling assumption / imposed an a priori restriction
 - that may be approximately correct

Restriction: The contributions of A and L to the mean of Y are admitted Model: $\mathrm{E}[Y A,L]= heta_0+ heta_1A+ heta_2L$	ditive © 0
True	
	0%
False	
	0%
Start the presentation to see live content. For screen share software, share the entire screen. O	Set help at pollev.com/app



What if randomization had been conditional on 2 covariates?

- ☐ Consider two dichotomous variables
 - L_1 Sex: Women (1), Men (0)
 - L_2 Age group: older than 50 (1), 50 or less (0)
- ☐ Suppose that treatment was randomly assigned with a different probability in each of the following 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 $(L_1=1, L_2=1)$

Standardization 39

Standardized mean in the treated

- ☐ To standardize (and adjust for confounding) we need to
 - compute the stratum-specific sample average $E[Y|A=1, L_1, L_2]$ in each of the 4 combinations of values of (L_1, L_2)
 - \blacksquare compute the prevalence of each of the 4 strata (L_1, L_2)
 - compute the weighted average of the 4 stratum-specific means:

$$\begin{split} & \text{E}[Y|A=1, L_1=0, L_2=0] \times \text{Pr}[L_1=0, L_2=0] + \\ & \text{E}[Y|A=1, L_1=0, L_2=1] \times \text{Pr}[L_1=0, L_2=1] + \\ & \text{E}[Y|A=1, L_1=1, L_2=0] \times \text{Pr}[L_1=1, L_2=0] + \\ & \text{E}[Y|A=1, L_1=1, L_2=1] \times \text{Pr}[L_1=1, L_2=1] \end{split}$$

See 3 standardization.R, lines 59-83

If randomization had been conditional on 2 dichotomous covariates

☐ We would need to estimate a total of 8 means

A (nonparametric) saturated linear model would have 8 parameters

■ The dimensionality of the problem starts to grow...

Standardization 41

Plan for today: Estimation of the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$

- 1. When confounding adjustment is not required
 - Marginally randomized experiments
- 2. When confounding adjustment is required
 - Conditionally randomized experiments, observational studies
 - Standardization by, say, 1-3 variables
 - Standardization by many variables

What if randomization had been conditional on 10 covariates?

☐ To standardize, we need to

- compute the stratum-specific sample average E[Y|A=1,L] in each of the $2^{10} = 1024$ combinations of values of the vector L
- \blacksquare compute the prevalence of L in each of the 1024 strata
- compute the weighted average of the 1024 stratum-specific means
 - ☐ A very long sum (an integral if some covariates were continuous)

Standardization 43

Nonparametric vs. parametric estimation of E[Y|A, L] for vector L

- ☐ A nonparametric estimator needs to estimate many parameters
 - If treatment A plus 10 dichotomous variables, 2¹¹= 2048 parameters
 - The curse of dimensionality
- ☐ A parametric estimator can get away with estimating far fewer parameters
 - Example: if 11 dichotomous variables, 11 parameters
 □ Under the assumption that their contributions to the mean of Y are additive
 - If the parametric model is correctly specified, large gains in variance (statistical efficiency)

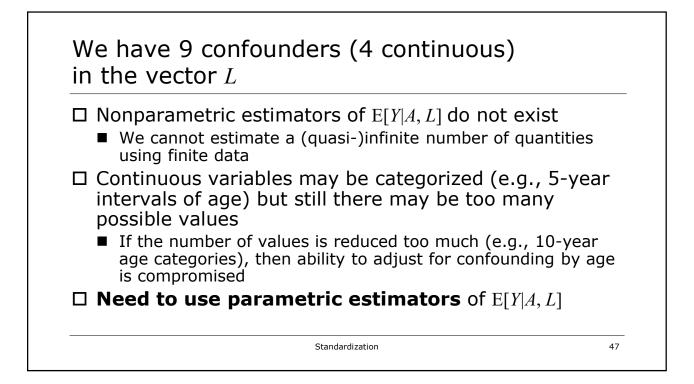
What if randomization had been conditional on many covariates?

- ☐ This is often the situation we consider in observational studies
- ☐ Investigators are often willing to assume that
 - an observational study with, say, 10 confounders is like a randomized experiment with randomization conditional on those 10 variables
 - all confounders are correctly measured

Standardization 45

Ours is an observational study

- ☐ Smoking cessation *A* was *not* conditionally randomized, but we are willing to assume that
 - all confounders were measured
 - Exchangeability within levels of sex, race, age, education, intensity and duration of smoking, exercise, active lifestyle, and body weight
- ☐ Then the average causal effect $E[Y^{a=1}] E[Y^{a=0}]$ can be consistently estimated by the difference of standardized (by L) averages



Parametric estimation

Nonsaturated linear model

- ☐ Fit model with linear+quadratic terms for continuous variables and few or no product terms
 - lacktriangle The predicted values from this model are estimates of the average outcome conditional on L
- \square Sum over all combinations of values of L
 - This integral can be approximated by using the empirical distribution of the confounders
 - That is, compute the average of predicted values for each individual in the population under treatment (A=1) and under no treatment (A=0)

Parametric estimation

Nonsaturated linear model

- ☐ Estimates of standardized mean
 - ~5.2 kg in the treated
 - ~1.7 kg in the untreated
- □ Difference: 3.5 kg See 3_standardization.R, lines 86-123
 - This difference would have a causal interpretation if all confounders had been included in the standardization procedure
 - Note the difference gets further from zero as more baseline covariates are adjusted for

Standardization 49

Parametric estimation

95% confidence interval via "bootstrapping"

- ☐ The lazy statistician's method
 - Sample with replacement to create a new sample of the same size as the study sample
 - Estimate the effect estimate in that sample
 - Repeat 1000 times
 - ☐ find percentiles 2.5 and 97.5 of the 1000 estimates and make them the limits of the 95% confidence interval
 - □ or compute the standard error of the 1000 estimates and use it to compute the limits of the 95% confidence interval
 - In our study the 95% CI is (2.5, 4.3) See 3_standardplusbootstrap.R

	sal question 2 usal effect of smoking cessation on death	
on	can use a logistic regression model condition treatment and confounders to estimate the death	
	Il of the above discussion applies except that tandardized means are replaced by standardized r	isks
	Standardization	5:

(Robins 1986)		
☐ General form	of standardization	
	tments, it's exactly the standardization scribed above	
	ised in the presence of time-varying and confounders	
•	discovered by computer discovered by computer ficial intelligence researchers	
☐ Cannot be use	d nonparametrically	

The parametric g-formula

- ☐ Estimate the components of the g-formula using models, and plug them in the formula
 - what we did above for a time-fixed treatment
- ☐ Challenges for time-varying treatments:
- 1. Computationally intensive
- 2. Conditional distributions of outcome and confounders are estimated via parametric models
 - Possibility of model misspecification

Standardization 53

Example: Lifestyle and risk of CHD Taubman et al. Int J Epidemiol 2009

Table 3 Simulated population risk estimates using the g-formula. Hypothetical interventions on entire cohort

Intervention	20-year risk	Population risk ratio	Population risk difference
(0) No intervention	3.68 (3.56 to 4.09)	1	0
(1) Quit smoking	3.01 (2.86 to 3.38)	0.82 (0.78 to 0.85)	-0.67 (-0.88 to -0.56)
(2) Exercise at least 30 min/day	2.90 (2.47 to 3.60)	0.79 (0.64 to 0.92)	-0.77 (-1.41 to -0.32)
(3) Keep diet score in the top 2 quintiles	3.27 (3.08 to 3.68)	0.89 (0.82 to 0.95)	-0.41 (-0.70 to -0.19)
(4) Consume at least 5g alcohol per day	3.19 (2.84 to 3.72)	0.87 (0.75 to 0.98)	-0.48 (-0.97 to -0.08)
(5) Maintain BMI <25	3.62 (3.45 to 4.11)	0.98 (0.93 to 1.04)	-0.06 (-0.28 to 0.14)
(6) 'Low-risk' lifestyle (1-3 combined)	2.22 (1.85 to 2.74)	0.60 (0.48 to 0.70)	-1.45 (-2.02 to -1.13)
(7) 'Low-risk' lifestyle (1-3 and 5 combined)	2.17 (1.78 to 2.69)	0.59 (0.47 to 0.70)	-1.51 (-2.06 to -1.13)
(8) 'Low-risk' lifestyle (1-3 and 4 combined)	1.88 (1.51 to 2.38)	0.51 (0.40 to 0.63)	-1.80 (-2.29 to -1.40)
(9) 'Low-risk' lifestyle (1–5 combined)	1.89 (1.46 to 2.41)	0.51 (0.39 to 0.64)	-1.79 (-2.34 to -1.41)

In summary, ass for causal infere	•	
☐ Exchangeability		
☐ Positivity		
☐ Consistency (inc	luding well-defined int	erventions)
☐ No model misspe	ecification	
	Standardization	55

Readings		
⊔ Causal Interence	e: What If. Chapter 13	
	Standardization	56

Progress report

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