STRATIFIED ANALYSIS: OUTCOME REGRESSION

Barbra Dickerman, Joy Shi, Miguel Hernán
DEPARTMENT OF EPIDEMIOLOGY



Before coming to class, you were expected to review

- ☐ the definitions of
 - average causal effect
 - confounding
 - confounder
- ☐ Recommended sources
 - EPI201/202 materials
 - Chapters 1 and 7 of the *Causal Inference: What If* book
 - Lessons 1 and 2 of the HarvardX *Causal Diagrams* course

Outcome Regression

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

- Define marginal and conditional causal effects
- Estimate conditional effects using a stratified analysis
- Estimate conditional effects using a parametric model
- Explain the bias-variance tradeoff in parametric modeling
- □ Key concepts
 - Counterfactual contrasts in the entire population
 - Counterfactual contrasts in subgroups of the population
 - Conditional exchangeability
 - Parametric and nonparametric estimators
 - Bias-variance tradeoff

Outcome Regression

3

The data

- ☐ We will be using a subset of the NHANES I Epidemiologic Follow-up Study (NHEFS)
 - More information on the NHEFS https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx
- □ Dataset is used throughout Part II of *Causal Inference: What If* and can be downloaded from the course website

Outcome Regression

CL. A .		I _ L:
Stuav	popu	ıatıon
– • • • • • • • • • • • • • • • • • • •	POPG	

□ 1629	cigarette	smo	kers
--------	-----------	-----	------

- □ 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

Outcome Regression

5

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

Outcome Regression

The causal questions of interest (informal version)

What is the effect of smoking cessation on

- 1. weight gain?
- 2. death?
- ☐ We will use these questions throughout the course to describe different methods for causal inference

Outcome Regression

7

Potential or counterfactual outcomes

- ☐ Precise causal questions require counterfactuals
- \square Under treatment a=1
 - \blacksquare $Y^{a=1}$ is an individual's weight gain if they had quit smoking
 - lacksquare $D^{a=1}$ indicates whether an individual would have died if they had quit smoking
- \square Under no treatment a=0
 - \blacksquare $Y^{a=0}$ is an individual's weight gain if they had not quit smoking
 - lacksquare $D^{a=0}$ indicates whether an individual would have died if they had not quit smoking

Outcome Regression

The causal effect of smoking cessation on

- ☐ Weight gain
 - Causal mean difference: $E[Y^{a=1}] E[Y^{a=0}]$ □ Additive scale (average causal effect)
- □ Death
 - Causal risk difference: $Pr[D^{a=1}=1] Pr[D^{a=0}=1]$ □ additive scale (average causal effect)
 - Causal risk ratio: $Pr[D^{a=1}=1] / Pr[D^{a=0}=1]$ □ multiplicative scale
 - Causal odds ratio: $(Pr[D^{a=1}=1] / Pr[D^{a=1}=0]) / (Pr[D^{a=0}=1] / Pr[D^{a=0}=0])$ □ multiplicative scale

Outcome Regression

9

The average causal effect can also be defined in subsets or strata of the population

- \square Select one stratum L=l
 - e.g., 65-year-old white women
- \square Mean weight gain in stratum L=l
 - if everybody had quit smoking: $E[Y^{a=1}|L=l]$
 - if nobody had quit smoking: $E[Y^{a=0}|L=l]$
- \square Conditional average causal effect in stratum L=l
 - $E[Y^{a=1}|L=l] E[Y^{a=0}|L=l]$

Outcome Regression

Some causal inference models only estimate conditional average causal effects

- 1. Stratified analysis (nonparametric)
- 2. Outcome regression (parametric)
- 3. Some propensity score methods (parametric)
- ☐ All these methods are based on stratification to adjust for confounding
 - Today we will talk about them

Outcome Regression

11

Stratification-based methods to estimate conditional average causal effects

- ☐ Most commonly used methods to adjust for confounding
 - Pick a random article and chances are the authors used some form of stratification-based method
- \square They require that the quitters (A=1) and the nonquitters (A=0) are exchangeable conditional on the measured variables L
 - Like all other methods for causal inference (except instrumental variable estimation)

Outcome Regression

Under conditional exchangeability, or no unmeasured confounding, given L

In each stratum L=l:

- \square The **mean weight gain if everybody had quit smoking** $\mathrm{E}[Y^{a=1}|L=l]$ is consistently estimated by the average weight gain among those who did quit smoking
 - $\hat{\mathbf{E}}[Y|A=1, L=l]$
- \square The **mean weight gain if nobody had quit smoking** $\mathrm{E}[Y^{a=0}|L=I]$ is consistently estimated by the average weight gain among those who did not quit smoking
 - \blacksquare $\hat{E}[Y|A=0, L=l]$

Outcome Regression

13

Example

- \square Suppose the only confounder L is biological sex
 - Men: *L*=0
 - Women: *L*=1
- ☐ Then we would estimate the difference in mean weight gain for treated vs. untreated
 - In men: $E[Y^{a=1}|L=0] E[Y^{a=0}|L=0]$
 - In women: $E[Y^{a=1}|L=1] E[Y^{a=0}|L=1]$
- ☐ By computing the corresponding sample averages

Outcome Regression

There are a few ways we can consider doing this

- □ Nonparametric estimation
 - Sample averages
 - Saturated outcome model
- □ Parametric estimation
 - Nonsaturated outcome model
- ☐ Let's take a look...

Outcome Regression

15

Nonparametric estimation

Sample average in Men

- □ 762 men
 - Out of 1566 individuals
- ☐ Mean weight gain in treated men
 - $\hat{E}[Y|A=1, L=0] = 4.8$
- ☐ Mean weight gain in untreated men
 - $\hat{E}[Y|A=0, L=0] = 2.0$
- □ Difference $\hat{E}[Y|A=1] \hat{E}[Y|A=0]$
 - 2.8 kg

□ 95% CI: 1.6, 4.1 (p-value < 0.01)

See 2.1_outcomereg.R, lines 10-13

Outcome Regression

Sample average in Women

- □ 804 women
 - Out of 1566 individuals
- ☐ Mean weight gain in treated women
 - $\hat{E}[Y|A=1, L=1] = 4.2$
- ☐ Mean weight gain in untreated women
 - $\hat{E}[Y|A=0, L=1]=2.0$
- □ Difference $\hat{E}[Y|A=1] \hat{E}[Y|A=0]$
 - 2.2 kg

□ 95% CI: 0.7, 3.6 (p-value < 0.01)

See 2.1_outcomereg.R, lines 16-19

Outcome Regression

17

An alternative estimation procedure

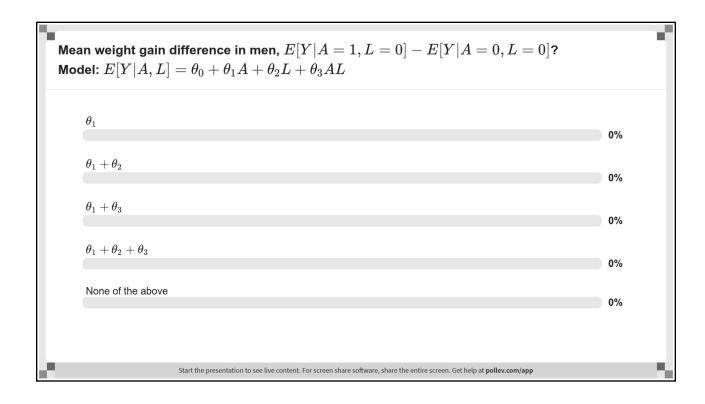
- ☐ We have computed the sample average of the outcome in 4 groups
 - 1. Men who did not quit smoking (A=0, L=0)
 - 2. Men who did quit smoking (A=1, L=0)
 - 3. Women who did not quit smoking (A=0, L=1)
 - 4. Women who did quit smoking (A=1, L=1)
- ☐ Let's now use outcome regression to estimate the same 4 quantities

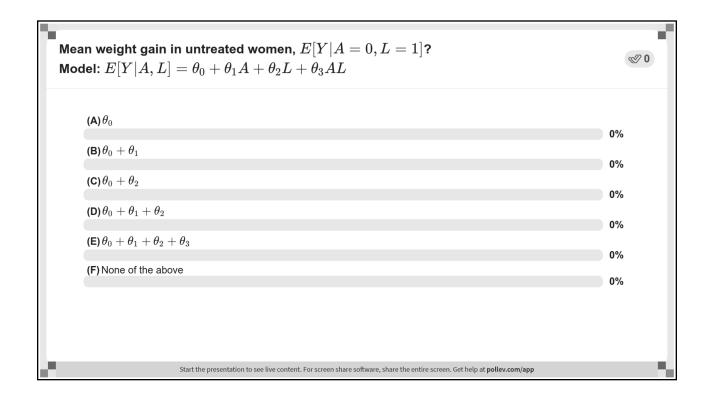
Outcome Regression

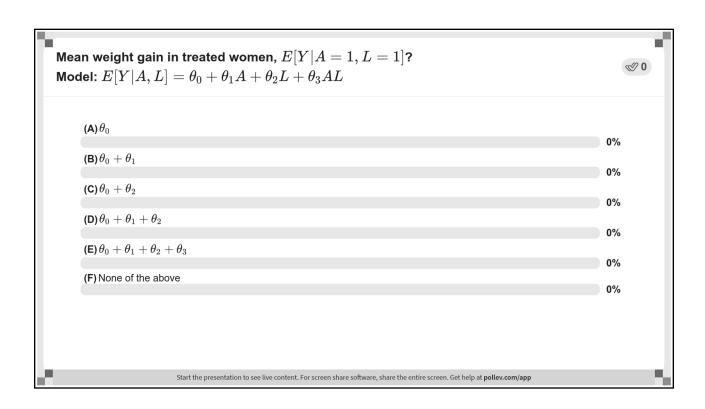
Saturated linear model

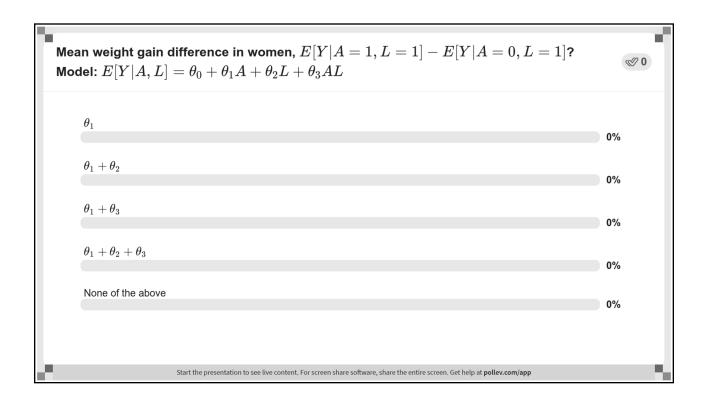
- ☐ Linear regression model
 - $\blacksquare \quad \mathbf{E}[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L + \theta_3 AL$
- ☐ Interpretation of parameters
 - Mean weight gain in untreated men
 - \square E[Y|A=0, L=0]= θ_0
 - Mean weight gain in treated men
 - \square E[*Y*|*A*=1, *L*=0]= $\theta_0 + \theta_1$

Outcome Regression









Saturated linear model

- ☐ Linear regression model
 - $\blacksquare \quad \mathbf{E}[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L + \theta_3 AL$
- □ Parameter estimates
 - $\blacksquare \hat{\theta}_0 = 2.00$
 - $\hat{\theta}_1 = 2.83$
 - $\hat{\theta}_2 = -0.03$
 - $\hat{\theta}_3 = -0.65$

See 2.1_outcomereg.R, lines 24-25

Outcome Regression

Saturated linear model

- ☐ An example of a saturated model
 - 4 parameters = 4 quantities to be estimated
 - $\ \square$ 1 mean for each covariate pattern defined by a combination of the values of A and L
 - Therefore no a priori restrictions
- ☐ The estimates from the model were exactly equal to the nonparametric estimates we obtained before
 - Because a saturated model is not really a model, just another way of obtaining nonparametric estimates (sample averages in this case)

Outcome Regression

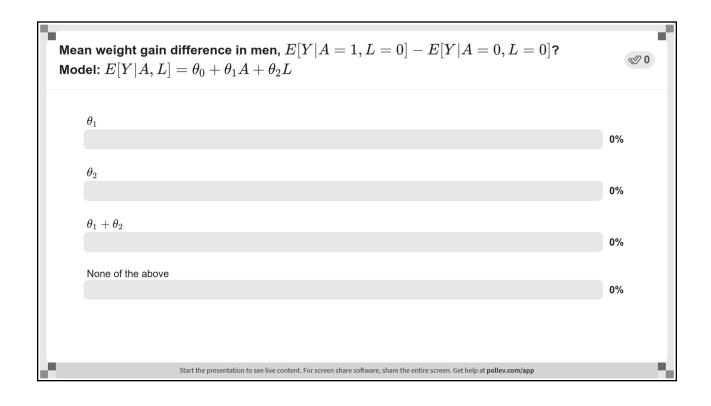
25

Parametric estimation

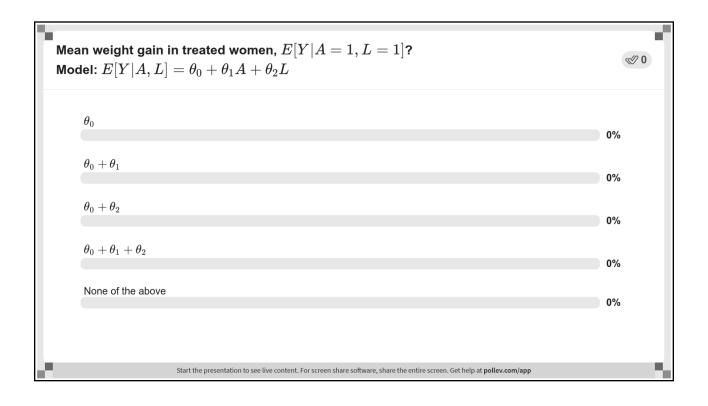
Nonsaturated linear model

- ☐ Linear regression model
 - $\blacksquare \quad \mathbf{E}[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L$
- ☐ Interpretation of parameters
 - Mean weight gain in untreated men
 - \square E[Y|A=0, L=0]= θ_0
 - Mean weight gain in treated men
 - \square E[Y|A=1, L=0]= $\theta_0 + \theta_1$

Outcome Regression



	ain in untreated women, $E[Y A=0,L=1]$? $[L]= heta_0+ heta_1A+ heta_2L$	₩ 0
$ heta_0$		0%
$ heta_0 + heta_1$		0%
$ heta_0 + heta_2$		0%
$ heta_0 + heta_1 + heta_2$		0%
None of the ab	ove	0%
	Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app	



	difference in women, $E[Y A=1,L=1]-E[Y A=0,L=1]$? $ext{2} = heta_0 + heta_1 A + heta_2 L$	€ 0
$ heta_1$		0%
$ heta_2$		0%
$ heta_1+ heta_2$		0%
None of the above		0%
	Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app	- 8

Parametric estimation

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

- □ Parameter estimates
 - $\hat{\theta}_0 = 2.09$
 - $\blacksquare \ \hat{\theta}_1 = 2.52$
 - $\hat{\theta}_2 = -0.20$
- ☐ These parameter estimates result in slightly different mean estimates
 - $\hat{E}[Y|A=1, L=1] = 4.4$
 - $\hat{E}[Y|A=0, L=1] = 1.9$
 - $\hat{E}[Y|A=1, L=0] = 4.6$
 - $\hat{\mathbf{E}}[Y|A=0, L=0] = 2.1$

See 2.1_outcomereg.R, lines 35-36

See 2.1_outcomereg.R, lines 33-34

Outcome Regression

31

Parametric estimation

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

- \square This model imposes a restriction on the values of the mean weight gain Y:
 - the difference in means between treated and untreated is the same for men and women
- ☐ Equivalently,
 - no additive effect modification by sex
 - lacktriangle the contributions of A and L to the mean of Y are additive
 - The parameter θ_3 is equal to zero

Outcome Regression

Nonparametric vs. parametric estimation Nonparametric no modeling assumptions no bias introduced by modeling assumptions Parametric modeling assumptions possible bias introduced by incorrect modeling assumptions

■ It's often the only thing you can do

■ Nonparametric estimators may have huge variance □ confidence intervals too wide to be useful

☐ Why would we use parametric models then?

Outcome Regression

33

Why go parametric? Continuous treatment A or confounders L

- Remember: a continuous variable can be viewed as a categorical variable with infinite categories
- ☐ Nonparametric estimators do not exist
 - We cannot estimate an infinite number of quantities (e.g., means) using finite data
- \square Need to use parametric estimators of E[Y|A, L]
- ☐ Continuous variables are often categorized
 - If too few categories (e.g., 20-year age categories), then ability to adjust for confounding may be compromised

Outcome Regression

Why go parametric? Multiple variables in vector L □ Suppose there is a dichotomous treatment A and 10 dichotomous variables in L □ A nonparametric estimator of E[Y|A, L] needs to estimate 2¹¹=2048 parameters ■ The curse of dimensionality □ A parametric estimator can get away with estimating far fewer parameters ■ Example: 12 parameters under the assumption that each covariate's contribution to the mean of Y is additive

Need to use parametric estimators in the presence of high-dimensionality

- ☐ Data may be high-dimensional because
 - many categorical variables
 - continuous variables
 - (time-varying variables)
 - all of the above

Outcome Regression

In summary, assumptions for causal inference with parametric models	
□ Exchangeability	
□ Positivity	
☐ Consistency (including well-defined interventions)	
☐ No model misspecification	
Outcome Regression	37

Assumptions needed for causal inference with models	
☐ Identifiability assumptions	
The assumptions that we would have to make even if we had an infinite amount of data	3
 Exchangeability, Positivity, Consistency (including well-defined interventions) 	
 Others for instrumental variable estimation 	
□ Modeling assumptions	
The assumptions that we have to make because we do not have an infinite amount of data	
☐ No model misspecification	
Outcome Regression 3	 38

Exchangeability assumption

If individuals with A=1 had had A=0, they would have had the same mean outcome as those who actually had A=0

and vice versa

The above has to be true for every subgroup of individuals with a different covariate pattern

■ Men age 50 with history of diabetes, women age 63 without history of diabetes, etc.

Outcome Regression

39

Positivity assumption

In each subgroup of the population defined by a covariate pattern,

Men age 50 with history of diabetes, women age 63 without history of diabetes, etc.

There must be some individuals with A=1 and some individuals with A=0

- The probability of treatment (and of no treatment) must be greater than zero in all levels of the confounders, i.e., positive
- We will take this condition for granted during this course

Outcome Regression

Consistency assumption

The interventions of interest (e.g., smoking cessation) must be sufficiently well-defined, and they need to correspond to the ones present in the data (e.g., A=1)

■ We will take this condition for granted during this course

Outcome Regression

41

If we had an infinite amount of data and the identifiability conditions held

- ☐ We could calculate the average causal effect of treatment on the outcome directly from the data
- ☐ That is, we could **identify** the average causal effect

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

Outcome Regression

But we never have an infinite amount of data

Therefore, we also need to make modeling assumptions regarding:

How covariates in the model relate to the outcome and/or the treatment

If the identifiability assumptions hold and our modeling assumptions happen to be correct, we can **estimate** the causal effect without bias

Outcome Regression

43

Examples of modeling assumptions (I)

- ☐ Continuous variable: The relation between the mean outcome and age is a
 - straight line, i.e., model includes only a linear term for age
 - curve, e.g., model includes also a quadratic (squared) term for age
 - step function, e.g., model includes indicators for quintiles of age
- □ Categorical variable: The relation between the mean outcome and education is a step function, with steps between categories of
 - same size, e.g., for a variable education with three levels, we can assume that the distance from level 1 to level 2 is the same as from level 2 to level 3
 - different size

Outcome Regression

Examples of modeling assumptions (II) Product terms ("interactions")

- □ Model includes no product terms between sex and diabetes
 - The contributions of sex and diabetes to the mean outcome are additive
- □ Model includes a product term between sex and diabetes
 - No assumptions about how sex and diabetes jointly contribute to the mean outcome

Outcome Regression

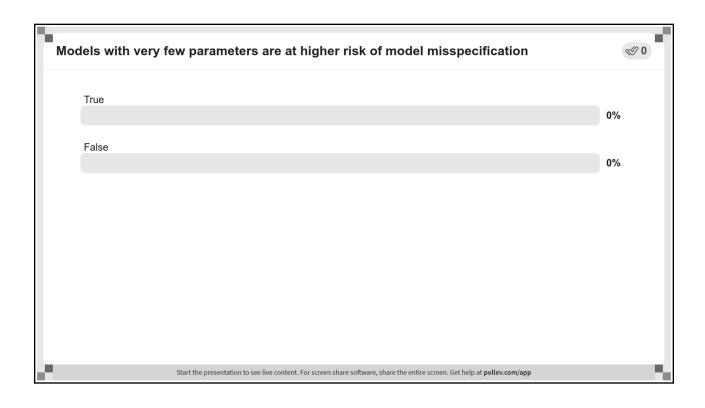
45

Bias-variance tradeoff

- \square A nonparametric estimator of $\mathrm{E}[Y|A,L]$ will not introduce bias because of incorrect modeling assumptions
 - There are no modeling assumptions
 - But estimates will be highly unstable (high variance)
- \square A parametric estimator of E[Y|A,L] may introduce bias because of incorrect modeling assumptions
 - But, IF the parametric model is correctly specified, it brings huge gains in variance

Outcome Regression

Models w	ith very few parameters make strong modeling assumptions	₩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	- 4



Models with very	few parameters yield more precise estimates	₩ 0
(A) True		0%
(B) False		076
		0%
	Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app	

Causal of smol	effect king cessation on death	
□ You w	ill estimate it in Homework 1	
■ All o	an use a logistic regression model of the above discussion applies except that erences of means can be replaced by odds ratios other effect measures	
	Outcome Regression	50

Progress report

- 1. Introduction to modeling
- 2. Stratified analysis: Outcome regression
- 3. Stratified analysis: Propensity scores

Outcome Regression