

Biost 518

Applied Biostatistics II

.....

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 3: Confounding, Effect Modification

January 8, 2007

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

Scientific Questions

-
- Most times:
 - Comparing distribution of response across groups defined by predictor of interest
 - Very often, other variables also need to be considered because
 - Comparison is different in strata
 - Groups being compared differ in other ways
 - Less variability of response if we control for other variables

2

Statistical Role

-
- Covariates other than the POI are included in the model as
 - Effect modifiers
 - Confounders
 - Precision variables

3

Effect Modification

.....

4

Effect Modifier

- The association between Response and POI differs in strata defined by effect modifier
 - Statistical term: “Interaction”
 - Depends on the measurement of effect
 - Summary measure
 - Mean, geometric mean, median, proportion, odds, hazard, etc.
 - Comparison across groups
 - Difference, ratio

5

Effect Modifier: Example 1a

- Serum LDL by sex (modified by smoking?)
 - Yes for mean, not so much for median
 - Difference or ratio

	<u>Mean</u>		<u>Median</u>	
	<u>Nsmk</u>	<u>Smk</u>	<u>Nsmk</u>	<u>Smk</u>
Men	120	122	120	115
Women	133	122	133	124
Difference	-13	0	-13	-9
Ratio	0.90	1.00	0.90	0.93 ₆

Effect Modifier: Example 1b

- Creatinine by stroke (modified by sex?)
 - Yes for difference, not so much for ratio
 - Mean or median

	<u>Mean</u>		<u>Median</u>	
	<u>Women</u>	<u>Men</u>	<u>Women</u>	<u>Men</u>
No stroke	0.72	1.08	0.7	1.1
Stroke	1.01	1.51	1.0	1.5
Diff	-0.29	-0.43	-0.3	-0.4
Ratio	0.71	0.72	0.70	0.73 ₇

Effect Modifier: Example 2a

- Stroke by smoking (modified by sex?)
 - Proportion: No for ratio, more for difference
 - Odds: Yes for difference, less for ratio

	<u>Proportion</u>		<u>Odds</u>	
	<u>Women</u>	<u>Men</u>	<u>Women</u>	<u>Men</u>
Nonsmok	0.10	0.16	0.03	0.19
Smoke	0.16	0.26	0.19	0.35
Diff	-0.06	-0.10	-0.10	-0.26
Ratio	0.62	0.62	0.47	0.54 ₈

Effect Modifier: Example 2b

- Stroke by smoking (modified by ASCVD?)
 - Proportion: Yes for difference, yes for ratio
 - Odds: Yes for difference, no for ratio

	<u>Proportion</u>		<u>Odds</u>	
	<u>None</u>	<u>ASCVD</u>	<u>None</u>	<u>ASCVD</u>
Nonsmok	0.02	0.33	0.02	0.50
Smoke	0.04	0.50	0.04	1.00
Diff	-0.02	-0.17	-0.02	-0.50
Ratio	0.50	0.67	0.50	0.50

Effect Modifier: Example 2c

- CHD by smoking (modified by sex?)
 - Proportion: Yes for difference, yes for ratio
 - Odds: Yes for difference, yes for ratio

	<u>Proportion</u>		<u>Odds</u>	
	<u>Women</u>	<u>Men</u>	<u>Women</u>	<u>Men</u>
Nonsmok	0.18	0.26	0.22	0.35
Smoke	0.05	0.24	0.05	0.32
Diff	0.13	0.02	0.17	0.03
Ratio	3.60	1.08	4.17	1.11

Effect Modifier: Example 2d

- CHD by ever smoke (modified by sex?)
 - Proportion: No for difference, no for ratio
 - Odds : No for difference, no for ratio

	<u>Proportion</u>		<u>Odds</u>	
	<u>Women</u>	<u>Men</u>	<u>Women</u>	<u>Men</u>
Never	0.16	0.25	0.19	0.33
Ever	0.16	0.26	0.19	0.35
Diff	0.00	-0.01	0.00	-0.02
Ratio	1.00	0.96	1.00	0.95

Aside: Be Careful with Ratios

- How close are two ratios?
 - 0.20 and 0.25 VERSUS 5.0 and 4.0 ?
 - 0.10 and 0.15 VERSUS 10.0 and 6.7 ?
- We might tend to consider a bigger difference when two ratios are each > 1 than when they are each < 1
 - "But that would be wrong."

Analysis of Effect Modification

.....

- When the scientific question involves effect modification, analyses must be within each stratum separately
 - If we want to estimate degree of effect modification or test for its existence:
 - A regression model will typically include
 - Predictor of interest (main effect)
 - Effect modifying variable (main effect)
 - A covariate modeling the interaction (usually product)

13

Ignoring Effect Modification

.....

- By design or mistake, we sometimes do not model effect modification
 - We might perform either
 - Unadjusted analysis:
 - POI only
 - Adjusted analysis:
 - POI and third variable, but no interaction term

14

Unadjusted Analyses

.....

- If effect modification exists, an unadjusted analysis will give different results according to the association between the POI and effect modifier in the sample
 - If POI, effect modifier not associated:
 - Unadjusted analysis tends toward some sort of weighted average of stratum specific effects
 - With means, exactly; with odds, hazards approximately
 - If POI, effect modifier associated in sample:
 - “Average effect” is confounded

15

Adjusted Analyses

.....

- If effect modification exists, an analysis adjusting only for the third variable (but not interaction) will tend toward a weighted average of the stratum specific effects
 - Hence, an association in one stratum and not the other will make an adjusted analysis look like an association
 - (providing sample size is large enough)

16

Confounding

.....

17

Simpson's Paradox

.....

- Given binary variables Y (response), X (POI), Z (strata), it is possible to have

$$\Pr(Y=1 | X=1, Z=1) > \Pr(Y=1 | X=0, Z=1)$$

$$\Pr(Y=1 | X=1, Z=0) > \Pr(Y=1 | X=0, Z=0)$$

but to have

$$\Pr(Y=1 | X=1) < \Pr(Y=1 | X=0)$$

18

Avoiding Simpson's Paradox

.....

- Suppose

$$\Pr(Y=1 | X=1, Z=1) > \Pr(Y=1 | X=0, Z=1)$$

$$\Pr(Y=1 | X=1, Z=0) > \Pr(Y=1 | X=0, Z=0)$$

- If either

$$\Pr(X=x, Z=z) = \Pr(X=x) \Pr(Z=z) \quad (X, Z \text{ indep})$$

OR

$$\Pr(Y=y, Z=z | X=1) = \Pr(Y=y | X=1) \Pr(Z=z | X=1) \quad (Y, Z \text{ cond indep})$$

then we must have

$$\Pr(Y=1 | X=1) > \Pr(Y=1 | X=0)$$

19

Confounding

.....

- Definition of confounding

– The association between a predictor of interest and the response variable is confounded by a third variable if

- The third variable is associated with the predictor of interest in the sample, AND
- The third variable is associated with the response
 - causally (in truth)
 - in groups that are homogeneous with respect to the predictor of interest, and
 - not in the causal pathway of interest

20

Adjustment for Covariates

.....

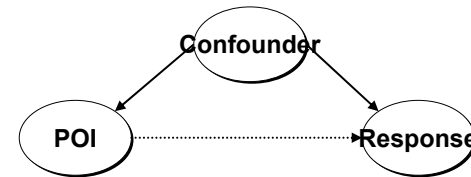
- We must consider our beliefs about the causal relationships among the measured variables
 - We will not be able to assess causal relationships in our statistical analysis
 - Inference of causation comes only from study design
 - However, consideration of hypothesized causal relationships helps us decide which statistical question to answer

21

Classical Confounder

.....

- A clear case of confounding is when some third variable is a “cause” of both the POI and response
 - We generally adjust for such a confounder

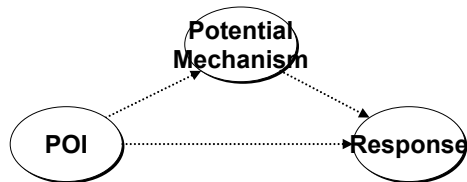


22

Causal Pathway

.....

- A variable in the causal pathway of interest is not a confounder
 - We would not adjust for such a variable (lest we lose ability to detect the effect)

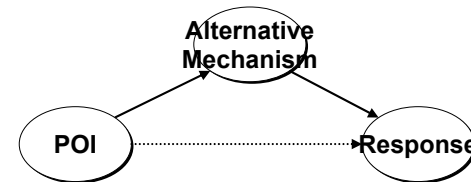


23

Causal Pathway

.....

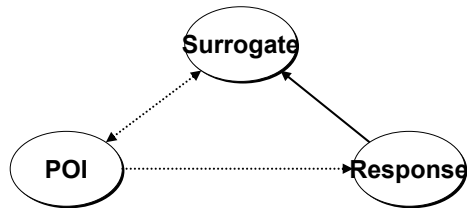
- We would want to adjust for a variable in a causal pathway not of interest
 - E.g., work stress causing ulcers by hormonal effects versus alcoholism



24

Surrogate for Response

- Adjustment for such a variable is a very BAD thing to do



25

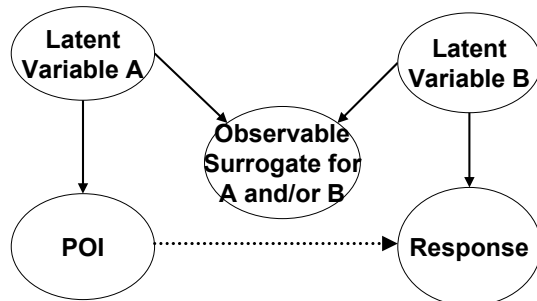
Unadjusted, Adjusted Analyses

- Confounding typically produces a difference between unadjusted and adjusted analyses, but those symptoms are not proof of confounding
 - Such a difference can occur times when there is no confounding

26

Complicated Causal Pathway

- Adjustment for Variable C would produce a spurious association (effect modification)



27

Symptoms of Confounding

- Estimates of association from unadjusted analysis are markedly different from estimates of association from adjusted analysis
 - Association within each stratum is similar to each other, but different from the association in the combined data

28

Nonlinear Summary Measures

.....

- Summary measures which are nonlinear functions of the mean sometimes show the above symptoms in the absence of confounding
 - Odds (and odds ratios)
 - Hazards (and hazard ratios)

29

Inference on Means

.....

- In linear regression, differences between adjusted and unadjusted analyses are diagnostic of confounding
 - Precision variables tend to change standard errors but not slope estimates
 - Effect modification would show differences between adjusted analysis and unadjusted analysis, but would also show different associations in the different strata

30

Inference on Odds, Hazards

.....

- In logistic and PH regression, differences between adjusted and unadjusted analyses are more difficult to judge
 - Comparisons in more homogeneous groups (i.e., after adjustment for a precision variable) drive slope estimates to the extreme (away from the null)

31

Precision Variables

.....

32

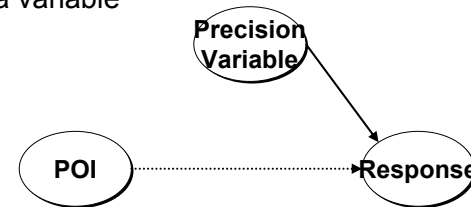
Precision

- Sometimes we choose the exact scientific question to be answered on the basis of which question can be answered most precisely
 - In general, questions can be answered more precisely if the within group distribution is less variable
 - Comparing groups that are similar with respect to other important risk factors decreases variability

33

Precision Variable

- The third variable is an independent “cause” of the response
 - We tend to gain precision if we adjust for such a variable



34

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

Typically : $se(\hat{\theta}) = \sqrt{\frac{V}{n}}$

(V related to average "statistical information")

Width of CI : $2 \times (crit\ val) \times se(\hat{\theta})$

Test statistic : $Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

35

Increasing Precision

- Options
 - Increase sample size
 - Decrease V
 - (Decrease confidence level)

36

Ex: Difference of Indep Means

.....

$$\text{ind } Y_{ij} \sim (\mu_i, \sigma_i^2), i = 1, 2; j = 1, \dots, n_i$$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = \mu_1 - \mu_2 \quad \hat{\theta} = \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$$

$$V = (r+1) \left[\sigma_1^2 / r + \sigma_2^2 \right] \quad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

37

Controlling Variation

.....

- In a two sample comparison of means, we might control some variable in order to decrease the within group variability
 - Restrict population sampled
 - Standardize ancillary treatments
 - Standardize measurement procedure

38

Ex: Linear Regression

.....

$$\text{ind } Y_i | X_i \sim (\beta_0 + \beta_1 \times X_i, \sigma_{Y|X}^2), i = 1, \dots, n$$

$$\theta = \beta_1 \quad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma_{Y|X}^2}{\text{Var}(X)} \quad se(\hat{\theta}) = \sqrt{\frac{\sigma_{Y|X}^2}{n \text{Var}(X)}}$$

39

Adjusting for Covariates

.....

- When comparing means using stratified analyses or linear regression, adjustment for precision variables decreases the within group standard deviation
 - Var (Y | X) vs Var (Y | X, W)

40

Ex: Linear Regression

ind $Y_i | X_i, W_i \sim (\beta_0 + \beta_1 \times X_i + \beta_2 \times W_i, \sigma_{Y|X,W}^2), i = 1, \dots, n$

$\theta = \beta_1 \quad \hat{\theta} = \hat{\beta}_1$ from LS regression

$$V = \frac{\sigma_{Y|X,W}^2}{\text{Var}(X)(1-r_{XW}^2)} \quad se(\hat{\theta}) = \sqrt{\frac{\sigma_{Y|X,W}^2}{n\text{Var}(X)(1-r_{XW}^2)}}$$

$$\sigma_{Y|X,W}^2 = \sigma_{Y|X}^2 - \beta_2^2 \text{Var}(W | X)$$

41

Precision with Proportions

- When analyzing proportions (means), the mean variance relationship is important
 - Precision is greatest when proportion is close to 0 or 1
 - Greater homogeneity of groups makes results more deterministic
 - (At least, I always hope for this)

42

Ex: Diff of Indep Proportions

ind $Y_{ij} \sim B(1, p_i), i = 1, 2; j = 1, \dots, n_i$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$\theta = p_1 - p_2 \quad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$

$$\sigma_i^2 = p_i(1 - p_i)$$

$$V = (r+1)[\sigma_1^2 / r + \sigma_2^2] \quad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

43

Precision with Odds

- When analyzing odds (a nonlinear function of the mean), adjusting for a precision variable results in more extreme estimates
 - odds = $p / (1-p)$
 - odds using average of stratum specific p is not the average of stratum specific odds

44

Hypothetical Example

- Stroke by smoking (in ASCVD strata)
 - No association between smoking and ASCVD in the sample: 10% smokers in each group
 - Not confounder (but clearly a precision variable)
 - Unadjusted OR “attenuated toward null”

	<u>No ASCVD</u>			<u>ASCVD</u>			<u>Combined</u>		
	<u>N</u>	<u>p</u>	<u>odds</u>	<u>N</u>	<u>p</u>	<u>odds</u>	<u>N</u>	<u>p</u>	<u>odds</u>
Smok	1000	0.04	0.04	100	0.50	1.00	1100	0.082	0.089
Nonsmok	10000	0.02	0.02	1000	0.33	0.50	11000	0.048	0.051
Ratio	OR= 2.00			OR= 2.00			OR= 1.75 ⁴⁵		

Diagnosing Confounding

Descriptive Statistics

46

Adjustment for Covariates

- We include predictors in an analysis for a variety of reasons
 - In order of importance
 - Scientific question
 - Predictor(s) of interest
 - Effect modifiers
 - Adjustment for confounding
 - Gain precision

47

Adjustment for Covariates

- Adjustment for covariates changes the question being answered by the statistical analysis
 - Adjustment can be used to isolate associations that are of particular interest

48

Adjustment for Covariates

.....

- When I consult with a scientist, it is often very difficult to decide whether the interest in additional covariates is due to confounding, precision, or effect modification
 - I tend to treat these variables differently in a statistical analysis

49

Scientific Question

.....

- Many times the scientific question dictates inclusion of particular predictors
 - Predictor(s) of interest
 - The scientific factor being investigated can be modeled by multiple predictors
 - » E.g., dummy variables, polynomials, etc.
 - Effect modifiers
 - The scientific question may relate to detection of effect modification
 - Confounders
 - The scientific question may have been stated in terms of adjusting for known (or suspected) confounders

50

Unanticipated Confounding

.....

- Other times, we explore our data to assess whether our results were confounded by some variable
 - Assessing the “independent effect” of the predictor of interest

51

Confounders

.....

- Variables (causally) predictive of outcome, but not in the causal pathway of interest
 - (Often assessed in the control group but thinking is best)
- Variables associated with the predictor of interest in the sample
 - Note that statistical significance is not relevant, because that tells us about associations in the population
- Detection must ultimately rely on our best knowledge about possible mechanisms

52

Confounding

.....

- Effect of confounding
 - A confounder can make the observed association between the predictor of interest and the response variable look
 - stronger than the true association,
 - weaker than the true association, or
 - even the reverse of the true association
 - “Qualitative confounding”

53

Diagnosing Confounding

.....

- Stratified analyses to distinguish between
 - Effect modifiers
 - Confounders
 - Precision variables

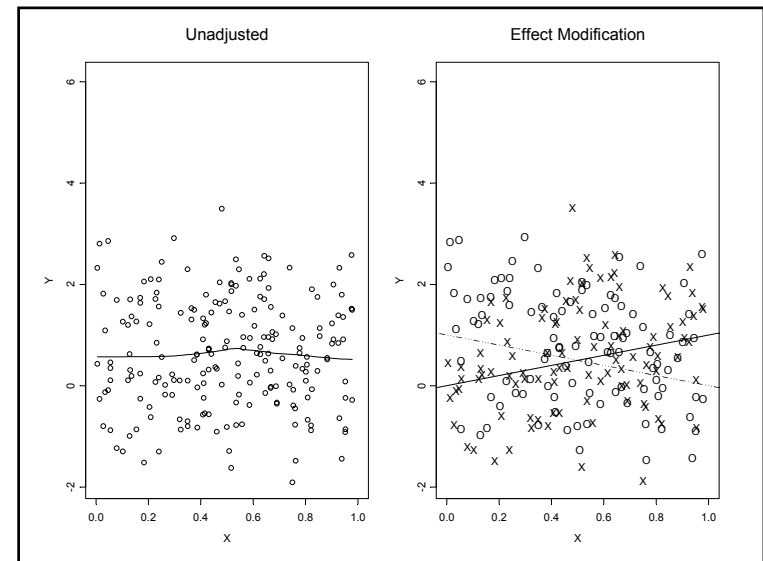
54

Effect modifiers

.....

- Estimates of treatment effect differ among the strata
 - When analyzing difference of means of continuous data
 - Stratified smooth curves of data are nonparallel
 - (Graphical techniques difficult in other settings)

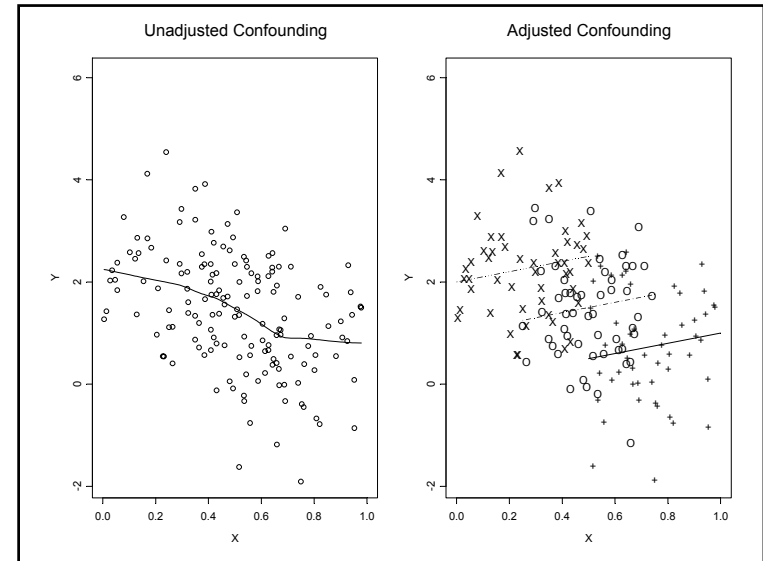
55



Confounders

- Estimates of treatment effect the same across strata, AND
 - Confounder is causally associated with Response, AND
 - Confounder associated with POI in the sample
- When analyzing difference of means of continuous data
 - Stratified smooth curves of data are parallel
 - Distribution of POI differs across strata
 - Unadjusted, adjusted analyses give different estimates

57



Precision Variables

- Estimates of treatment effect the same across strata, AND
 - Variable is causally associated with Response, AND
 - Variable not associated with POI in the sample
- When analyzing difference of means of continuous data
 - Stratified smooth curves of data are parallel
 - Distribution of POI same across strata
 - Unadjusted, adjusted analyses give similar estimates

59

