Biost 518
Applied Biostatistics II

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

Lecture 3:
Confounding, Effect Modification

January 9, 2006

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

Statistical Role

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

Effect Modification

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

Effect Modifier: Example 1a

- Serum LDL by sex (modified by smoking?)

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

Effect Modifier: Example 1b

- Creatinine by stroke (modified by sex?)

	Mean		Median	
	Women	Men	Women	Men
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

Effect Modifier: Example 2a

- Stroke by smoking (modified by sex?)

	Proportion		Odds	
	Women	Men	Women	Men
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

Effect Modifier: Example 2b

- Stroke by smoking (modified by ASCVD?)

	Proportion		Odds	
	None	ASCVD	None	ASCVD
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		Odds	
	Women	Men	Women	Men
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		Odds	
	Women	Men	Women	Men
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
      - Effect modifier
      - A covariate modeling the interaction (usually product)

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

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)

Confounding

Simpson's Paradox

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

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

but to have

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

Avoiding Simpson's Paradox

- Suppose

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

- Then if either

$$\Pr(X=x, Z=z) = \Pr(X=x) \Pr(Z=z) \text{ OR}$$
$$\Pr(Y=y, Z=z \mid X=1) = \Pr(Y=y \mid X=1) \Pr(Z=z \mid X=1)$$

then we must have

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

Confounding

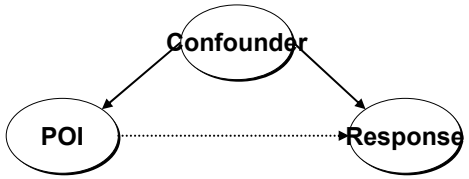
- Definition of confounding
  - The association between a predictor of interest and the response variable is confounded by a third variable if
    - The third variable is associated with the predictor of interest in the sample, AND
    - The third variable is associated with the response
      - causally (in truth)
      - in groups that are homogeneous with respect to the predictor of interest, and
      - not in the causal pathway of interest

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

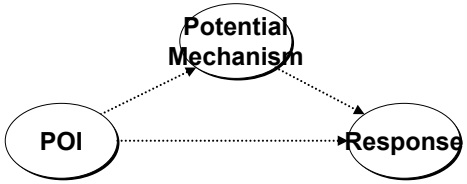
Classical Confounder

- 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



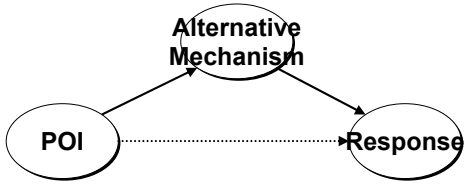
Causal Pathway

- 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)



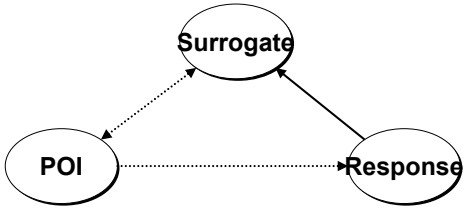
Causal Pathway

- 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



Surrogate for Response

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

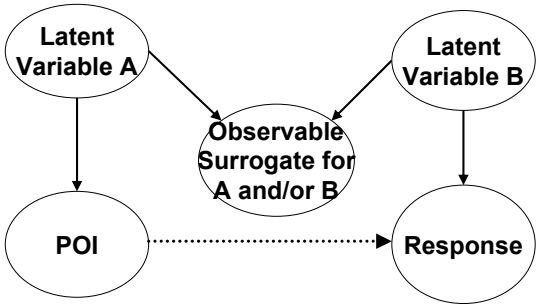


Unadjusted, Adjusted Analyses

- 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

Complicated Causal Pathway

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



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)

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

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

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)

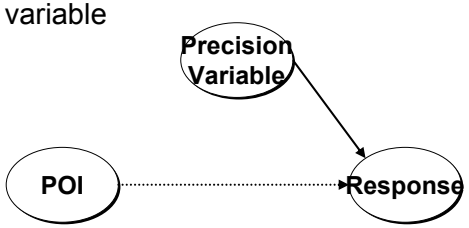
Precision Variables

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

Precision Variable

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



Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

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

$$(V \text{ related to average "statistical information"})$$

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

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

Increasing Precision

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

Ex: Difference of Indep Means

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

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

$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

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

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

Ex: Linear Regression

$$ind\ Y_i \mid X_i \sim \left(\beta_0 + \beta_1 \times X_i, \sigma^2_{Y|X}\right), i = 1, \ldots, n$$

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

$$V = \frac{\sigma^2_{Y|X}}{Var(X)} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2_{Y|X}}{nVar(X)}}$$

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)

Ex: Linear Regression

$$ind\ Y_i \mid X_i, W_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_2 \times W_i, \sigma^2_{Y|X,W}\right), i = 1, \ldots, n$$

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

$$V = \frac{\sigma^2_{Y|X,W}}{Var(X)\left(1 - r^2_{XW}\right)} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2_{Y|X,W}}{nVar(X)\left(1 - r^2_{XW}\right)}}$$

$$\sigma^2_{Y|X,W} = \sigma^2_{Y|X} - \beta_2^2 Var(W \mid X)$$

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)

Ex: Diff of Indep Proportions

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

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

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

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

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

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

Hypothetical Example

- Stroke by smoking (in ASCVD strata)
  - No association between smoking and ASCVD in the sample
    - Not confounder (but clearly a precision variable)

	No ASCVD			ASCVD			Combined		
	N	p	odds	N	p	odds	N	p	odds
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	

Diagnosing Confounding

Descriptive Statistics

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

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

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

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

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

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

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"

Diagnosing Confounding

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

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)

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

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