Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 7:
Covariate Adjustment

February 2, 2015

1

Lecture Outline

- Toy Example
- Effect Modification
- Confounding
- Precision
- Adjustment for Covariates
  - Borrowing Information Across Strata

2

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

3

Statistical Role

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

4

Toy Example

5

Example

- A hypothetical agricultural experiment is conducted to assess the effect of fertilizer on the size of fruit produced
- Plants are randomly assigned to receive either fertilizer or a sham treatment
  - Randomization in some sense precludes the possibility of confounding
- Response variable
  - At the end of the study, the diameter of the fruit produced by the plants is measured.

6

Example: Predictor of Interest

- The scientific question translates into a statistical question comparing the distribution of fruit sizes across groups defined by fertilizer treatment
- Predictor of interest:
  - A binary variable indicating whether the corresponding fruit was obtained from a plant that received fertilizer (1) or a sham treatment (0)

Example: Hypothetical Data (Case 1)

Fruit sizes by treatment group

	Fert	Sham	Diff
	3.7, 12.5,	41.6, 10.3,	
	13.7, 44.2,	0.9, 40.5,	
	43.8, 43.5,	9.8, 10.2,	
	4.3, 14.0,	11.1, 1.1,	
	4.6, 43.9,	39.9, 1.3,	
	13.8, 4.2	40.7, 1.4	
Mean	20.5	17.4	3.1
SD	17.7	17.6	

8

Example: Conclusions (Case 1)

- No conclusive evidence that fertilizer improves size
- The difference in the average size of fruit (mean difference 3.1) was not very large compared to the variability in the size of the fruit within groups
  - Var (Size | Trt) = 311.5 (SD = 17.65)
  - (P value = 0.67)
- Thus with these small sample sizes, we cannot rule out that the difference in means was not just a chance observation when no real effect exists
  - A larger sample size might make such an observed difference conclusive
  - BUT: With a larger sample size, the greater precision might make the estimate closer to the truth, which might be the null

9

Example: Adjusted Analysis (Case 1)

Fruit sizes by treatment group and type of fruit

	Fert	Sham	Diff
Berry	3.7, 4.3,	0.9, 1.1,	
	4.6, 4.2	1.3, 1.4	
Mean(SD)	4.2 (0.37)	1.2 (0.22)	3.0
Apple	13.8, 12.5,	9.8, 10.2,	
	13.7, 14.0,	11.1, 10.3,	
Mean(SD)	13.5 (0.68)	10.4 (0.54)	3.1
Melon	44.2, 43.8,	41.6, 40.5,	
	43.5, 43.9	39.9, 40.7	
Mean(SD)	43.8 (0.29)	40.7 (0.70)	3.1

10

Combining Estimates Across Strata

- Comparisons across strata or studies
- This is easy, if estimates are independent and approximately normally distributed
- Weighted averages using importance weights, weights by sample size, or weights to achieve greatest precision
  - Regression models are essentially computing stratified estimates

For independent strata

$$\hat{\theta}_1 \dot{\sim} N\left(\theta_1, se_1^2\right); \quad \hat{\theta}_2 \dot{\sim} N\left(\theta_2, se_2^2\right)$$

Weighted average : For $w_1 + w_2 = 1$

$$\hat{\theta} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2 \dot{\sim} N\left(w_1\theta_1 + w_2\theta_2, \quad w_1^2 se_1^2 + w_2^2 se_2^2\right)$$

11

Example: Adjusted Conclusions (Case 1)

- This second analysis suggests very conclusive evidence that fertilizer improves size of fruit
- More precision was gained by comparing similar types of fruits ("Apples with apples")
  - Var (Size | Trt, Fruit) = 0.25 (SD = 0.50)
- The average difference of 3.1 across types of fruit is large compared to the within group standard deviation of 0.50
  - (P value < .0001)
- (Randomization did protect us from confounding: Each treatment group had four plants of each kind)

Example: Case 2 - Confounding

- We can use this example to illustrate how confounding would appear different
- In Case 1, we imagined that randomization worked perfectly (perhaps we stratified on type of plant)
- If we used complete randomization, we might have been unlucky and ended up with imbalance between treatment groups with respect to type of plant

Example: Hypothetical Data (Case 2)

Fruit sizes by treatment group

	Fert	Sham	Diff
	3.7, 12.5,	41.6, 10.3,	
	13.7, 44.2,	0.9, 40.5,	
	3.8, 43.5,	9.8, 10.2,	
	4.3, 14.0,	11.1, 1.1,	
	4.6, 43.9,	39.9, 41.3,	
	13.8, 4.2	40.7, 1.4	
Mean	17.2	20.7	-3.5
SD	16.6	18.1	

14

Example: Conclusions (Case 2)

- No conclusive evidence that fertilizer improves size of fruit

- The difference in the average size of fruit (mean difference -3.1) was not very large compared to the variability in the size of the fruit (standard deviation 16.6 and 18.1 in the two groups)
  - (P value = 0.62)

- In fact, the point estimate of treatment effect actually suggests that the fertilizer treatment makes things worse

15

Example: Adjusted Analysis (Case 2)

Fruit sizes by treatment group and type of fruit

	Fert	Sham	Diff
Berry	3.7, 4.3, 3.8, 4.6, 4.2	0.9, 1.1, 1.4	
Mean(SD)	4.1 (0.37)	1.1 (0.25)	3.0
Apple	13.8, 12.5, 13.7, 14.0,	9.8, 10.2, 11.1, 10.3,	
Mean(SD)	13.5 (0.68)	10.4 (0.54)	3.1
Melon	44.2, 43.5, 43.9	41.6, 40.5, 41.3,39.9,40.7	
Mean(SD)	43.9 (0.35)	40.8 (0.67)	3.1

16

Example: Adjusted Conclusions (Case 2)

- This second analysis suggests very conclusive evidence that fertilizer improves size of fruit
- More accuracy was gained by comparing similar types of fruits (“Apples with apples”)
  - In this case, also gained precision, though not as much as when fruit type was balanced
- The average difference of 3.1 across types of fruit is large compared to the standard deviations with groups defined by type of fruit and treatment
  - ($P < .0001$)

17

Example: Case 3 – Effect Modification

- We can also use this example to illustrate how effect modification would appear different
- In Cases 1 and 2, we imagined that the treatment worked equally well for all types of fruit
- We can now examine what would happen if that were not the case

Example: Hypothetical Data (Case 3)

Fruit sizes by treatment group

	Fert	Sham	Diff
	3.7, 12.5,	45.6, 10.3,	
	13.7, 44.2,	0.9, 44.5,	
	43.8, 43.5,	9.8, 10.2,	
	4.3, 14.0,	11.1, 1.1,	
	4.6, 43.9,	43.9, 1.3,	
	13.8, 4.2	44.7, 1.4	
Mean	20.5	18.7	1.8
SD	17.7	19.6	

19

Example: Conclusions (Case 3)

- No conclusive evidence that fertilizer improves size of fruit
- The difference in the average size of fruit (mean difference 1.8) was not very large compared to the variability in the size of the fruit (standard deviation 17.6 and 19.6 in the two groups)
  - (P value = 0.82)
- Thus with these small sample sizes, we cannot rule out that the difference in means was not just a chance observation when no real effect exists
  - A larger sample size might make such an observed difference conclusive
  - BUT: With a larger sample size, the greater precision might make the estimate closer to the truth, which might be the null

20

Example: Adjusted Analysis (Case 3)

Fruit sizes by treatment group and type of fruit

	Fert	Sham	Diff
Berry	3.7, 4.3, 4.6, 4.2	0.9, 1.1, 1.3, 1.4	
Mean(SD)	4.2 (0.37)	1.2 (0.22)	3.0
Apple	13.8, 12.5, 13.7, 14.0,	9.8, 10.2, 11.1, 10.3,	
Mean(SD)	13.5 (0.68)	10.4 (0.54)	3.1
Melon	44.2, 43.8, 43.5, 43.9	45.6, 44.5, 43.9, 44.7	
Mean(SD)	43.8 (0.29)	44.7 (0.70)	-0.8

21

Example: Adjusted Conclusions (Case 3)

- A stratified analysis suggests the question about fertilizer effect should be answered by stratum
- Two basic approaches to analysis are possible
- Average the stratum specific effect of fertilizer across strata
  - Treatment effect of 1.8 is large compared to within group variation (P=.0009)
- Analyze each stratum separately
  - Improvement of 3.1 for berries, apples is large compared to within group variation (P <.0001, P<.0001)
  - Decrease of 0.8 for melons is marginal (P=0.032 without adjustment for multiple comparisons)

22

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?)
  - Yes for mean, not so much for median
  - Difference or ratio

	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

25

Effect Modifier: Example 1b

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

	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: No for ratio, more for difference
  - Odds: Yes for difference, less for ratio

	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

27

Effect Modifier: Example 2b

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

	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

28

Effect Modifier: Example 2c

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

	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

29

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

	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

30

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."

31

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)

32

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

33

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 in sample:
  - 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

34

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)

35

Confounding

36

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

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

37

Avoiding Simpson's Paradox

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

- If either

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

OR

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

then we must have

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

38

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

39

Adjustment for Covariates

- We must consider our beliefs about the causal relationships among the measured variables

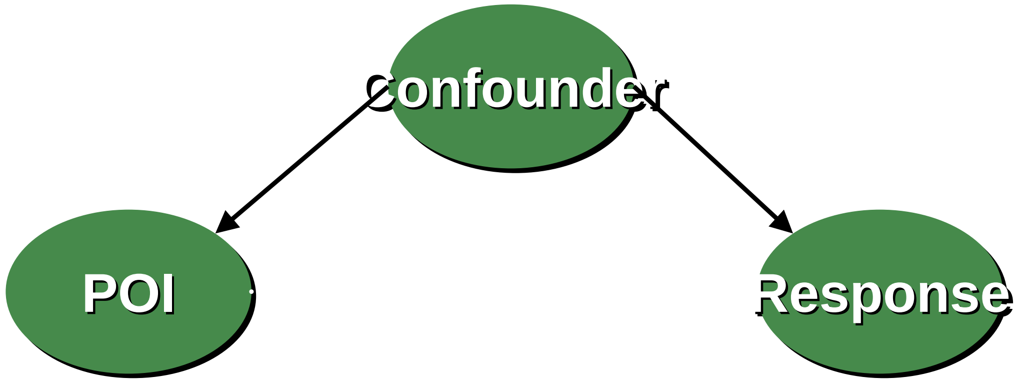
- We will not be able to assess causal relationships in our statistical analysis
  - Inference of causation comes only from study design

- However, consideration of hypothesized causal relationships helps us decide which statistical question to answer

40

Classical Confounder

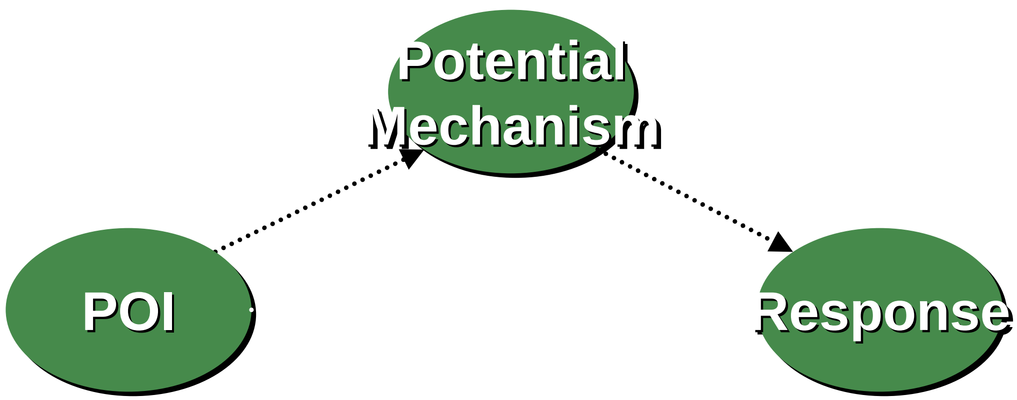
- A clear case of confounding is when some third variable is a “cause” of both the POI and response
- We generally adjust for such a confounder



41

Causal Pathway

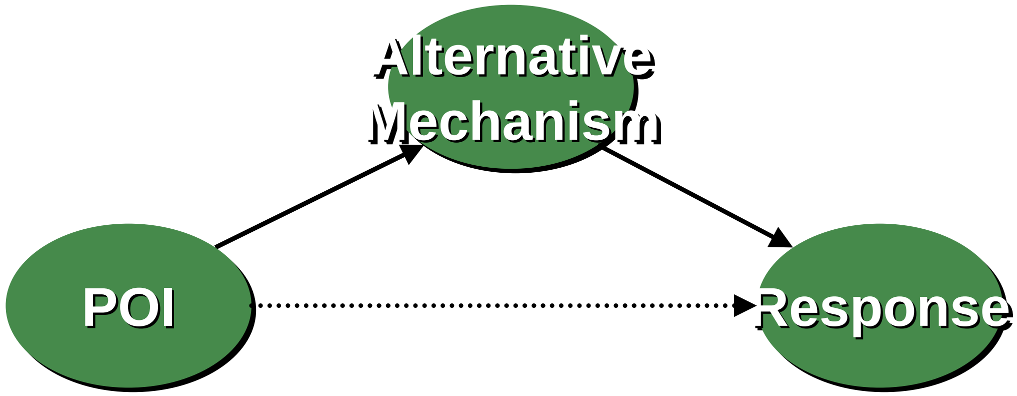
- A variable in the causal pathway of interest is not a confounder
- We would not adjust for such a variable (lest we lose ability to detect the effect)



42

Causal Pathway

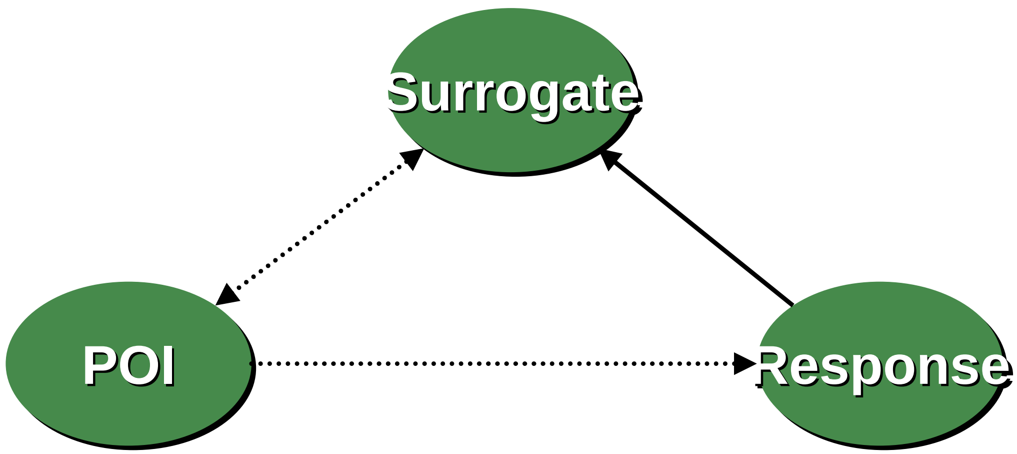
- We would want to adjust for a variable in a causal pathway not of interest
- E.g., work stress causing ulcers by hormonal effects versus alcoholism



43

Surrogate for Response

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



44

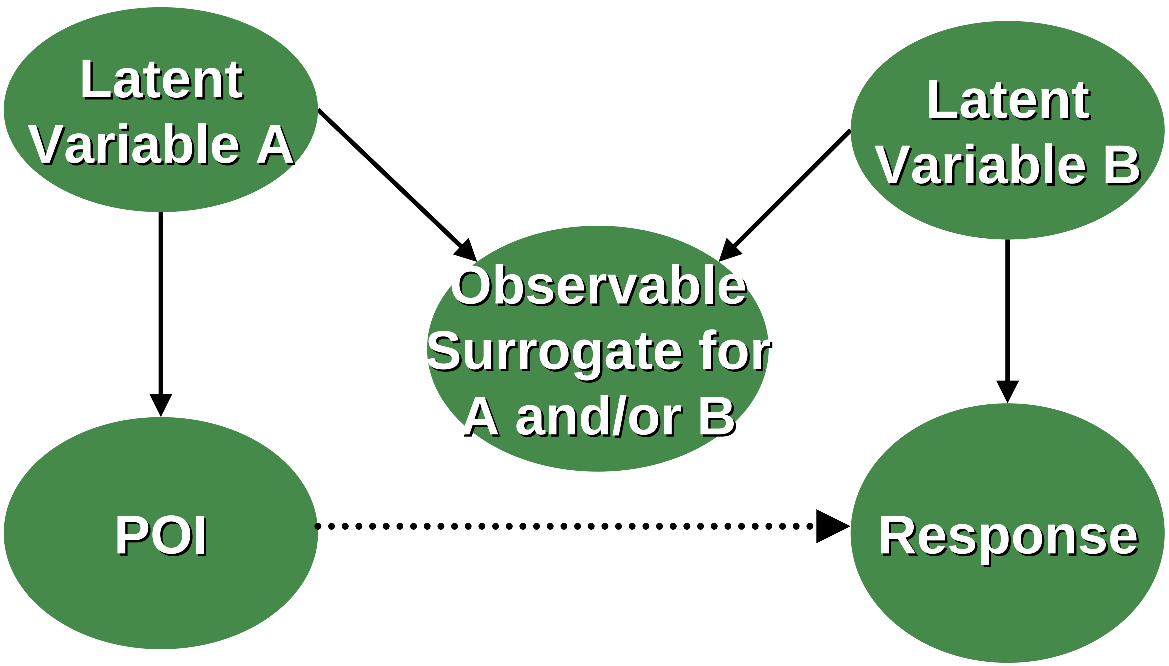
Unadjusted, Adjusted Analyses

- 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)



46

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

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)

- Such measures are called “noncollapsible”
  - Collapsing the strata into one group does not give the same answer, even if there is no confounding
  - (This happens when trying to “collapse” over a precision variable)

48

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

51

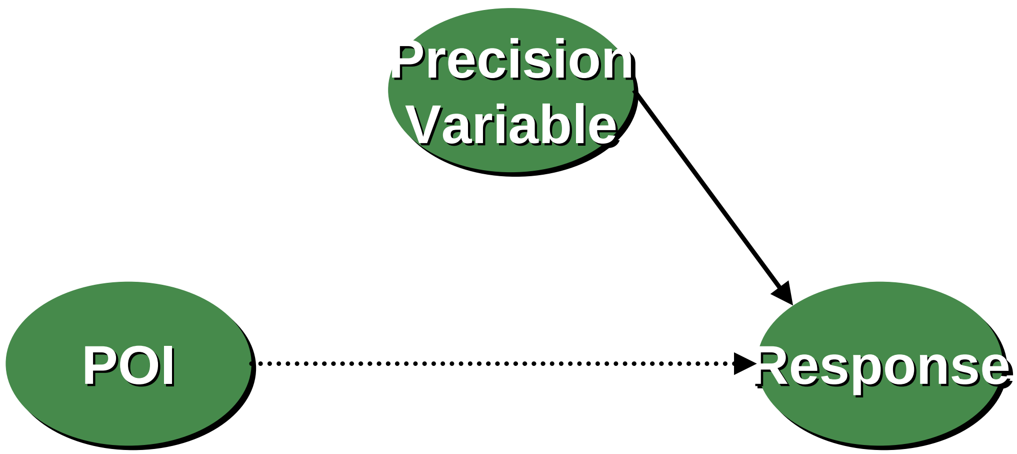
Precision

- 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

52

Precision Variable

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



53

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

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

(V related to average "statistical information")

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

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

54

Increasing Precision

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

55

Ex: Difference of Indep Means

- Distribution free inference based on central limit theorem
  - Need to know variance of observations and relative group sizes

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

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

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

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

56

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

57

Ex: Linear Regression

- Distribution free inference based on a central limit theorem
  - Need to know variance of observations within groups (MSE) and variance of predictor *X*

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

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

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

58

Controlling Variation

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

59

Ex: Linear Regression w/ Adjustment

- Distribution free inference based on a central limit theorem
  - Need to know variance of observations within groups (MSE), variance of predictor *X,* and correlation of *X* with any additional covariates

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

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

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

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

60

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)

61

Ex: Difference of Indep Proportions

- Inference based on central limit theorem
  - Variance from mean-variance relationship; need to know relative group sizes

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

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

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

$$\text{Define } \sigma_i^2 = p_i(1-p_i)$$

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

62

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

63

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"

	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	

64

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

66

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
- 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

68

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

69

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

70

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”

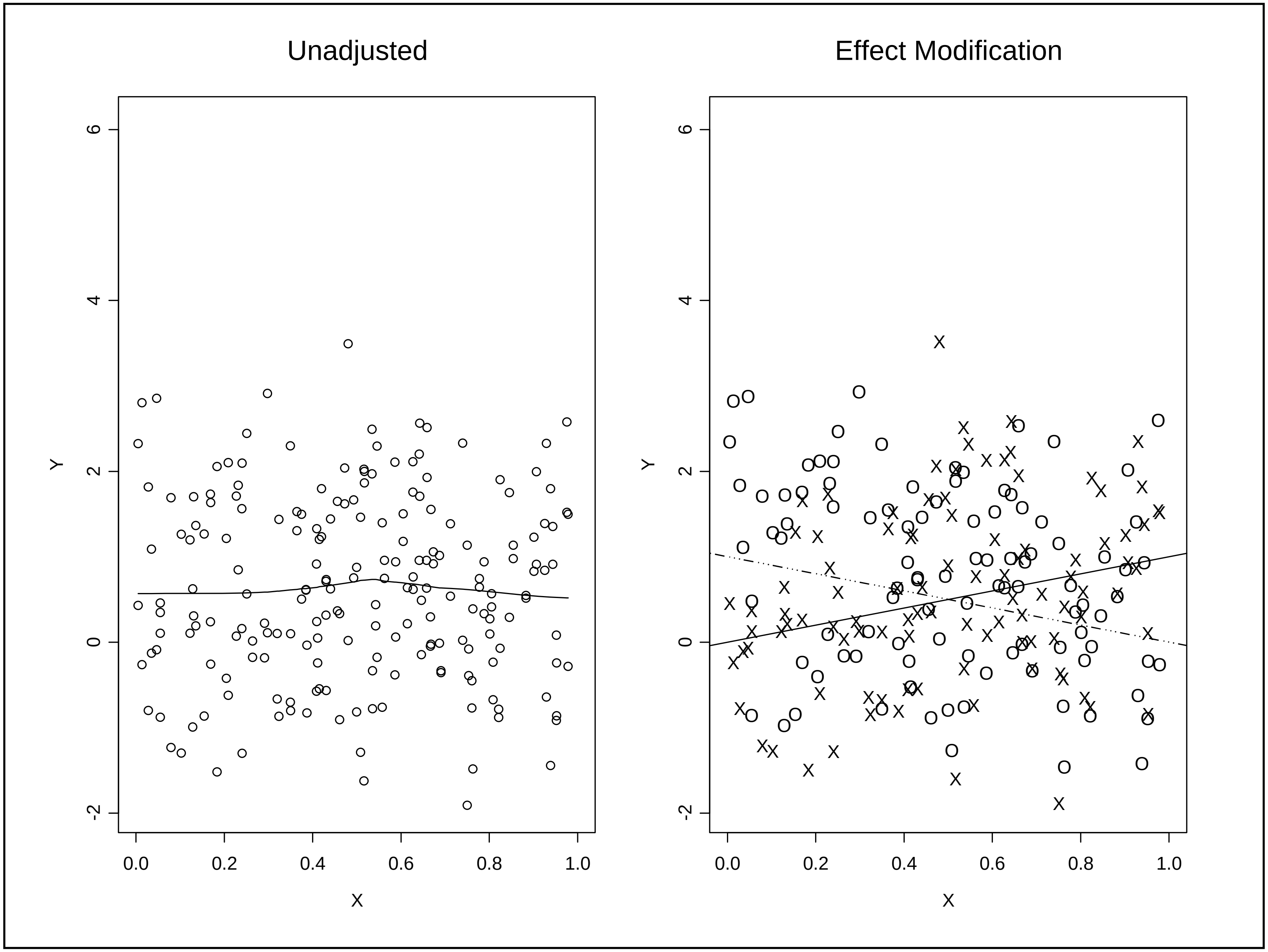
71

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)

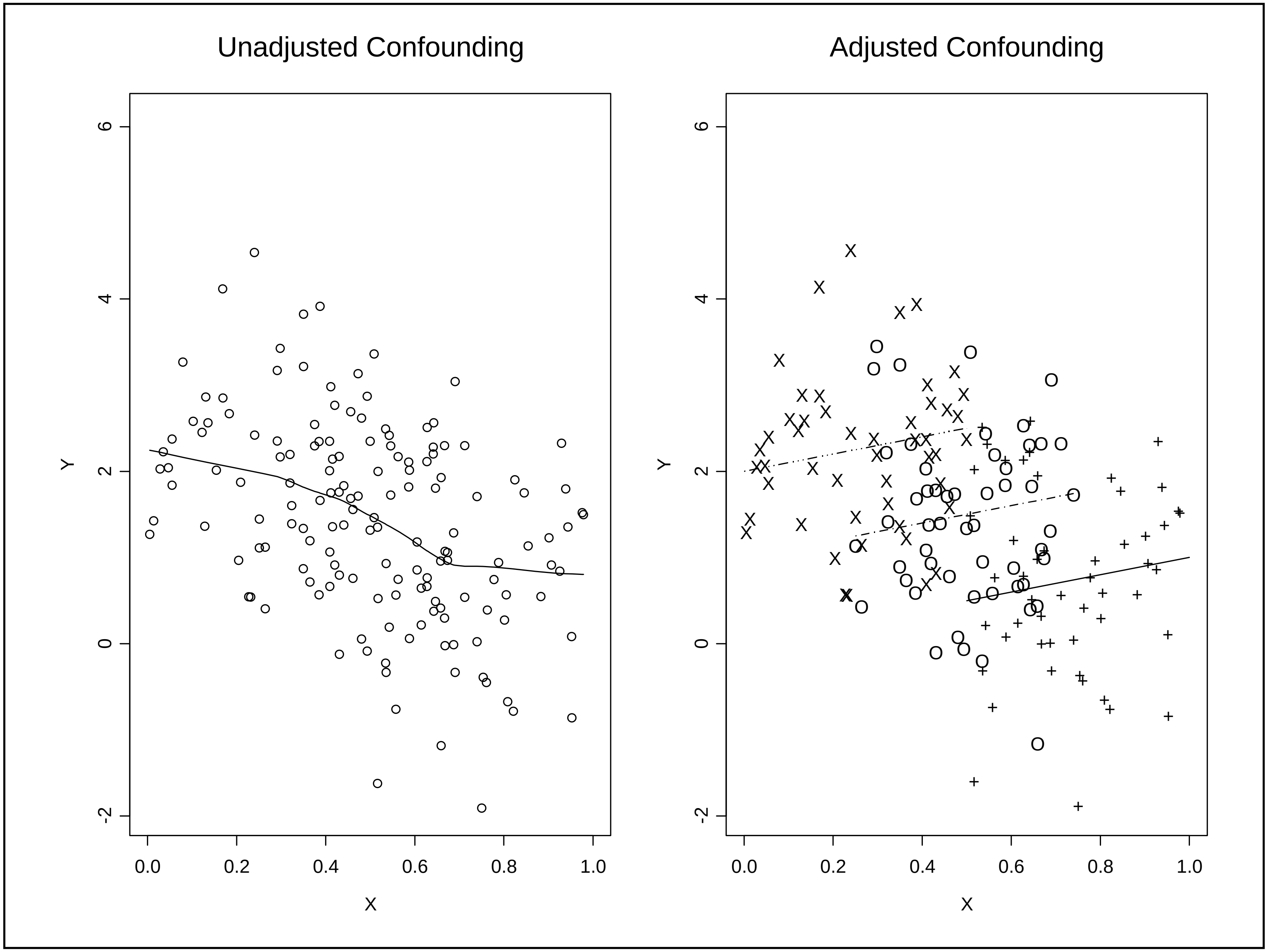

Unadjusted
Effect Modification
Y
X
0.0
0.2
0.4
0.6
0.8
1.0
-2
0
2
4
6

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

75

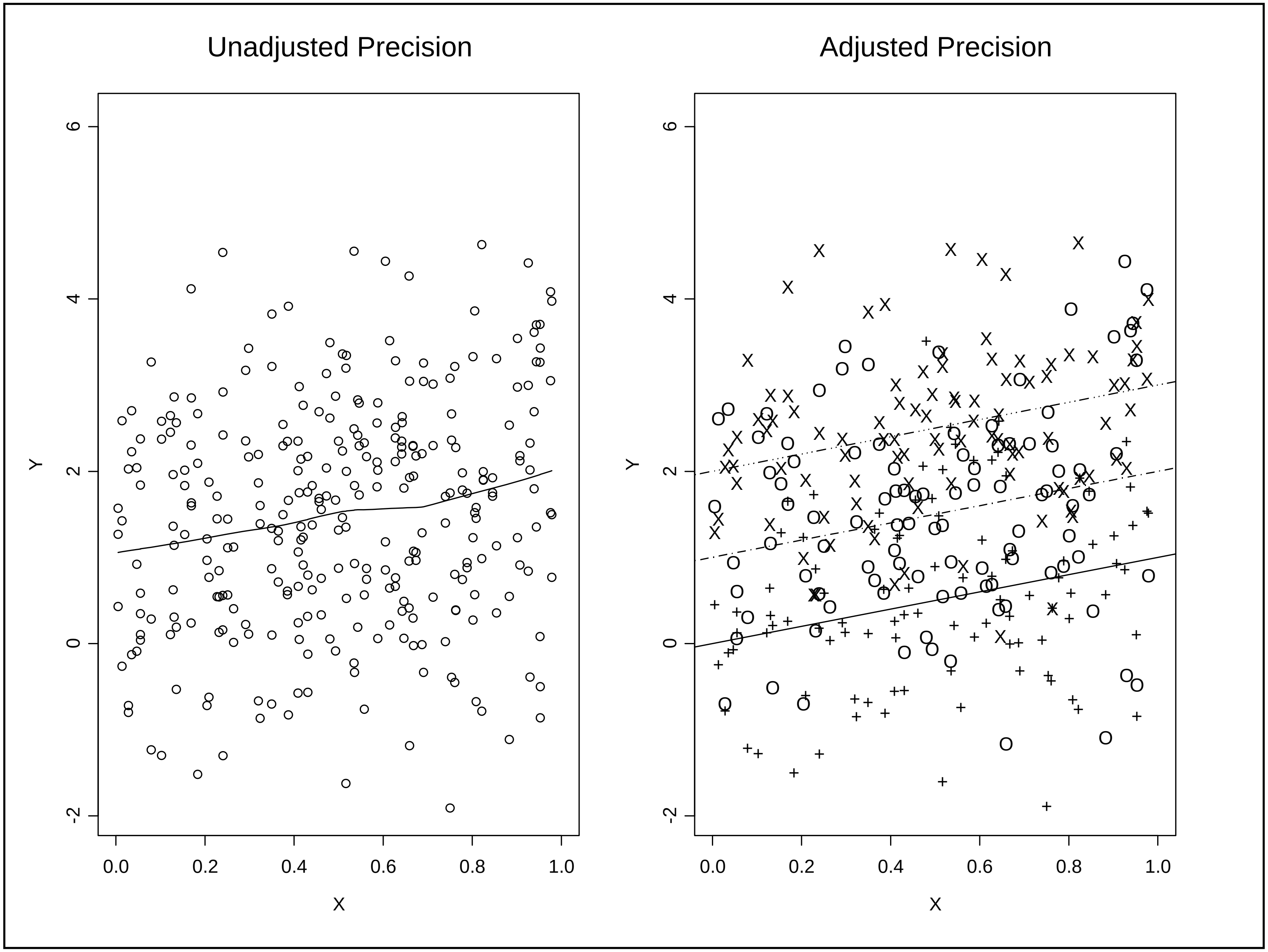

Unadjusted Confounding
Adjusted Confounding
Y
X
-2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

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

77


Unadjusted Precision
Adjusted Precision
Y
X
6
4
2
0
-2
0.0
0.2
0.4
0.6
0.8
1.0

Adjusting for Covariates

Borrowing Information Across Strata

Two Sample Comparisons

- When we have only two populations of interest, we test for associations by comparing summary measures
  - Difference of means
  - Ratio of means
  - Ratio of geometric means
  - Difference of proportions
  - Ratio of proportions
  - Odds ratio
  - Hazard ratio
  - (and we have others)

Borrowing Information Across *X*

- When we have multiple groups, we often want to summarize associations by making comparisons between groups that differ by a single unit
  - We will have multiple ways we can make such a comparison
  - We will generally not have much data for each individual comparison

- One solution is to take the average of all such comparisons
  - We can compare the means between all adjacent age groups
  - Then we can average all of those comparisons

81

Example: Association Between FEV, Age

Mean FEV in age group a : μ_a

Differences across adjacent ages :

$$\Delta_{4,3} = \mu_4 - \mu_3 \qquad \hat{\Delta}_{4,3} = \bar{Y}_4 - \bar{Y}_3$$

$$\Delta_{5,4} = \mu_5 - \mu_4 \qquad \hat{\Delta}_{5,4} = \bar{Y}_5 - \bar{Y}_4$$

$$\vdots \qquad \vdots$$

$$\Delta_{19,18} = \mu_{19} - \mu_{18} \qquad \hat{\Delta}_{19,18} = \bar{Y}_{19} - \bar{Y}_{18}$$

Use weighted average for overall effect

$$\theta = \frac{\sum_{a=4}^{19} w_a \Delta_{a,a-1}}{\sum_{a=4}^{19} w_a} \qquad \hat{\theta} = \frac{\sum_{a=4}^{19} w_a \hat{\Delta}_{a,a-1}}{\sum_{a=4}^{19} w_a}$$

82

Choice of Weights

- We can show that if there is a constant association in each comparison and homoscedasticity, weighting by the "statistical information" is most often best
  - "statistical information" is the inverse of the variance

Differences across adjacent ages:

$$\hat{\Delta}_{a,a-1} = \overline{Y}_a - \overline{Y}_{a-1} \mathrel{\dot\sim} N\left(\beta, \sigma^2\left(\frac{1}{n_a} + \frac{1}{n_{a-1}}\right)\right)$$

Weighted average for overall effect based on harmonic means of sample sizes (σ^2 cancels out)

$$\hat{\beta} = \frac{\sum_{a=4}^{19} w_a \hat{\Delta}_{a,a-1}}{\sum_{a=4}^{19} w_a} \qquad w_a = \frac{n_a n_{a-1}}{n_a + n_{a-1}}$$

83

Borrowing Information: More Efficient

- The prior approach did not use all the comparisons we could have made
- If there was (something close to) a straight line association across means, we could also compare groups that had a greater difference in *X*
- But now, not all such comparisons are estimating the same number
  - We need to divide by the difference in *X* for each comparison
  - Our weighting scheme should still account for the statistical information in each comparison

84

Choice of Weights

- We can show that if there is a constant association in each comparison and homoscedasticity, weighting by the "statistical information" is most often best
  - "statistical information" is the inverse of the variance

Differences across arbitrary age groups:

$$\hat{\Delta}_{a,b} = \frac{\overline{Y}_a - \overline{Y}_b}{a-b} \,\dot{\sim}\, N\left(\beta, \frac{\sigma^2}{(a-b)^2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)\right)$$

Weighted average for overall effect based on harmonic means of sample sizes (σ^2 cancels out)

$$\hat{\beta} = \frac{\sum_{a=3}^{19} \sum_{b=3}^{19} w_{ab} \hat{\Delta}_{a,b}}{\sum_{a=3}^{19} \sum_{b=3}^{19} w_{ab}} \qquad w_{ab} = (a-b)^2 \frac{n_a n_b}{n_a + n_b}$$

85

Least Squares Regression

- We can also express this formula for each observation, rather than for group means
  - The weights do not depend on the sample sizes, because the sample sizes are 1 for each comparison

Comparison for each possible pair:

$$\hat{\Delta}_{i,j} = \frac{Y_i - Y_j}{X_i - X_j} \dot{\sim} N\left(\beta, \frac{2\sigma^2}{\left(X_i - X_j\right)^2}\right)$$

Weighted average for overall effect based on harmonic means of sample sizes ($2\sigma^2$ cancels out)

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \hat{\Delta}_{i,j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \qquad w_{ij} = \left(X_i - X_j\right)^2$$

86

Ordinary Least Squares: Usual Formula

- This formulation of ordinary least squares estimates (OLSE) can be shown to be equal to the usual formula

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 \times X_i \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

But

$$S_{XY} = \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) = \sum_{i=1}^{n}(Y_i - \overline{Y})X_i = \sum_{i=1}^{n}Y_i(X_i - \overline{X})$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i - Y_j)(X_i - X_j) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2\frac{(Y_i - Y_j)}{(X_i - X_j)}$$

$$S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}X_i(X_i - \overline{X}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2$$

87

Other Regression Models

- I can write down exact formulas for linear regression estimates
  - (Even so, I do prefer using a computer to solve them)
- For other regression models, no “closed form solution” exists
  - We have to find regression coefficient estimates in a computerized iterative search
- Just the same: The spirit of fitting a regression model remains the same
  - We are estimating slopes by combining estimates across all relevant comparisons

88

Adjustment for Covariates

- The same principles about “borrowing information” hold
- We are in essence estimating associations within strata defined by the third variable (a confounder or precision variable)
- We then average those stratum specific estimated associations
  - Weights proportional to statistical information from strata

$$E(Y_i \mid X_i, Z_i = z) = \beta_{0z} + \beta_{1z} \times X_i \qquad \hat{\beta}_{1z} = \frac{S_{XYz}}{S_{XXz}} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X}) 1_{[Z_i = z]}}{\sum_{i=1}^{n} (X_i - \overline{X})^2 1_{[Z_i = z]}}$$

Adjusted estimate as weighted average of stratum specific estimates

$$E(Y_i \mid X_i, Z_i = z) = \beta_0 + \beta_1 \times X_i + \beta_2 \times Z_i$$

$$\hat{\beta}_1 = \frac{\sum_z w_z \hat{\beta}_{1z}}{\sum_z w_z} \qquad w_z = \frac{\hat{\sigma}^2_{Y|X,z}}{s\hat{e}^2(\hat{\beta}_{1z})} \quad (\text{homoscedasticity : constant } \sigma^2_{Y|X,z})$$

89