

# Applied Regression Analysis

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

## Session 6

---

---

---

---

---

---

---

---

# Applied Regression Analysis

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

## Part 3: Adjustment for Covariates

---

---

---

---

---

---

---

---

# Lecture Outline

- .....
- Topics:
    - Multiple Regression Model
    - Reasons for Adjusting for Covariates
    - FEV Example

---

---

---

---

---

---

---

---

## Multiple Regression Model

.....

---

---

---

---

---

---

---

---

## Multiple Regression Model

.....

- We often model the mean response across groups defined by multiple predictors
  - Simple regression: 1 predictor
    - E.g., compare the distribution of FEV across age groups
  - Multiple regression: 2 or more predictors
    - E.g., compare the distribution of FEV across groups defined by age, height, and smoking status

---

---

---

---

---

---

---

---

## Interpretation of Regression Parameters

.....

- Difference in interpretation of slopes
- Unadjusted Model :  $E[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$
- $\beta_1$  = Diff in mean Y for groups differing by 1 unit in X
    - (The distribution of W might differ across groups being compared)
- Adjusted Model :  $E[Y_i | X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$
- $\gamma_1$  = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

---

---

---

---

---

---

---

---

## Relationship Between Models

- Relationship between the adjusted and unadjusted slopes
  - The slope of the unadjusted model will tend to be

$$\beta_1 = \gamma_1 + r_{XW} \frac{\sigma_W}{\sigma_X} \gamma_2$$

- Hence, adjusted and unadjusted slopes for X are estimating the same quantity only if

- $r_{XW} = 0$  (X and W are uncorrelated), OR
- $\gamma_2 = 0$  (there is no association between W and Y after adjusting for X)

---

---

---

---

---

---

---

---

## Relationship Between Models

- Relationship between the precision of the adjusted and unadjusted models

Unadjusted Model  $[\text{se}(\hat{\beta}_1)]^2 = \frac{\text{Var}(Y|X)}{n\text{Var}(X)}$

Adjusted Model  $[\text{se}(\hat{\gamma}_1)]^2 = \frac{\text{Var}(Y|X,W)}{n\text{Var}(X)(1-r_{XW}^2)}$

$$\text{Var}(Y|X) = r_2^2 \text{Var}(W|X) + \text{Var}(Y|X,W)$$

---

---

---

---

---

---

---

---

## Relationship Between Models

- Relationship between the precision of the adjusted and unadjusted models
  - An association between Y and W (after adjustment for X) tends toward increased precision of the adjusted model relative to the unadjusted model
  - Correlation between X and W tends toward decreased precision of the adjusted model relative to the unadjusted model

---

---

---

---

---

---

---

---

## Impact on Covariate Adjustment

- Our focus on why we adjust for covariates is thus on
  - The scientific interpretation of the slopes
  - The bias of the estimates relative to the scientific parameter of interest
  - The precision of the estimates of association

---

---

---

---

---

---

---

---

## Reasons for Adjusting for Covariates

.....

---

---

---

---

---

---

---

---

## Adjustment for Covariates

- In order to assess whether we adjust 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

---

---

---

---

---

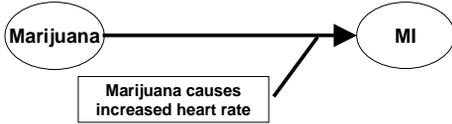
---

---

---

# Causation versus Association

- Example: Scientific interest in causal pathways between marijuana use and heart attacks (MI)
  - Pictorial representation of hypothetical causal effect of marijuana on MI that might be of scientific interest



---

---

---

---

---

---

---

---

# Causation versus Association

- Statistical analysis can only detect associations reflecting causation in either direction
  - Only experimental design and understanding of the variables allows us to infer cause and



- Statistical analysis will identify causation in either direction

---

---

---

---

---

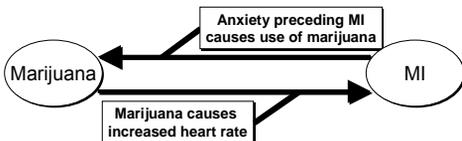
---

---

---

# Causation versus Association

- In an observational study, we cannot thus be sure which causative mechanism an association might represent
  - Either of these mechanisms will result in an association between marijuana use and MI



---

---

---

---

---

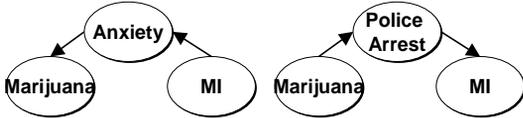
---

---

---

## Causation versus Association

- Thus, in using statistical associations to try to investigate causation, we must further consider the role other variables might play
  - A statistical association can exist between two variables due to a network of causal pathways in either direction between the two variables



Applied Regression Analysis,  
June, 2003

16

---

---

---

---

---

---

---

---

---

---

## Causation versus Association

- Furthermore, an association between two variables exists if they are each caused by a third variable
  - This is the classic case of a confounder that we would like to adjust for in order to avoid finding spurious associations when looking for cause and effect



Applied Regression Analysis,  
June, 2003

17

---

---

---

---

---

---

---

---

---

---

## Causation versus Association

- But not all such networks of causal pathways will produce an association
  - Two variables are not associated just because they each are the cause of a third variable
    - E.g., no association between marijuana use and MI if the following are the only pathways



Applied Regression Analysis,  
June, 2003

18

---

---

---

---

---

---

---

---

---

---

## Causation versus Association

- Adjustment for the third variable in this case can produce a spurious association in this example
  - Missing days off work is informative about MI incidence among those who do not use marijuana
    - Among people missing work, marijuana users will have lower incidence of MI
      - The incidence of MI will likely be similar between marijuana users and nonusers who do not miss work
  - The resulting interaction will seem to be an association in an adjusted analysis



Applied Regression Analysis,  
June, 2003

19

---

---

---

---

---

---

---

---

## Causation versus Association

- In the previous example, we might know not to adjust for Days Off Work, because that occurs after the response
  - We regard that causes of events must be in the correct temporal sequence
    - However, there are situations where this criterion can be hard to judge
    - Furthermore, there are situations where similarly inappropriate adjustment of variables can occur with variables measured before the event

Applied Regression Analysis,  
June, 2003

20

---

---

---

---

---

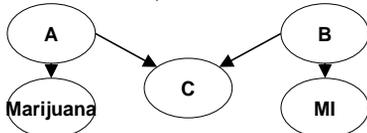
---

---

---

## Causation versus Association

- Similar problems can arise from more complicated causal pathways
  - Adjustment for Variable C would produce a spurious association
    - Note that the association between C and marijuana and C and MI are not causal, but C can occur before an MI



Applied Regression Analysis,  
June, 2003

21

---

---

---

---

---

---

---

---

## Causation versus Association

---

---

---

---

---

---

---

---

---

---

- Sometimes we can isolate particular pathways of scientific interest by including a third variable into an analysis
  - “Adjusting” for an effect of a third variable
    - Strata are defined based on the value of the third variable
    - Comparisons of the response distribution across groups defined by the predictor of interest are made within strata
    - The effects within strata are then averaged in some way to obtain the adjusted association

## Causation versus Association

---

---

---

---

---

---

---

---

---

---

- Clearly, such adjustment makes most sense only when the association between response and predictor of interest is the same in each stratum
  - If there are different effects across strata, modeling an interaction would be indicated
    - Essentially, the question should be answered in each stratum separately

## Causation versus Association

---

---

---

---

---

---

---

---

---

---

- 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

- We include predictors in a regression model 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 changes the question being answered by the statistical analysis
    - Adjustment can be used to isolate associations that are of particular interest

---

---

---

---

---

---

---

---

## 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

---

---

---

---

---

---

---

---

## 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

---

---

---

---

---

---

---

---

# Confounding

---

- 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
  - In linear regression, these symptoms are diagnostic of confounding
    - Effect modification would show differences between adjusted analysis and unadjusted analysis, but would also show different associations in the different strata

---

---

---

---

---

---

---

---

# Confounding

---

- Note that confounding produces a difference between unadjusted and adjusted analyses, but those symptoms are not proof of confounding
  - Must consider possible causal pathways
    - (recall M-shaped causal diagram)
  - Summary measures which are nonlinear functions of the mean sometimes show the above symptoms in the absence of confounding
    - (relevant to odds ratios)

---

---

---

---

---

---

---

---

# 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

---

---

---

---

---

---

---

---

# Confounding

- Some times the scientific question of greatest interest is confounded by unexpected associations in the data
  - Confounders
    - Variables (causally) predictive of outcome, but not in the causal pathway of interest
      - (Often assessed in the control group)
    - 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
  - Detecting confounders must ultimately rely on our best knowledge about possible mechanisms

---

---

---

---

---

---

---

---

# 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

- Two special cases to consider when attempting to gain precision in a model
  - If stratified randomization or matched sampling was used in order to address possible confounding and / or precision issues, the added precision will NOT be realized UNLESS the stratification or matching variables are adjusted for in the analysis
  - If baseline measurements are available, it is more precise to adjust for those variables as a covariate than to analyze the change

---

---

---

---

---

---

---

---

## 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
  - We illustrate the difference between precision variables, confounders, and effect modifiers in the following hypothetical example

---

---

---

---

---

---

---

---

## 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.

---

---

---

---

---

---

---

---

## 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 receiving 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	
<b>Mean</b>	20.5	17.4	3.1
<b>SD</b>	17.7	17.6	

Applied Regression Analysis,  
June, 2003

37

---

---

---

---

---

---

---

---

---

---

## 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
    - $\text{Var}(\text{Size} | \text{Trt}) = 311.5$  (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)

Applied Regression Analysis,  
June, 2003

38

---

---

---

---

---

---

---

---

---

---

## Example: Adjusted Analysis (Case 1)

Fruit sizes by treatment group and type of fruit

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

Applied Regression Analysis,  
June, 2003

39

---

---

---

---

---

---

---

---

---

---

## 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")
    - $\text{Var}(\text{Size} \mid \text{Trt}, \text{Fruit}) = 0.25$  ( $\text{SD} = 0.50$ )
  - The average difference of 3.1 across types of fruit is large compared to the within group standard deviation of 0.50
    - ( $P \text{ value} < .0001$ )
  - (Randomization did protect us from confounding: Each treatment group had four plants of each kind)

Applied Regression Analysis,  
June, 2003

40

---

---

---

---

---

---

---

---

---

---

## 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

Applied Regression Analysis,  
June, 2003

41

---

---

---

---

---

---

---

---

---

---

## 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	

Applied Regression Analysis,  
June, 2003

42

---

---

---

---

---

---

---

---

---

---

## 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

---

---

---

---

---

---

---

---

---

---

## 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

---

---

---

---

---

---

---

---

---

---

## 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)

---

---

---

---

---

---

---

---

---

---

## 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	
<b>Mean</b>	20.5	18.7	1.8
<b>SD</b>	17.7	19.6	

---

---

---

---

---

---

---

---

---

---

## 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)

---

---

---

---

---

---

---

---

---

---

## Example: Adjusted Analysis (Case 3)

Fruit sizes by treatment group and type of fruit

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

Applied Regression Analysis,  
June, 2003

49

---

---

---

---

---

---

---

---

---

---

## 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 ( $P=0.0009$ )
    - 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)

Applied Regression Analysis,  
June, 2003

50

---

---

---

---

---

---

---

---

---

---