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 : $E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$

  - β_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 : $E\left[Y_i \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

  - γ_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 $\left[se\left(\hat{\beta}_1\right)\right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right)\right]^2 = \frac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1 - r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

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 undadjusted 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

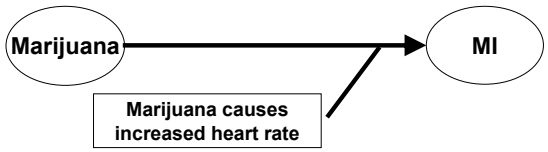
Reasons for Adjusting for Covariates

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

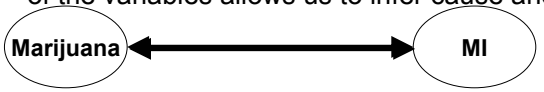
Causation versus Association

- 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

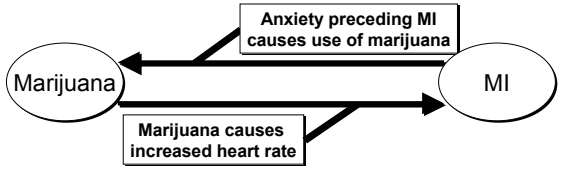
- Statistical analysis can only detect associations reflecting causation in either direction
  - Only experimental design and understanding of the variables allows us to infer cause and



  - Statistical analysis will identify causation in either direction

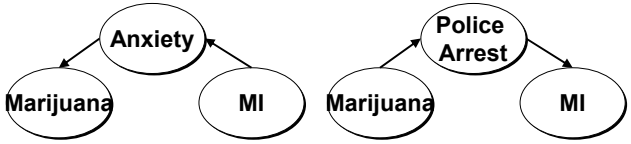
Causation versus Association

- 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



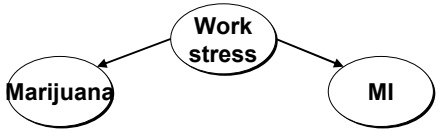
Causation versus Association

- 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



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



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

Causation versus Association

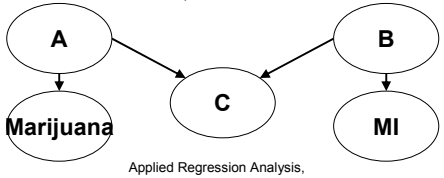
- Adjustment for the third variable in this case can produce a spurious association in this example
  - Missing days off work is informative about MI incidence among those who do not use marijuana
    - Among people missing work, marijuana users will have lower incidence of MI
      - The incidence of 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

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

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



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	
Mean	20.5	17.4	3.1
SD	17.7	17.6	

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)

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

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

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	
Mean	20.5	18.7	1.8
SD	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
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

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)

Adjusting for Covariates: Confounding, Precision, Effect Modification

Confounding, Precision, Effect Modification

- Discriminating between confounding, precision, and effect modifying variables
  - Is the estimate of association between response and the predictor of interest the same in all strata?
    - Effect modifier: NO; Confounder, precision: YES
  - Is the third variable causally associated with the response after adjusting for the predictor of interest?
    - Confounder, precision: YES
  - Is the third variable associated with the predictor of interest?
    - Confounder: YES; Precision: NO

Interpretation of Regression Parameters

- Difference in interpretation of slopes

Unadjusted Model : $E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 \times X_i$

  - β_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 : $E\left[Y_i \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

  - γ_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 $\left[se\left(\hat{\beta}_1\right)\right]^2 = \frac{Var\left(Y \mid X\right)}{nVar\left(X\right)}$

Adjusted Model $\left[se\left(\hat{\gamma}_1\right)\right]^2 = \frac{Var\left(Y \mid X, W\right)}{nVar\left(X\right)\left(1 - r_{XW}^2\right)}$

$$Var\left(Y \mid X\right) = \gamma_2^2 Var\left(W \mid X\right) + Var\left(Y \mid X, W\right)$$

Example: Unadjusted Analysis (Case 1: A Precision Variable)

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**	

Example: Adjusted Analysis (Case 1: A Precision Variable)

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	41.6, 40.5, 39.9, 40.7	
Mean(SD)	43.8 (0.29)	40.7 (0.70)	3.1

Example: Unadjusted Analysis (Case 2: A Confounder)

Fruit sizes by treatment group

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

Example: Adjusted Analysis (Case 2: A Confounder)

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: Unadjusted Analysis (Case 3: An Effect Modifier)

Fruit sizes by treatment group

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

Example: Adjusted Analysis (Case 3: An Effect Modifier)

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

FEV Example

Scientific Question

- Association between smoking and lung function in children
  - Longterm smoking is associated with lower lung function
  - Are similar effects observed in short term smoking in children?

Causal Pathway of Interest

- We are interested in whether smoking will cause a decrease in lung function as measured by FEV

Causation versus Association

- Statistical analyses, however, can only detect associations between smoking and FEV

  - In a randomized trial, we could infer from the design that any association must be causal
  - In an observational study, we must try to isolate causal pathways of interest by adjusting for covariates

Study Design

- Observational study
  - Measurements on 654 healthy children
    - Predictor of interest: Self-reported smoking
    - Response: FEV
    - Additional covariates
      - Effect modifiers
      - Potential confounders
      - Precision variables

Additional Covariates: Effect Modifiers

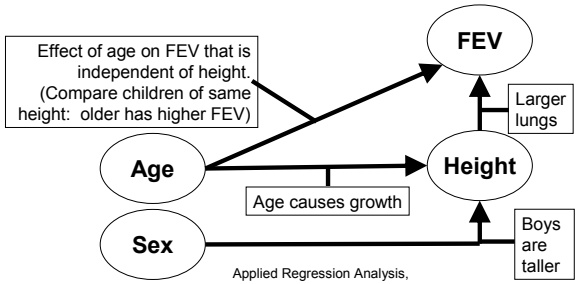
- There are no covariates currently of scientific interest for their potential for effect modification
  - First things first
    - Not generally advisable to go looking for different effects of smoking in subgroups before we have established that an effect exists overall
      - (We may sometimes delay discovery of important facts, but most times this seems the logical strategy)

Additional Covariates: Confounders

- Think about potential confounders
  - Necessary requirements for confounders
    - Associated causally with response
    - Associated with predictor of interest in sample
  - Prior to looking at data, we cannot be sure of the second criterion
    - But, clearly, any strong predictor of the response has the potential to be a confounder
      - So first consider known predictors of response
    - Furthermore, in an observational study, known associations in the population will likely also be in the sample

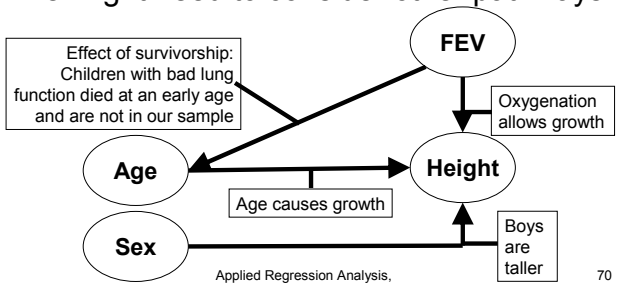
Predictors of FEV

- "Known" predictors of FEV



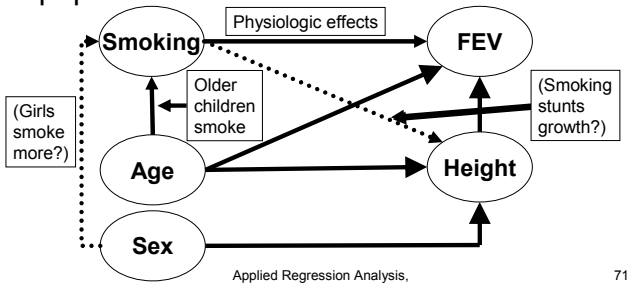
An Aside: What is "Known"?

- In an observational, cross-sectional study, we might need to consider other pathways



Associations with Smoking

- "Known" associations with smoking in the population

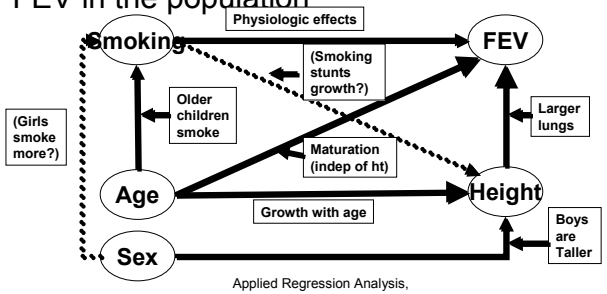


Adjusting for Potential Confounders

- Investigating the effect of smoking on FEV in children
  - We are scientifically interested in the possibility that smoking might cause decreased FEV
  - We are not scientifically interested in showing that FEV status might influence smoking behavior
    - (Of course, this is one possible explanation of an observed association, and so we must try to rule this out)

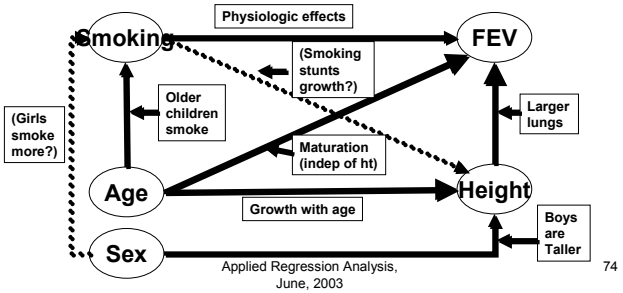
Associations with Smoking, FEV

- "Known" associations with smoking and FEV in the population



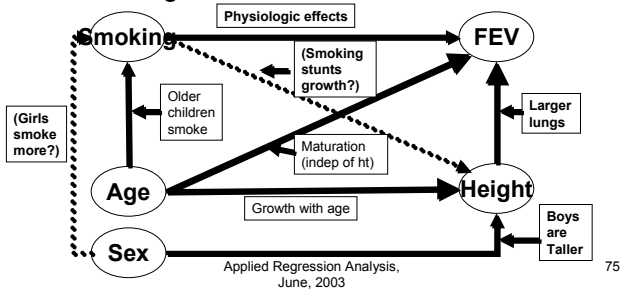
Pathways Tested in Unadjusted Analysis

- Comparing nonsmokers to smokers in observational study



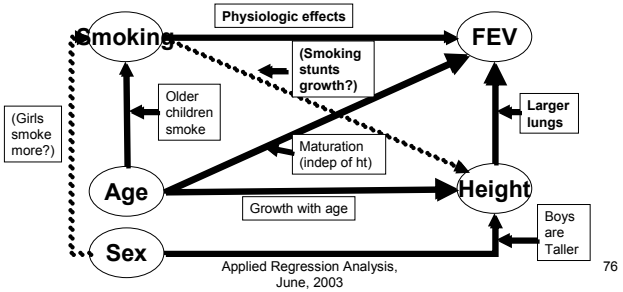
Pathways Tested Adjusting for Age

- Comparing nonsmokers to smokers of same age in observational study removes major confounding



Pathways Tested Adjusting for Age, Sex

- Comparing nonsmokers to smokers of same age and sex removes all confounding



Additional Covariates: Precision

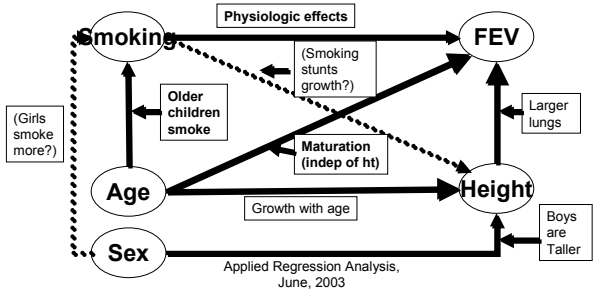
- Think about major predictors of response
  - In an observational study, all predictors of response should be considered potential confounders
  - However, even if strong predictors of response are not confounding (i.e., not associated with POI in sample), we might want to consider adjusting the analysis to gain precision

Additional Covariates: Precision

- In the FEV study, height is probably the strongest predictor of the response
  - The amount of air exhaled in 1 second (FEV) involves
    - Lung size (may not be of as much interest)
    - Lung function (probably more affected by smoking)
  - Height is a reasonable surrogate for lung size
    - Adjusting for height may allow comparisons that are more directly related to lung function

Pathways Tested Adjusting for Height

- Comparing nonsmokers to smokers of same height gains precision, but still has confounding

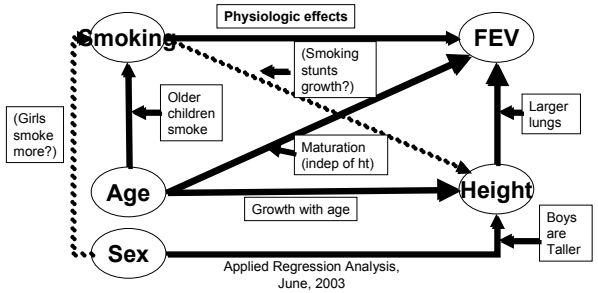


Additional Covariates: Precision

- After adjusting for age, however, height is primarily a precision variable
  - After adjusting for age, there may be some residual confounding through any tendency for one sex to smoke more
    - (In our data, we have approximately equal numbers of boys and girls who smoke, so such confounding may not be such an issue)

Pathways Tested Adjusting for Age, Height

- Comparing nonsmokers to smokers of same age and height removes confounding and gains precision

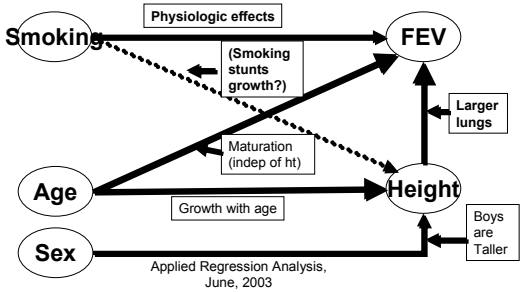


Additional Covariates: Precision

- If we adjust for height, we do lose one of the ways that smoking might have affected FEV
  - We can consider a hypothetical randomized clinical trial (RCT) of smoking (don't try this at home)
    - Consider randomizing 10 year olds to smoke or not
      - Stratify on height at 10 years old to gain precision
    - At the end of 5 years, we might anticipate lower FEV in the smokers due to
      - Shorter smokers (if smoking stunts growth)
      - Lower FEV when comparing children of same height
    - Statistical analyses could adjust for baseline height to gain precision
      - Secondary analyses might adjust for final height to tease out mechanisms

Causal Pathways of Interest in RCT

- RCT would test all causal pathways, and might have precision if we match heights at baseline



Planned Analyses: Covariate Adjustment

- Based on these issues, a priori we might plan an analysis adjusting for age and height (and sex?)
  - If that had not been specified a priori, I would perform the unadjusted analysis and then report the observed confounding from exploratory analyses
    - Data driven analyses always provide less confidence than prespecified analyses
  - In order to illustrate the effects of adjusting for confounders and precision variables, I will explore several analyses
    - Variable smoker coded 0= nonsmokers, 1= smokers

Planned Analyses: Summary Measure

- Based on the scientific relationship between FEV and its strongest predictor (height), we will compare geometric means rather than means
  - Geometric means will likely be estimated with greater precision, because the standard deviation of FEV measurements is likely proportional to the mean
  - Such an analysis is easily performed and interpreted
    - Linear regression on log FEV
    - Interpret exponentiated regression parameters as multiplicative effects

Planned Analyses: Sampled Ages

- We will restrict our analyses to children 9 and older
  - The dataset included children as young as 3!
    - The youngest smoker was 9
  - Dilemma
    - Younger children may help predict “normal” FEV, if our modeling of age and height is correct
    - If we are wrong, then we may not remove all confounding
  - Reasoning behind decision
    - We only have 65 smokers, so that is the limiting factor in precision of our analysis
      - Having young nonsmokers does not add much

Unadjusted Analysis: Stata Output

```
. regress logfev smoker if age>=9, robust

Number of obs =     439
Root MSE      =  .24765
              |         Robust
logfev |   Coef. St Err    t    P>|t|   [95% CI]
smoker |   .102  .0317   3.23  0.001   .040   .165
 _cons |  1.058  .0129  81.82  0.000  1.033  1.084
```

Unadjusted Analysis: Interpretation.....

- Smoking effect
  - Geometric mean of FEV is 10.8% higher in smokers than in nonsmokers (95% CI: 4.1% to 17.9% higher)
    - These results are atypical of what we might expect with no true difference between groups: P = 0.001
    - (Calculations: $e^{0.102}$= 1.108; $e^{0.040}$= 1.041; $e^{0.165}$= 1.179)
      - (Note that exp (x) is approx 1+x for x close to 0)
  - (Because smoker is a binary (0-1) variable, this analysis is nearly identical to a two sample t test allowing for unequal variances)

Unadjusted Analysis: Interpretation.....

- Intercept
  - –Geometric mean of FEV in nonsmokers is 2.88 l/sec
    - The scientific relevance is questionable here, because we do not really know the population our sample represents
      - Comparing smokers to nonsmokers is more useful than looking at either group by itself
    - (Calculations: $e^{1.058}$= 2.881)
    - (The P value is of no importance whatsoever, it is testing that the log geometric mean is 0 or that the geometric mean is 1. Why would we care?)
  - –(Because *smoker* is a binary variable, the estimate corresponds to the sample geometric mean)

Age Adjusted Analysis: Stata Output.....

```
. regress logfev smoker age if age>=9, robust

Number of obs =     439
Root MSE      =  .20949
             |         Robust
logfev |   Coef. St Err    t    P>|t|   [95% CI]
smoker |  -.051  .0344  -1.49  0.136  -.119   .016
   age |   .064  .0051  12.37  0.000   .053   .074
 _cons |  0.352  .0575   6.12  0.000   .239   .465
```


Age Adjusted Analysis: Interpretation.....

- Smoking effect
  - Geometric mean of FEV is 5.0% lower in smokers than in nonsmokers of the same age (95% CI: 12.2% lower to 1.6% higher)
    - These results are not atypical of what we might expect with no true difference between groups of the same age: P = 0.136
      - Lack of statistical significance is also evident because the confidence interval contains 1 (as a ratio) or 0 (as a percent difference)
    - (Calculations: $e^{-0.051}$= 0.950; $e^{-0.119}$= 0.888; $e^{0.016}$= 1.016)
      - (Note that exp (x) is approx 1+x for x close to 0)

Age Adjusted Analysis: Interpretation.....

- Age effect
  - –Geometric mean of FEV is 6.6% higher for each year difference in age between two groups with similar smoking status(95% CI: 5.5% to 7.6% higher for each year difference in age)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between age groups having similar smoking status: P < 0.0005

Age Adjusted Analysis: Interpretation.....

- Intercept
  - –Geometric mean of FEV in newborn nonsmokers is 1.42 l/sec
    - Intercept corresponds to the log geometric mean in a group having all predictors equal to 0
    - There is no scientific relevance is here, because we are extrapolating outside our data
    - (Calculations: $e^{0.352} = 1.422$)

Age Adjusted Analysis: Comments.....

- Comparing unadjusted and age adjusted analyses
  - – Marked difference in effect of smoking suggests that there was indeed confounding
    - Age is a relatively strong predictor of FEV
    - Age is associated with smoking in the sample
      - – Mean (SD) of age in analyzed smokers: 11.1 (2.04)
      - – Mean (SD) of age in analyzed nonsmokers: 13.5 (2.34)
  - – Effect of age adjustment on precision
    - Lower Root MSE (.209 vs .248) would tend to increase precision of estimate of smoking effect
    - Association between smoking and age tends to lower precision
    - Net effect: Less precision (SE 0.034 vs 0.031)

Age, Height Adjusted Analysis: Stata Output.....

```
. regress logfev smoker age loght if age>=9, robust

Number of obs =     439
Root MSE      =  .14407
             |               Robust
      logfev |      Coef.   St Err      t    P>|t|     [95% CI]
      smoker |    -.054    .0241    -2.22   0.027    -.101    -.006
         age |     .022    .0035     6.18   0.000     .015     .028
       loght |   2.870    .1280    22.42   0.000    2.618    3.121
       _cons |  -11.095    .5153   -21.53   0.000  -12.107  -10.082
```

Age, Height Adjusted Analysis: Interpretation

- Smoking effect
  - Geometric mean of FEV is 5.2% lower in smokers than in nonsmokers of the same age and height (95% CI: 9.6% to 0.6% lower)
    - These results are atypical of what we might expect with no true difference between groups of the same age and height: P = 0.027
    - (Calculations: $e^{-0.054}$= .948; $e^{-0.101}$= .904; $e^{-0.006}$= .994)
  - Note the wording "same age and height" even though I adjusted using a log transformation of height.
    - Equal log heights lead to equal heights

Age, Height Adjusted Analysis: Interpretation

- Age effect
  - Geometric mean of FEV is 2.2% higher for each year difference in age between two groups with similar height and smoking status (95% CI: 1.5% to 2.9% higher for each year difference in age)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between age groups having similar height and smoking status: P < 0.0005
  - Note that there is clear evidence that height confounded the age effect estimated in the analysis which modeled only smoking and age
    - But there is a clear independent effect of age on FEV

Age, Height Adjusted Analysis: Interpretation

- Height effect
  - Geometric mean of FEV is 31.5% higher for each 10% difference in height between two groups with similar ages and smoking status (95% CI: 28.3% to 34.6% higher for each 10% difference in height)
    - These results are highly atypical of what we might expect with no true difference in the geometric mean FEV between height groups having similar age and smoking status: P < 0.0005
    - (Calculations: $1.1^{2.867}$= 1.315)
  - Note that the regression coefficient of 2.870 (95% CI 2.618 to 3.121) is consistent with the scientifically derived value of 3.0

Age, Height Adjusted Analysis: Interpretation

- Intercept
  - Geometric mean of FEV in newborn nonsmokers who are 1 inch high is 0.000015 l/sec
    - Intercept corresponds to the log geometric mean in a group having all predictors equal to 0
      - Nonsmokers
      - Age 0 (newborn)
      - Log height 0 (height 1 inch)
    - There is no scientific relevance is here, because there are no such people in our sample OR the population

Age, Height Adjusted Analysis: Comments

- Comparing age and age-height adjusted analyses
  - No difference in effect of smoking suggests there was no more confounding after age adjustment
  - Effect of height adjustment on precision
    - Lower Root MSE (.144 vs .209) would tend to increase precision of estimate of smoking effect
    - Little association between smoking and height after adjustment for age will not tend to lower precision
    - Net effect: Higher precision (SE 0.024 vs 0.034)

Height Adjusted Analysis: Stata Output

```
. regress logfev smoker loght if age>=9, robust

Number of obs =     439
Root MSE      =  .14907
             |         Robust
logfev |   Coef. St Err    t    P>|t|    [95% CI]
smoker |  -.015  .0231  -0.64  0.522   -.060    .031
 loght |  3.236  .1199  27.00  0.000   3.000   3.471
 _cons |-12.375  .4968 -24.91  0.000 -13.352 -11.399
```


Height Adjusted Analysis: Comments

- Comparing height and age-height adjusted analyses
  - Marked difference in effect of smoking suggests there was still confounding by age after height adjustment
  - Effect of age adjustment on precision
    - Only slightly lower Root MSE (.144 vs .149) suggests that age adds less precision to the model than height

Age, Height, Sex Adjusted: Stata Output

```
. regress logfev smoker age loght maleif age>=9,
        robust

Number of obs =     439
Root MSE      =  .14407
             |         Robust
logfev |   Coef. St Err    t    P>|t|    [95% CI]
smoker |  -.051  .0244  -2.08  0.038   -.099   -.003
   age |   .022  .0035   6.35  0.000    .015    .029
 loght |  2.818  .1399  20.14  0.000   2.543   3.093
  male |   .015  .0151   0.99  0.323   -.015    .045
 _cons |-10.895  .5609 -19.43  0.000 -11.997  -9.793
```

Age, Height, Sex Adjusted: Comments

- Comparing age-height-sex and age-height adjusted analyses
  - No suggestion of further confounding by sex
  - Effect of sex adjustment on precision
    - Root MSE (.144 vs .144) suggests that sex adds virtually no precision to the model

Final Comments

- Choosing the model for analysis
  - Confirmatory vs Exploratory analyses
    - Every statistical model answers a different question
    - Data driven choice of analyses requires later confirmatory analyses
    - Best strategy
      - Choose appropriate primary analysis based on scientific question identified a priori
        - » Provide most robust statistical inference regarding this question
      - Further explore your data to generate new hypotheses and speculate on mechanisms
        - » Regard these statistics as descriptive

Final Disclaimer

- In presenting 5 different analyses of the FEV data, I did not mean to suggest that I would choose from among these
  - Instead, I wanted to show how regression could be used to address confounding and provide greater precision
  - I would have chosen the analysis based on age and height adjustment a priori, and reported those results as my primary analysis