

Applied Regression Analysis

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

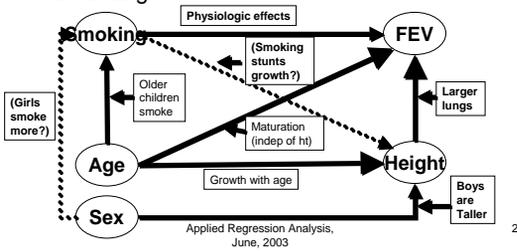
Session 8

Applied Regression Analysis,
June, 2003

1

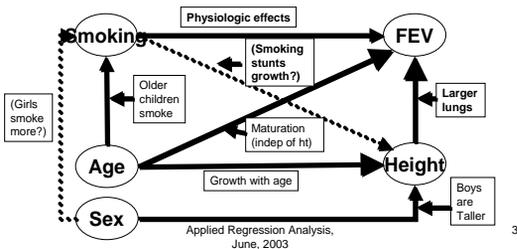
Pathways Tested Adjusting for Age.....

- 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

Applied Regression Analysis,
June, 2003

4

Additional Covariates: Precision

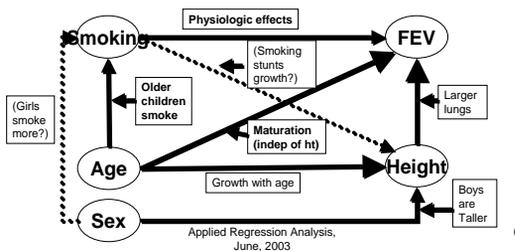
- In the FEV study, height is probably the strongest predictor of the response
 - The amount of air exhaled in 1 second (FEV) involves
 - Lung size (may not be of as much interest)
 - Lung function (probably more affected by smoking)
 - Height is a reasonable surrogate for lung size
 - Adjusting for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

Applied Regression Analysis,
June, 2003

5

Pathways Tested Adjusting for Height

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



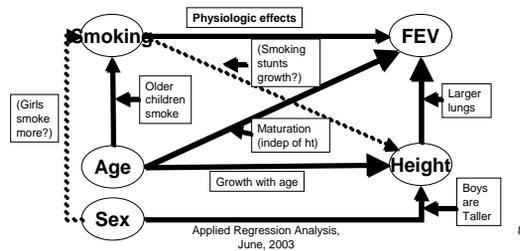
6

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

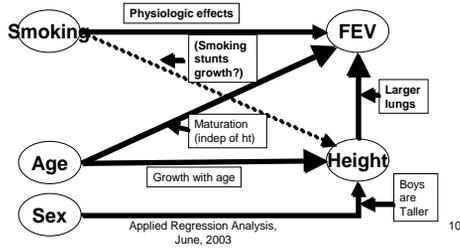


Additional Covariates: Precision

- 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

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)

Applied Regression Analysis,
June, 2003

16

Age Adjusted Analysis: Stata Output.....

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

Number of obs =      439
Root MSE      =    .20949

logfev |          Robust
-----+-----
smoker |   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
```

Applied Regression Analysis,
June, 2003

17

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)

Applied Regression Analysis,
June, 2003

18

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: $P < 0.0005$

Applied Regression Analysis,
June, 2003

19

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: $e^{0.352} = 1.422$)

Applied Regression Analysis,
June, 2003

20

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)

Applied Regression Analysis,
June, 2003

21

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
    
```

Applied Regression Analysis,
June, 2003

22

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

Applied Regression Analysis,
June, 2003

23

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

Applied Regression Analysis,
June, 2003

24

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: $1.12^{.867} = 1.315$)
 - 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

Applied Regression Analysis,
June, 2003

25

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

Applied Regression Analysis,
June, 2003

26

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)

Applied Regression Analysis,
June, 2003

27

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
    
```

Applied Regression Analysis,
June, 2003

28

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

Applied Regression Analysis,
June, 2003

29

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
    
```

Applied Regression Analysis,
June, 2003

30

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