

Biost 518

Applied Biostatistics II

.....

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

Lecture 7: Multiple Regression: Interpreting Adjusted Analyses

January 22, 2007

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

Lecture Outline

.....

- Adjustment for confounders / precision
- Linear Regression Example
 - FEV and smoking in children

2

Adjustment for Confounders, Precision Variables

.....

3

Adjustment for Covariates

.....

- We “adjust” for other covariates
 - Define groups according to
 - Predictor of interest, and
 - Other covariates
 - Compare the distribution of response across groups which
 - differ with respect to the Predictor of Interest, but
 - are the same with respect to the other covariates
 - “holding other variables constant”

4

Unadjusted vs Adjusted Models

- Adjustment for covariates changes the scientific question
 - Unadjusted models
 - Slope compares parameters across groups differing by 1 unit in the modeled predictor
 - Groups may also differ with respect to other variables
 - Adjusted models
 - Slope compares parameters across groups differing by 1 unit in the modeled predictor but similar with respect to other modeled covariates

5

Interpretation of Slopes

- Difference in interpretation of slopes

Unadjusted Model : $g[\theta | X_i] = \beta_0 + \beta_1 \times X_i$

- β_1 = Compares θ for groups differing by 1 unit in X
 - (The distribution of W might differ across groups being compared)

Adjusted Model : $g[\theta | X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

- γ_1 = Compares θ for groups differing by 1 unit in X, but agreeing in their values of W

6

Comparing models

Unadjusted $g[\theta | X_i, W_i] = \beta_0 + \beta_1 \times X_i$

Adjusted $g[\theta | X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

Science: When is $\gamma_1 = \beta_1$?

 When is $\hat{\gamma}_1 = \hat{\beta}_1$?

Statistics: When is $se(\hat{\gamma}_1) = se(\hat{\beta}_1)$?

 When is $s\hat{e}(\hat{\gamma}_1) = s\hat{e}(\hat{\beta}_1)$?

7

General Results

- These questions can not be answered precisely in the general case
 - However, in linear regression we can derive exact results
 - These will serve as a basis for later examination of
 - Logistic regression
 - Poisson regression
 - Proportional hazards regression

8

Linear Regression

- Difference in interpretation of slopes

Unadjusted Model : $E[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$

- 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 : $E[Y_i | X_i, W_i] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$

- γ_1 = Diff in mean Y for groups differing by 1 unit in X, but agreeing in their values of W

9

Relationships: True Slopes

- The slope of the unadjusted model will tend to be

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

- Hence, true adjusted and unadjusted slopes for X are estimating the same quantity only if
 - $\rho_{XW} = 0$ (X and W are truly uncorrelated), OR
 - $\gamma_2 = 0$ (no association between W and Y after adjusting for X)

10

Relationships: Estimated Slopes

- The estimated slope of the unadjusted model will be

$$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X (r_{YX} - r_{YW} r_{XW})} \right] \right)$$

- Hence, estimated adjusted and unadjusted slopes for X are equal only if
 - $r_{XW} = 0$ (X and W are uncorrelated in the sample, which can be arranged by experimental design), OR
 - $\hat{\gamma}_2 = 0$ (which cannot be predetermined, because Y is random)

11

Relationships: True SE

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

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

$$Var(Y|X) = \gamma_2^2 Var(W|X) + Var(Y|X,W)$$

$$\sigma_{Y|X}^2 = \gamma_2^2 \sigma_{W|X}^2 + \sigma_{Y|X,W}^2$$

12

Relationships: True SE

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

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

$$Var(Y|X) = \gamma_2^2 Var(W|X) + Var(Y|X,W)$$

Thus, $se(\hat{\beta}_1) = se(\hat{\gamma}_1)$ if $r_{XW} = 0$

AND

$$\gamma_2 = 0 \quad \text{OR} \quad Var(W|X) = 0$$

13

Relationships: Estimated SE

Unadjusted Model $[s\hat{e}(\hat{\beta}_1)]^2 = \frac{SSE(Y|X)/(n-2)}{(n-1)s_X^2}$

Adjusted Model $[s\hat{e}(\hat{\gamma}_1)]^2 = \frac{SSE(Y|X,W)/(n-3)}{(n-1)s_X^2(1-r_{XW}^2)}$

$$SSE(Y|X) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i)^2$$

$$SSE(Y|X,W) = \sum (Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i)^2$$

14

Relationships: Estimated SE

Unadjusted Model $[s\hat{e}(\hat{\beta}_1)]^2 = \frac{SSE(Y|X)/(n-2)}{(n-1)s_X^2}$

Adjusted Model $[s\hat{e}(\hat{\gamma}_1)]^2 = \frac{SSE(Y|X,W)/(n-3)}{(n-1)s_X^2(1-r_{XW}^2)}$

Thus, $s\hat{e}(\hat{\beta}_1) = s\hat{e}(\hat{\gamma}_1)$ if $r_{XW} = 0$

AND

$$SSE(Y|X)/(n-2) = SSE(Y|X,W)/(n-3)$$

15

Residual Squared Error

$$SSE(Y|X) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i)^2$$

$$SSE(Y|X,W) = \sum (Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i)^2$$

When calculated on the same data :

$$SSE(Y|X) \geq SSE(Y|X,W)$$

16

Relationships: Estimated SE

$$SSE(Y | X) = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \times X_i)^2$$

$$SSE(Y | X, W) = \sum (Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \times X_i - \hat{\gamma}_2 \times W_i)^2$$

Now $\hat{\beta}_1 = \hat{\gamma}_1$ if

$\hat{\gamma}_2 = 0$, in which case $SSE(Y | X) = SSE(Y | X, W)$

OR

$r_{XW} = 0$, and $SSE(Y | X) > SSE(Y | X, W)$ if $\hat{\gamma}_2 \neq 0$

17

Special Cases

- Behavior of unadjusted and adjusted models according to whether
 - X and W are uncorrelated
 - W is associated with Y after adjustment for X

	$r_{XW} = 0$	$r_{XW} \neq 0$
$\gamma_2 \neq 0$	Precision	Confounding
$\gamma_2 = 0$	Irrelevant	Var Inflation

18

Precision: Linear Regression

- E.g., X , W independent in population (or completely randomized experiment) AND W associated with Y independent of X

$$\rho_{XW} = 0 \quad \gamma_2 \neq 0$$

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \approx \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) > se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

19

Precision: Logistic Regression

- Adjusting for a precision variable
 - Deattenuates slope away from the null
 - Standard errors reflect mean-variance relationship
 - Substantially increased power only in extreme cases
 - » (OR > 5 for equal samples sizes of binary W)

	<u>True Value</u>	<u>Estimates</u>
Slopes $\beta_1 > 0$:	$\beta_1 < \gamma_1$	$\hat{\beta}_1 < \hat{\gamma}_1$
$\beta_1 < 0$:	$\beta_1 > \gamma_1$	$\hat{\beta}_1 < \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

20

Precision: Poisson Regression

- Adjusting for a precision variable
 - No effect on the slope (similar to linear regression)
 - log ratios are linear in log means
 - Standard errors reflect mean-variance relationship
 - Virtually no effect on power

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \approx \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) \approx se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \approx s\hat{e}(\hat{\gamma}_1)$

21

Precision: PH Regression

- Adjusting for a precision variable
 - Deattenuates slope away from the null
 - Standard errors stay fairly constant
 - (Complicated result of binomial mean-variance)

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 > 0 : \beta_1 < \gamma_1$	$\hat{\beta}_1 < \hat{\gamma}_1$
	$\beta_1 < 0 : \beta_1 > \gamma_1$	$\hat{\beta}_1 > \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) \approx se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \approx s\hat{e}(\hat{\gamma}_1)$

22

Lin Reg: Stratified Randomization

- Stratified randomization in a designed experiment

$$r_{XW} = 0 \quad \gamma_2 \neq 0$$

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) = se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) > s\hat{e}(\hat{\gamma}_1)$

23

Confounding: Linear Regression

- Causally associated with response and associated with POI in sample

$$r_{XW} \neq 0 \quad \gamma_2 \neq 0$$

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1 + \rho_{XW} \frac{\sigma_X}{\sigma_W} \gamma_2$	$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X (r_{YX} - r_{YW} r_{XW})} \right] \right)$
Std Errs	$se(\hat{\beta}_1) \begin{cases} > \\ = \\ < \end{cases} se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \begin{cases} > \\ = \\ < \end{cases} s\hat{e}(\hat{\gamma}_1)$

24

Relationships: True SE

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

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

$$Var(Y|X) = \gamma_2^2 Var(W|X) + Var(Y|X, W)$$

$$\sigma_{Y|X}^2 = \gamma_2^2 \sigma_{W|X}^2 + \sigma_{Y|X, W}^2$$

25

Confounding: Other Regression

- With logistic, Poisson, PH regression we cannot write down a formula, but
 - As with linear regression, anything can happen

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \gamma_2$	$\hat{\beta}_1 \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} s\hat{e}(\hat{\gamma}_1)$

26

Variance Inflation

- Associated with POI in sample, but not associated with response

$r_{XW} \neq 0$ $\gamma_2 = 0$

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X(r_{XX} - r_{YW}r_{XW})} \right] \right)$

Std Errs	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$
----------	--	--

27

Var Inflation: Other Regressions

- With logistic, Poisson, PH regression we cannot write down a formula, but
 - Similar to linear regression

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

28

Irrelevant Variables

.....

- Uncorrelated with POI in sample, and not associated with response
 - Slight loss of precision in all regressions

$$r_{xw} = 0 \quad \gamma_2 = 0$$

	<u>True Value</u>	<u>Estimates</u>
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1$
Std Errs	$se(\hat{\beta}_1) = se(\hat{\gamma}_1)$	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$

29

Stata Example

.....

FEV and Smoking in Children

30

Stata: Multiple Regression

.....

- In Stata, we use the same commands as were used for simple regression
 - We just list more variable names
 - Interpretation of CI, P values for coefficient estimates now relate to new scientific interpretation of intercept and slopes
 - Test of entire regression model also provided
 - A test that all slopes are equal to 0

31

FEV Dataset

.....

- Association between lung function and self reported smoking in children
 - Compare geometric means of FEV of children who smoke to comparable nonsmokers
 - Restrict analysis to children 9 yo and older
 - No smokers less than 9
 - Still about 6 : 1 ratio of nonsmokers to smokers
 - Little precision gained by keeping younger children
 - Borrowing information from young kids problematic if not a linear relationship between log(FEV) and predictors
 - » With confounding, want to get model correct

32

Compare Alternative Models

- Real life:
 - We should choose a single model in advance of looking at the data
- Academic exercise for this lecture
 - Observe what happens to parameter estimates and SE across models
 - Smoking
 - Smoking adjusted for age
 - Smoking adjusted for age and height

33

Ex: FEV and Smoking

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

```
Number of obs = 439
F( 1, 437) = 10.45
Prob > F = 0.0013
R-squared = 0.0212
Root MSE = .24765
```

		Robust				
<u>logfev</u>		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

34

Unadjusted 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)

35

Unadjusted 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)

36

Ex: Adjusted for Age

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

```
Number of obs =    439
F( 2, 437) =    82.28
Prob > F      =    0.0000
R-squared     =    0.3012
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	

37

Age Adjusted 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: $e^{0.352} = 1.422$)

38

Age Adjusted 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: $P < 0.0005$

39

Age Adjusted 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)

40

Age Adjusted 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 (adj SE 0.034 vs unadj SE 0.031)

41

Ex: Adjusted for Age, Height

```

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

                                Number of obs =    439
                                F( 3, 437) = 284.22
                                Prob > F    = 0.0000
                                R-squared    = 0.6703
                                Root MSE  = .14407

```

		Robust				
	Coef.	St Err	t	P> t	[95% CI]	
logfev						
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	

42

Age, Ht Adjusted 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

43

Age, Ht Adjusted 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

44

Age, Ht Adjusted 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

45

Age, Ht Adjusted 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

46

Age, Ht Adjusted 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 (adj SE 0.024 vs unadj SE 0.034)

47