Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 8:

Adjustment for Confounders, Precision

February 6, 2015

1

Lecture Outline

- Stratified analyses
  - Example: Mantel-Haenszel statistic
- Multiple regression
- Unadjusted vs adjusted analyses
  - Four important questions of statistics
- Examples in Stata:
  - FEV and smoking in children (geometric means and linear regr)
  - Mortality and LDL (odds and logistic regression)

2

Scientific Questions

- Most times:
  - Comparing distribution of response across groups defined by predictor of interest
- Very often, other variables also need to be considered because
  - Comparison is different in strata
  - Groups being compared differ in other ways
  - Less variability of response if we control for other variables

3

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

Statistical Role

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

5

Statistical Methods

- Two general methods to adjust for additional covariates
- Stratified analyses
  - Combines information about association between response and POI across strata
  - Will not borrow information about (or even estimate) association between response and adjustment variables
- Multiple regression
  - Can (but does not have to) borrow information about associations between response and all modeled variables

6

Stratified Analyses

7

Stratified Analyses

- Divide the data into strata based on all combinations of the "adjustment" covariates
  - E.g., every combination of sex, age, race, etc.
- In each stratum, perform an analysis comparing response across POI groups
- Use (weighted) average of estimated associations across groups

8

Stratified Estimates

- This is easy, if estimates are independent and approximately normally distributed

For independent strata $k = 1, \ldots, K$

Estimate in stratum k: $\hat{\theta}_k \sim N(\theta_k, se_k^2)$

Weight for stratum k: w_k

Stratified estimate:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \sim N\left(\theta = \frac{\sum_{k=1}^{K} w_k \theta_k}{\sum_{k=1}^{K} w_k}, \frac{\sum_{k=1}^{K} w_k^2 se_k^2}{\left(\sum_{k=1}^{K} w_k \right)^2} \right)$$

9

Choosing Weights

- Criteria
  - Scientific relevance of stratified estimate ("importance weights")
  - Statistical precision of stratified estimate ("efficiency weights")
- Should be based on statistical role of "adjustment" variables
  - Effect modifiers
  - Confounding
  - Precision

10

Importance Weights

- Sometimes we anticipate effect modification by some variables
- But often we do not choose to report estimate of association between response and POI in each stratum separately
  - E.g., political polls, age adjusted incidence rates
- We want to estimate an "average association" for a population

11

Choice of Importance Weights

- Typically relate to the size of the corresponding stratum in a population of interest
  - Some general target population, or
  - An agreed upon "standard population" for comparison
- Example: in ecologic studies comparing incidence of hip fracture across countries
  - Hip fracture rates increase with age
  - Industrialized countries and developing world have very different age distributions
  - Choose standard age distribution to remove confounding by age
    - E.g., what would be rates if all countries had US age distribution?
- The weights actually used in the analysis might be
  - Population sizes from external sources (census, prior studies)
  - Sample sizes in current study

12

Aside: Oversampling

- In political polls or epidemiologic studies, we sometimes oversample some strata in order to gain precision
  - With binary data, greatest uncertainty when probability is near 0.5
- For fixed maximal sample size, we gain most precision if stratum sample size is proportional to weight times standard deviation of measurements in the stratum

13

Aside: Optimal Sample Sizes for Fixed N

- Typical case for stratified estimates

For independent strata $k = 1, \ldots, K$

Sample size in stratum k: n_k

Estimate in stratum k: $\hat{\theta}_k \sim N\left(\theta_k, se_k^2 = \frac{V_k}{n_k}\right)$

Importance weight for stratum k: w_k

Optimal sample size when fixed $N = \sum_{k=1}^{K} n_k$:

$$\frac{w_1\sqrt{V_1}}{n_1} = \frac{w_2\sqrt{V_2}}{n_2} = \cdots = \frac{w_K\sqrt{V_K}}{n_K}$$

14

Pure Confounders, Precision

- In the absence of effect modification, we have "pure" confounding or "pure" precision
- The role of scientific vs statistical criteria for weights changes
- Scientific Criteria
  - The true association is the same in each stratum
    - We are free to consider statistical criteria
- Statistical Criteria
  - Maximize precision of stratified estimate by minimizing standard error

15

Pure Confounders, Precision

- Association between response and POI is the same in every stratum

For independent strata $k = 1, \ldots, K$

Estimate in stratum k: $\hat{\theta}_k \sim N\left(\theta_k = \theta, se_k^2\right)$

Weight for stratum k: w_k

Stratified estimate:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \sim N\left(\theta, \frac{\sum_{k=1}^{K} w_k^2 se_k^2}{\left(\sum_{k=1}^{K} w_k\right)^2}\right)$$

16

Optimal Weights for Smallest SE

- Typical case for stratified estimates

For independent strata $k = 1, \ldots, K$

Sample size in stratum k : n_k

Estimate in stratum k : $\hat{\theta}_k \sim N\left(\theta_k, se_k^2 = \frac{V_k}{n_k}\right)$

Efficiency weight for stratum k : $w_k = \frac{n_k}{V_k}$

Smallest standard error for $\hat{\theta} = \sum_{k=1}^{K} w_k \hat{\theta}_k$:

$$\frac{w_1 V_1}{n_1} = \frac{w_2 V_2}{n_2} = \cdots = \frac{w_K V_K}{n_K} = 1$$

17

Choice of Weights

- Stratum specific estimate most often a difference of means of (possibly transformed) observations
  - "Efficient score statistic" is the efficient transformation of data
- Exact form of *V* includes within group variances and sample size

Differences across POI in stratum k :

$$\hat{\theta}_k = \overline{U}_{k1} - \overline{U}_{k0} \sim N\left(\theta, \left(\frac{\sigma_{k1}^2}{n_{k1}} + \frac{\sigma_{k0}^2}{n_{k0}}\right)\right) \qquad w_k = \frac{n_{k1} n_{k0}}{n_{k1}\sigma_{k0}^2 + n_{k0}\sigma_{k1}^2}$$

(This can generalize to a regression coefficient in each stratum)

18

Choice of Weights: Ignoring Mean-Vrnc

- Sometimes we assume that the mean-variance relationship is not too great
  - Weights depend only on sample sizes within each POI group
  - Population based sampling and balance: importance weights

Differences across POI in stratum k :

$$\hat{\theta}_k = \overline{U}_{k1} - \overline{U}_{k0} \sim N\left(\theta, \left(\frac{\sigma_{k1}^2}{n_{k1}} + \frac{\sigma_{k0}^2}{n_{k0}}\right)\right) \qquad w_k = \frac{n_{k1} n_{k0}}{n_{k1}\sigma_{k0}^2 + n_{k0}\sigma_{k1}^2}$$

If any mean - variance relationship is ignored : harmonic mean of sample sizes within stratum

$$\hat{\theta} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \qquad w_k = \frac{n_{k1} n_{k0}}{n_{k1} + n_{k0}}$$

19

Ex: Mantel-Haenszel Statistic

- Hypothesis test comparing odds (proportions) across two groups
- Adjust for confounding in a stratified analysis
- Weights chosen for statistical precision
  - Not quite the most optimal weights but close
  - Actual statistic uses stratum specific standard errors computed using hypergeometric distribution rather than binomial distribution

20

Ex: Mantel-Haenszel Statistic

- Approximate weighting of difference in proportions based on harmonic means of sample sizes in each stratum
  - Usually viewed as weighted odds ratios (next slide)

Sample size in stratum k : n_{1k}, n_{0k}

Estimates in stratum k : $\hat{p}_{1k}, \hat{p}_{0k}$

Weighted average across strata : $\hat{\theta} = \dfrac{\sum w_k (\hat{p}_{1k} - \hat{p}_{0k})}{\sum w_k}$

Precision weight for stratum k :

$$w_k = \frac{\frac{n_{1k} n_{0k}}{n_{1k} + n_{0k}}}{\sum_{k=1}^{K} \frac{n_{1k} n_{0k}}{n_{1k} + n_{0k}}}$$

21

Ex: Mantel-Haenszel Statistic: OR

- Approximate weighting of odds ratios yields same test of independence

Data in stratum k : POI

	$Y=1$	$Y=0$	
1	a_k	b_k	n_{1k}
0	c_k	d_k	n_{0k}
	m_{1k}	m_{0k}	N_k

Estimates in stratum k : $\hat{\theta}_k = \dfrac{a_k d_k}{b_k c_k}$

Weighted average across strata : $\hat{\theta} = \dfrac{\sum w_k \hat{\theta}_k}{\sum w_k}$

Weight for stratum k : $w_k = \dfrac{b_k c_k}{n_{1k} + n_{0k}}$

22

Ex: Stata Mantel-Haenszel

- Odds of being full professor by sex

```
. cc full female if year==95, by(field)

field |    OR     [95% ConfInt]      M-H Weight
 Arts |   .538     .293    .984       16.545 (exact)
Other |   .254     .187    .344       91.645 (exact)
 Prof |   .343     .164    .705       14.426 (exact)
Crude |   .290     .227    .372              (exact)
  M-H |   .303     .238    .386
Test of homogeneity: chi2(2)= 5.47  Pr>chi2 = 0.0648
Test that OR =1    :         MH chi2(1) =     99.10
                             Pr>chi2    =    0.0000
```

23

Ex: Interpretation

- The odds of a female faculty member being a full professor is 69.7% less than the odds of a male faculty member in her same academic field having that rank (OR = 0.303)
- Based on a Mantel-Haenszel based 95% CI, the observed data would not be atypical in a setting in which the true odds ratio were anywhere between 0.238 to 0.386.
- The highly statistically significant two-sided P < 0.0001 suggests that we can with high confidence reject the null hypothesis that women faculty and men faculty in the same field have the same odds of being a full professor.
- (The chi squared test of homogeneity is a test for effect modification. I would not "pre-test" for homogeneity, though I might descriptively note that in post-hoc analyses there was insufficient evidence to say the OR differed by field.)

24

Multiple Regression

25

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric survival regr

26

General Regression

- General notation for variables and parameter

Y_i	Response measured on the ith subject
X_i	Value of the POI for the ith subject
$W_{1i}, W_{2i}, \ldots$	Value of adjustment variables for the ith subject
θ_i	Parameter of distribution of Y_i

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

27

Multiple Regression

- General notation for multiple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \cdots$$

$g(\ )$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope for Pred of Interest X"

β_j "Slope for covariate W_{j-1}"

- The link function is usually either none (means) or log (geom mean, odds, hazard)

28

Borrowing Information

- Use other groups to make estimates in groups with sparse data
- Intuitively: 67 and 69 year olds would provide some relevant information about 68 year olds
- Assuming straight line relationship on (possibly transformed) covariates tells us how to adjust data from other (even more distant) age groups
  - If we do not know about the exact functional relationship, we might want to borrow information only close to each group
    - Linear splines, or other flexible modeling methods

29

Defining "Contrasts"

- Define a comparison across groups to use when answering scientific question
- If straight line relationship in parameter, slope for POI is difference in parameter between groups differing by 1 unit in X when all other covariates in model are equal
- If nonlinear relationship in parameter, slope is average difference in parameter between groups differing by 1 unit in X "holding covariates constant"
  - Statistical jargon: a "contrast" across the groups

30

Comparison of Models

- The major difference between regression models is interpretation of the parameters
  - Summary: Mean, geometric mean, odds, hazards
  - Comparison of groups: Difference, ratio
- Issues related to inclusion of covariates remain the same for all regression models
  - Address the scientific question
    - Predictor of interest; Effect modifiers
  - Address confounding
  - Increase precision

31

Interpretation of Parameters

- Intercept
  - Corresponds to a population with all modeled covariates equal to zero
    - Most often outside range of data; quite often impossible; very rarely of interest by itself
- Slope
  - A comparison between groups differing by 1 unit in corresponding covariate, but agreeing on all other modeled covariates
    - Sometimes impossible to use this definition when modeling interactions or complex curves

32

Stratification vs Regression

- Generally, any stratified analysis could be performed as a regression model
- But stratification adjusts for covariates and all interactions among those covariates
  - E.g, sex, race, and the sex-race interaction
- Beware: Our habit in regression is to just adjust for the covariates (the “main effect”), and consider interactions less often

33

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

34

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
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
```

35

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
```

36

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
      logfev |      Coef.   Std. Err      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   -.054     .0241    -2.22   0.027    -.1014   -.0066
         age |    .022     .0035     6.18   0.000     .0151    .0289
       loght |   2.870     .1280    22.42   0.000     2.618    3.121
       _cons |  -11.095    .5153   -21.53   0.000    -12.107  -10.082
------------------------------------------------------------------------------
```

37

Adjustment for Confounders, Precision Variables

Four Important Questions of Statistics

38

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”

39

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

40

Interpretation of Slopes

- Difference in interpretation of slopes

$$\text{Unadjusted Model}: \quad g\left[\theta \mid X_i\right] = \beta_0 + \beta_1 \times X_i$$

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

$$\text{Adjusted Model}: \quad g\left[\theta \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

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

41

Comparing models

- Four important questions

Unadjusted $g\left[\theta \mid X_i\right] = \beta_0 + \beta_1 \times X_i$

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

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

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

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

When is $\hat{se}\left(\hat{\gamma}_1\right) = \hat{se}\left(\hat{\beta}_1\right)$?

42

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

43

Linear Regression

- Difference in interpretation of slopes

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

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

$$\text{Adjusted Model}: \quad E\left[Y_i \mid X_i, W_i\right] = \gamma_0 + \gamma_1 \times X_i + \gamma_2 \times W_i$$

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

44

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)

45

Relationships: Estimated Slopes

- The estimated slope of the unadjusted model will be

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

- 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)

46

Relationships: True SE

- Relationship to within group variance (RMSE) and correlation among predictors

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

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

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

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

47

Relationships: True SE

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

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

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

Thus, $se\left(\hat{\beta}_1\right) = se\left(\hat{\gamma}_1\right)$ if

$$r_{XW} = 0$$

AND

$$\gamma_2 = 0 \quad \text{OR} \quad Var\left(W \mid X\right) = 0$$

48

Relationships: Estimated SE

- Estimated SE also involves degrees of freedom to estimate RMSE

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

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

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

49

Relationships: Estimated SE

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

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

Thus, $s\hat{e}\left(\hat{\beta}_1\right) = s\hat{e}\left(\hat{\gamma}_1\right)$ if

$r_{XW} = 0$

AND

$$SSE\left(Y \mid X\right)/(n-2) = SSE\left(Y \mid X,W\right)/(n-3)$$

50

Residual Squared Error

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

When calculated on the same data :

$$SSE\left(Y \mid X\right) \geq SSE\left(Y \mid X,W\right)$$

51

Relationships: Estimated SE

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

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

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

OR

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

52

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

53

Precision: Linear Regression

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

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

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

54

Precision: Logistic Regression

- Adjusting for a precision variable
  - Deattenuates slope away from the null (not collapsible)
  - Standard errors reflect mean-variance relationship
    - Substantially increased power only in extreme cases
      - (OR > 5 for equal samples sizes of binary W)

		True Value	Estimates
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		$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

55

Precision: Poisson Regression

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

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

56

Precision: PH Regression

- Adjusting for a precision variable
  - Deattenuates slope away from the null (not collapsible)
  - Standard errors stay fairly constant
    - (Complicated result of binomial mean-variance)

		True Value	Estimates
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		$se(\hat{\beta}_1) \approx se(\hat{\gamma}_1)$	$\hat{se}(\hat{\beta}_1) \approx \hat{se}(\hat{\gamma}_1)$

57

Lin Reg: Stratified Randomization

- Stratified randomization in a designed experiment

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

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

58

Confounding: Linear Regression

- Causally associated with response and associated with POI in sample

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

	True Value	Estimates
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	$se(\hat{\beta}_1) \left\{\begin{matrix}>\\=\\<\end{matrix}\right\} se(\hat{\gamma}_1)$	$\hat{se}(\hat{\beta}_1) \left\{\begin{matrix}>\\=\\<\end{matrix}\right\} \hat{se}(\hat{\gamma}_1)$

59

Relationships: True SE

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

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

$$Var(Y \mid X) = \gamma_2^2 Var(W \mid X) + Var(Y \mid X, W)$$
$$\sigma^2_{Y|X} = \gamma_2^2 \sigma^2_{W|X} + \sigma^2_{Y|X,W}$$

60

Confounding: Other Regression

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

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

61

Variance Inflation

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

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

	True Value	Estimates
Slopes	$\beta_1 = \gamma_1$	$\hat{\beta}_1 = \hat{\gamma}_1 \left(1 + \hat{\gamma}_2 r_{XW} \left[\frac{s_W}{s_X (r_{YX} - r_{YW} r_{XW})} \right] \right)$
Std Errs	$se(\hat{\beta}_1) < se(\hat{\gamma}_1)$	$s\hat{e}(\hat{\beta}_1) < s\hat{e}(\hat{\gamma}_1)$

62

Var Inflation: Other Regressions

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

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

63

Irrelevant Variables

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

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

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

64

Stata Example

FEV and Smoking in Children

65

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

66

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

67

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

68

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
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
```

69

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)

70

Unadjusted Interpretation: Smoker Slope

- 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)

71

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
```

72

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: $e^{0.352} = 1.422$)

73

Age Adjusted Interpretation: Age Slope

- 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

74

Age Adjusted Interpretation: Smoker Slope

- 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)

75

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 nonsmokers: 11.1 (2.04)
    - Mean (SD) of age in analyzed smokers: 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)

76

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
      logfev |      Coef.   Std. Err      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   -.054     .0241    -2.22   0.027    -.1014   -.0066
         age |    .022     .0035     6.18   0.000     .0151    .0289
       loght |   2.870     .1280    22.42   0.000     2.618    3.121
       _cons |  -11.095    .5153   -21.53   0.000    -12.107  -10.082
------------------------------------------------------------------------------
```

77

Age, Ht Adj 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

78

Age, Ht Adj Interpretation: Age Slope

- 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

79

Age, Ht Adj Interpretation: Height Slope

- 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
  - (I truly would not have cared if it were not, however)

80

Age, Ht Adj Interpretation: Smoker Slope

- 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

81

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)

82

Stata Example

Odds of Mortality and LDL

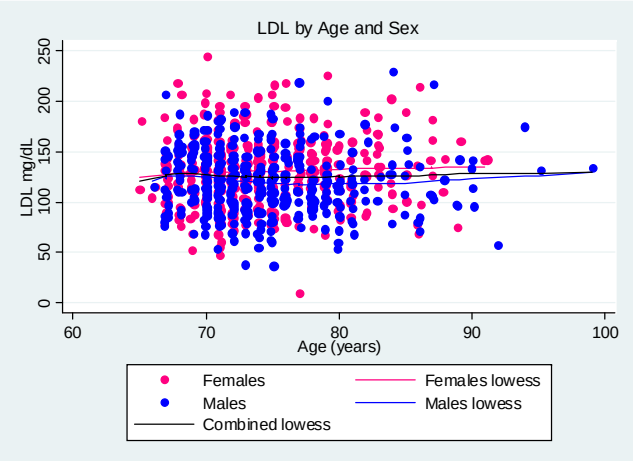
83

Mortality and LDL

- Association between death from all causes and serum LDL at time of study enrollment
  - Recall no censoring prior to 5 years in MRI dataset
- Compare odds of mortality across groups defined by LDL status
  - Model POI using log transform (base 1.1 logarithm)
- Consider adjustment for age, sex
  - Causal associations between mortality and age, sex
    - Well known
  - Associations with POI in sample
    - Regression slope from age on log LDL: -0.0355 / 10% diff in LDL
      - Not very important scientifically
    - Geom mean LDL: females: 126 mg/dL; males 116 mg/dL
      - More substantial association

84

LDL vs Age, Sex



85

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
    - Log LDL
    - Log LDL adjusted for age
    - Log LDL adjusted for sex
    - Log LDL adjusted for age, sex

86

Unadjusted vs Adjusted for Age

```
.g logldl = log(ldl) / log(1.1)
. logistic deadin5 logldl
Logistic regression                     Number of obs   =        725
                                        LR chi2(1)      =       9.26
                                        Prob > chi2     =     0.0023
Log likelihood = -319.05912             Pseudo R2       =     0.0143
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Intvl]
------------+-----------------------------------------------------------------
     logldl |   .9097017   .0282909    -3.04   0.002     .8559    .9669


. logistic deadin5 logldl age
Logistic regression                     Number of obs   =        725
                                        LR chi2(2)      =      27.22
                                        Prob > chi2     =     0.0000
Log likelihood = -310.07946             Pseudo R2       =     0.0420
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf.
      Interval]
------------+-----------------------------------------------------------------
     logldl |   .9096465   .0287098    -3.00   0.003     .8551    .9677
        age |    1.07742   .0187562     4.28   0.000     1.041    1.115
```

87

Interpretation: Age adjustment

- Unadjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL, the odds of all cause mortality are estimated to be 9.03% lower in the group having higher LDL
    - OR = 0.9097 (95% CI 0.8559 to 0.9669; two sided P = 0.002)
- Age adjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL but having the same age, the odds of all cause mortality are estimated to be 9.04% lower in the group having higher LDL
    - OR = 0.9096 (95% CI 0.8551 to 0.9677; two sided P = 0.003)
- Note that not much precision is gained in logistic regression

88

Unadjusted vs Adjusted for Sex

```
.g logldl = log(ldl) / log(1.1)
. logistic deadin5 logldl
Logistic regression                     Number of obs   =        725
                                        LR chi2(1)      =       9.26
                                        Prob > chi2     =     0.0023
Log likelihood = -319.05912             Pseudo R2       =     0.0143
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Intvl]
------------+-----------------------------------------------------------------
     logldl |   .9097017   .0282909    -3.04   0.002     .8559    .9669
------------------------------------------------------------------------------


. logistic deadin5 logldl male
Logistic regression                     Number of obs   =        725
                                        LR chi2(2)      =      19.97
                                        Prob > chi2     =     0.0000
Log likelihood = -313.70647             Pseudo R2       =     0.0308
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Intvl]
------------+-----------------------------------------------------------------
     logldl |   .9200492   .0287831    -2.66   0.008      .8653    .9782
       male |   1.968378   .4141986     3.22   0.001     1.3031    2.973
------------------------------------------------------------------------------
```

89

Interpretation: Sex adjustment

- Unadjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL, the odds of all cause mortality are estimated to be 9.03% lower in the group having higher LDL
    - OR = 0.9097 (95% CI 0.8559 to 0.9669; two sided P = 0.002)
- Sex adjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL but having the same sex, the odds of all cause mortality are estimated to be 8.00% lower in the group having higher LDL
    - OR = 0.9200 (95% CI 0.8653 to 0.9782; two sided P = 0.008)
- Owing to confounding, adjusted estimate is different (and some loss of precision)

90

Unadjusted vs Adjusted for Age, Sex

```
.g logldl = log(ldl) / log(1.1)
. logistic deadin5 logldl
Logistic regression                     Number of obs   =        725
                                        LR chi2(1)      =       9.26
                                        Prob > chi2     =     0.0023
Log likelihood = -319.05912             Pseudo R2       =     0.0143
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Intvl]
------------+-----------------------------------------------------------------
     logldl |   .9097017   .0282909    -3.04   0.002     .8559    .9669
------------------------------------------------------------------------------


. logistic deadin5 logldl male age
Logistic regression                     Number of obs   =        725
                                        LR chi2(3)      =      37.10
                                        Prob > chi2     =     0.0000
Log likelihood = -305.13721             Pseudo R2       =     0.0573
------------------------------------------------------------------------------
    deadin5 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Intvl]
------------+-----------------------------------------------------------------
     logldl |   .9208159   .0291924    -2.60   0.009      .8653    .9798
       male |   1.935794   .4128427     3.10   0.002     1.274     2.940
        age |   1.076029   .0189026     4.17   0.000     1.040     1.114
------------------------------------------------------------------------------
```

91

Interpretation: Age, Sex adjustment

- Unadjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL, the odds of all cause mortality are estimated to be 9.03% lower in the group having higher LDL
    - OR = 0.9097 (95% CI 0.8559 to 0.9669; two sided P = 0.002)
- Age, sex adjusted association between mortality and LDL
  - When comparing two groups differing by 10% in their serum levels of LDL but having the same age and sex, the odds of all cause mortality are estimated to be 8.00% lower in the group having higher LDL
    - OR = 0.9208 (95% CI 0.8653 to 0.9798; two sided P = 0.009)
- Owing to confounding by sex, adjusted estimate is different
  - No real gain in precision from adjusting for age

92