

Biost 518
Applied Biostatistics II
.....
Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 5:
Adjustment for Covariates

January 27, 2006

1

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

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

2

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”

3

Statistical Role
.....

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

4

Statistical Methods

.....

- Adjustment for additional covariates
 - Stratified analyses
 - Multiple regression

5

Statistical Methods

.....

- Adjustment for additional covariates
 - Stratified analyses
 - Combines information about association between response and POI across strata
 - Will not borrow information about (or even estimate) about 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 $k = 1, \dots, K$

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
 - Statistical precision of stratified estimate
- Should be based on statistical role of “adjustment” variables
 - Effect modifiers
 - Confounding
 - Precision

10

Presence of Effect Modification

- Scientific Criteria
 - Sometimes we anticipate effect modification by some variables, but
 - 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

Presence of Effect Modification

- Choose weights according to importance
 - Size of the corresponding stratum in a population of interest
 - The real population, or
 - Some standard population used for comparisons
 - E.g., 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 a standard age distribution to remove confounding by age

12

Aside: Oversampling

- In political polls or epidemiologic studies, we sometimes oversample some strata in order to gain precision
 - 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: Oversampling

- Typical case for stratified estimates

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

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 $N = \sum_{k=1}^K n_k$:

$$\frac{w_1 \sqrt{V_1}}{n_1} = \frac{w_2 \sqrt{V_2}}{n_2} = \dots = \frac{w_K \sqrt{V_K}}{n_K} \quad 14$$

Confounders, Precision

- 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

Confounders, Precision

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

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

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

Weight for stratum k : w_k

Stratified estimate :

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

16

Optimal Weights for Small SE

- Typical case for stratified estimates

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

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 $N = \sum_{k=1}^K n_k$:

$$\frac{w_1 \sqrt{V_1}}{n_1} = \frac{w_2 \sqrt{V_2}}{n_2} = \dots = \frac{w_K \sqrt{V_K}}{n_K} \quad 17$$

Confounders, Precision

- Often we ignore the aspect that variability might differ across strata
 - Just choose weights by sample size for each stratum
 - Note that if we sampled randomly from the population of interest, this would also be appropriate for importance weights in the presence of effect modification

18

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)

19

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

Sample size in stratum k :

$$n_{1k}, n_{0k}$$

Estimates in stratum k :

$$\hat{p}_{1k}, \hat{p}_{0k}$$

Precision weight for stratum k :

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

20

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 = 0.0000₂₁

Multiple Regression

22

Types of Variables

- Binary data
 - E.g., sex, death
- Nominal data: unordered, categorical data
 - E.g., race, marital status
- Ordinal categorical data
 - E.g., stage of disease
- Quantitative data
 - E.g., age, blood pressure
- Right censored data
 - E.g., time to death (when not everyone has died)₂₃

Summary Measures

- The measures commonly used to summarize and compare distributions vary according to the types of data
 - Means: binary; quantitative
 - Medians: ordered; quantitative; censored
 - Proportions: binary; nominal
 - Odds: binary; nominal
 - Hazards: censored
 - hazard = instantaneous rate of failure

24

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

25

General Regression

- General notation for variables and parameter

Y_i	Response measured on the i th subject
X_i	Value of the POI for the i th subject
W_{1i}, W_{2i}, \dots	Value of adjustment variables for the i th 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.

26

Multiple Regression

- General notation for simple regression model

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

$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)

27

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 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

28

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

29

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
 - Address the scientific question
 - Predictor of interest; Effect modifiers
 - Address confounding
 - Increase precision

30

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

31

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
 - Our habit in regression is to just adjust for the covariates (the “main effect”), and consider interactions less often

32

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

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					
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

34

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

35

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.	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

36