Biost 518
Applied Biostatistics II

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

Lecture 6:
Adjustment for Covariates

January 25, 2008

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

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”

Statistical Role

• Covariates other than the POI are included in the model as
  – Effect modifiers
  – Confounders
  – Precision variables
Statistical Methods

- Adjustment for additional covariates
  - Stratified analyses
  - Multiple regression

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

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
Stratified Estimates

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

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

Estimate in stratum $k$:

$$\hat{\theta}_k \sim N(\theta_k, \text{se}_k^2)$$

Weight for stratum $k$:

$$w_k$$

Stratified estimate:

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

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

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

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
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

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

Aside: Oversampling

- Typical case for stratified estimates
  For independent strata \( k = 1, \ldots, K \)
  \[
  \text{Sample size in stratum } k : \quad n_k \\
  \text{Estimate in stratum } k : \quad \hat{\theta}_k \sim N\left( \theta, \frac{\sigma_k^2}{n_k} \right) \\
  \text{Importance weight for stratum } k : \quad w_k \\
  \text{Optimal sample size when fixed } N = \sum_{k=1}^{K} n_k : \\
  \frac{w_1 \sqrt{V_1}}{n_1} = \frac{w_2 \sqrt{V_2}}{n_2} = \cdots = \frac{w_K \sqrt{V_K}}{n_K}
  \]

Confounders, Precision

- Association between response and POI is the same in every stratum
  For independent strata \( k = 1, \ldots, K \)
  \[
  \hat{\theta}_k \sim N\left( \theta, \frac{\sigma_k^2}{w_k} \right) \\
  \text{Weight for stratum } k : \quad w_k \\
  \text{Stratified estimate : } \\
  \hat{\theta} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \sim N\left( \theta, \frac{\sum_{k=1}^{K} w_k^2 \sigma_k^2}{\left(\sum_{k=1}^{K} w_k\right)^2} \right)
  \]
Optimal Weights for Small SE

• Typical case for stratified estimates
  For independent strata $k = 1, \ldots, K$
  Sample size in stratum $k : n_k$
  Estimate in stratum $k : \hat{\theta}_k \sim \mathcal{N}(\theta_k, \sigma_k^2 = \frac{V_k}{n_k})$
  Importance weight for stratum $k : w_k$
  Optimal sample size when fixed $N = \sum_{k=1}^{K} n_k :$
  \[
  \frac{w_1 \sqrt{V_1}}{n_1} = \frac{w_2 \sqrt{V_2}}{n_2} = \ldots = \frac{w_K \sqrt{V_K}}{n_K}
  \]

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

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)
Ex: Stata Mantel-Haenszel

- Odds of being full professor by sex

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

<table>
<thead>
<tr>
<th>field</th>
<th>OR</th>
<th>[95% ConfInt]</th>
<th>M-H Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>.538</td>
<td>.293 .984</td>
<td>16.545 (exact)</td>
</tr>
<tr>
<td>Other</td>
<td>.254</td>
<td>.187 .344</td>
<td>91.645 (exact)</td>
</tr>
<tr>
<td>Prof</td>
<td>.343</td>
<td>.164 .705</td>
<td>14.426 (exact)</td>
</tr>
<tr>
<td>Crude</td>
<td>.290</td>
<td>.227 .372</td>
<td>(exact)</td>
</tr>
<tr>
<td>M-H</td>
<td>.303</td>
<td>.238 .386</td>
<td></td>
</tr>
</tbody>
</table>

Test of homogeneity: chi2(2)= 5.47  Pr>chi2 = 0.0648
Test that OR =1 :       MH chi2(1) = 99.10  Pr>chi2 = 0.0000

Multiple Regression

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)

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
Regression Models

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

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
  $W_{i,1}, W_{i,2}, \ldots$ Value of adjustment variables for the $i$th subject
  $\theta_i$ Parameter of distribution of $Y_i$

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

Multiple Regression

- General notation for simple regression model
  $$g(\hat{\theta}_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{i,1} + \beta_3 \times W_{i,2} + \cdots$$
  $g()$ "link" function used for modeling
  $\beta_0$ "Intercept"
  $\beta_1$ "Slope for Pred of Interest $X$"
  $\beta_j$ "Slope for covariate $W_{j-1}$"

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

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
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

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

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

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
**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

**Ex: FEV and Smoking**

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

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>logfev</td>
<td>Coef. St. Err</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
</tr>
<tr>
<td>smoker</td>
<td>.102 .0317</td>
<td>3.23</td>
<td>0.001</td>
<td>.040</td>
</tr>
<tr>
<td>_cons</td>
<td>1.058 .0129</td>
<td>81.82</td>
<td>0.000</td>
<td>1.033</td>
</tr>
</tbody>
</table>

**Ex: Adjusted for Age**

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

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>logfev</td>
<td>Coef. St. Err</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
</tr>
<tr>
<td>smoker</td>
<td>-.051 .0344</td>
<td>-1.49</td>
<td>0.136</td>
<td>-.119</td>
</tr>
<tr>
<td>age</td>
<td>.064 .0051</td>
<td>12.37</td>
<td>0.000</td>
<td>.053</td>
</tr>
<tr>
<td>_cons</td>
<td>0.352 .0875</td>
<td>6.12</td>
<td>0.000</td>
<td>.239</td>
</tr>
</tbody>
</table>

**Ex: Adjusted for Age, Height**

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

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>logfev</td>
<td>Coef. St. Err</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
</tr>
<tr>
<td>smoker</td>
<td>-.054 .0241</td>
<td>-2.22</td>
<td>0.027</td>
<td>-.101</td>
</tr>
<tr>
<td>age</td>
<td>.022 .0035</td>
<td>6.18</td>
<td>0.000</td>
<td>.015</td>
</tr>
<tr>
<td>loght</td>
<td>2.870 .1280</td>
<td>22.42</td>
<td>0.000</td>
<td>2.618</td>
</tr>
<tr>
<td>_cons</td>
<td>-11.095 .5153</td>
<td>-21.53</td>
<td>0.000</td>
<td>-12.107</td>
</tr>
</tbody>
</table>