**Module 3:**
An Example of a Two-stage Model; NMMAPS Study

Instructor: Elizabeth Johnson
Course Developed: Francesca Dominici and Michael Griswold
The Johns Hopkins University
Bloomberg School of Public Health

---

**NMMAPS Example of Two-Stage Hierarchical Model**

- National Morbidity and Mortality Air Pollution Study (NMMAPS)
- Daily data on cardiovascular/respiratory mortality in 10 largest cities in U.S.
- Daily particulate matter (PM10) data
- Log-linear regression estimate relative risk of mortality per 10 unit increase in PM10 for each city
- Estimate and statistical standard error for each city
Semi-Parametric Poisson Regression

\[ \log \mu_t = \beta_0 + \beta x_{t-t} + s_1(\text{time, } \lambda_1) + s_2(\text{temp, } \lambda_2) + \text{others} \]

Relative Risks* for Six Largest Cities

<table>
<thead>
<tr>
<th>City</th>
<th>RR Estimate (% per 10 micrograms/ml)</th>
<th>Statistical Standard Error</th>
<th>Statistical Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>0.25</td>
<td>0.13</td>
<td>.0169</td>
</tr>
<tr>
<td>New York</td>
<td>1.4</td>
<td>0.25</td>
<td>.0625</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.60</td>
<td>0.13</td>
<td>.0169</td>
</tr>
<tr>
<td>Dallas/Ft Worth</td>
<td>0.25</td>
<td>0.55</td>
<td>.3025</td>
</tr>
<tr>
<td>Houston</td>
<td>0.45</td>
<td>0.40</td>
<td>.1600</td>
</tr>
<tr>
<td>San Diego</td>
<td>1.0</td>
<td>0.45</td>
<td>.2025</td>
</tr>
</tbody>
</table>

*Approximate values read from graph in Daniels, et al. 2000. AJE
City-specific MLEs for Log Relative Risks

Notation

- \( \hat{\beta}_c \) is the city-specific relative risk estimate
- \( \beta_c \) is the "true" city-specific relative risk
- \( \alpha \) is the average relative risk over all the cities
- \( e_c \) is the statistical error
- \( d_c \) is the deviation of the true city-specific relative risk from the overall mean
Sources of Variation

- $\hat{\beta}_c = \beta_c + e_c$
- $\beta_c = \alpha + d_c$
- $\hat{\beta}_c = \alpha + d_c + e_c$
- $\text{var}(\hat{\beta}_c) = \text{var}(e_c) + \text{var}(d_c) = v_c + NV$

City-specific MLEs for Log Relative Risks (*) and True Values (o)
City-specific MLEs for Log Relative Risks (* and True Values (o))

City-specific MLEs for Log Relative Risks
City-specific MLEs for Log Relative Risks

Notation

- \( v_c \) is the statistical variance
- \( e_c \) is the statistical error
- \( d_c \) is the deviation of city-specific true relative risk from the average
- \( \text{var}(e_c) \) is the random noise
- \( NV = \text{var}(d_c) \) is the variance of the true relative risks across cities, also called natural variance or heterogeneity
- \( TV_c = NV + \text{var}(e_c) \) is the total variance
Estimating Overall Mean

- Idea: give more weight to more precise values

- Specifically, weight estimates inversely proportional to their variances

Estimating the Overall Mean

We can estimate the average relative risk over all the cities by:

\[ \hat{\alpha} = \left( \sum_c w_c \right)^{-1} \sum_c w_c \hat{\beta}_c \]
\[ \text{var}(\hat{\alpha}) = \left( \sum_c h_c \right)^{-1} \]

The weights are calculated as follows:

\[ h_c = \frac{1}{v_c + \bar{N}V} \]
\[ w_c = \frac{h_c}{\sum_c h_c}, \sum_c w_c = 1 \]
\[ \bar{N}V = \frac{1}{N-1} \sum_c (\hat{\beta}_c - \bar{\beta})^2 - \frac{1}{N} \sum_c v_c \]
Calculations for Empirical Bayes Estimates

<table>
<thead>
<tr>
<th>City</th>
<th>Log RR (bc)</th>
<th>Stat Var (vc)</th>
<th>Total Var (TVc)</th>
<th>1/TVc</th>
<th>wc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.25</td>
<td>.0169</td>
<td>.0994</td>
<td>10.1</td>
<td>.27</td>
</tr>
<tr>
<td>NYC</td>
<td>1.4</td>
<td>.0625</td>
<td>.145</td>
<td>6.9</td>
<td>.18</td>
</tr>
<tr>
<td>Chi</td>
<td>0.60</td>
<td>.0169</td>
<td>.0994</td>
<td>10.1</td>
<td>.27</td>
</tr>
<tr>
<td>Dal</td>
<td>0.25</td>
<td>.3025</td>
<td>.385</td>
<td>2.6</td>
<td>.07</td>
</tr>
<tr>
<td>Hou</td>
<td>0.45</td>
<td>.160</td>
<td>.243</td>
<td>4.1</td>
<td>.11</td>
</tr>
<tr>
<td>SD</td>
<td>1.0</td>
<td>.2025</td>
<td>.285</td>
<td>3.5</td>
<td>.09</td>
</tr>
<tr>
<td>Overall</td>
<td>0.65</td>
<td></td>
<td>37.3</td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[ \alpha = .27 \times 0.25 + .18 \times 1.4 + .27 \times 0.60 + .07 \times 0.25 + .11 \times 0.45 + 0.9 \times 1.0 = 0.65 \]

\[ \text{Var}(\alpha) = \frac{1}{\text{Sum}(1/TVc)} = 0.164^2 \]

Software in R

```r
beta.hat <- c(0.25, 1.4, 0.50, 0.25, 0.45, 1.0)
se <- c(0.13, 0.25, 0.13, 0.55, 0.40, 0.45)
NV <- var(beta.hat) - mean(se^2)
TV <- se^2 + NV
tmp <- 1/TV
ww <- tmp/sum(tmp)
v.alphahat <- sum(ww)^{-1}
alpha.hat <- v.alphahat*sum(beta.hat*ww)
```
Two Extremes

- Natural variance >> Statistical variances
  - Weights $w$ approximately constant $= 1/n$
  - Use ordinary mean of estimates regardless of their relative precision

- Statistical variances >> Natural variance
  - Weight each estimator inversely proportional to its statistical variance

Sensitivity of Inferences to Natural Variance

![Graph showing sensitivity of inferences to natural variance](image.png)
Estimating Relative Risk for Each City

- Disease screening analogy
  - Test result from imperfect test
  - Positive predictive value combines prevalence with test result using Bayes theorem
- Empirical Bayes estimator of the true value for a city is the conditional expectation of the true value given the data $E(\beta_c | \hat{\beta}_c)$

Empirical Bayes Estimation

$$
\tilde{\beta}_c = E[\beta_c | \hat{\beta}_c] = \theta_c \hat{\beta}_c + (1 - \theta_c) \hat{\alpha}
$$

$$
\theta_c = \frac{NV}{NV + var(e_c)}
$$

$$
var(\tilde{\beta}_c | \alpha) \approx \theta_c^2 var(e_c)
$$
### Calculations for Empirical Bayes Estimates

<table>
<thead>
<tr>
<th>City</th>
<th>Log RR</th>
<th>Stat Var (vc)</th>
<th>Total Var (TVc)</th>
<th>1/TVc</th>
<th>wc</th>
<th>$\theta_c$</th>
<th>RR.EB</th>
<th>se RR.EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA</td>
<td>0.25</td>
<td>.0169</td>
<td>.0994</td>
<td>10.1</td>
<td>.27</td>
<td>.83</td>
<td>0.32</td>
<td>0.17</td>
</tr>
<tr>
<td>NYC</td>
<td>1.4</td>
<td>.0625</td>
<td>.145</td>
<td>6.9</td>
<td>.18</td>
<td>.57</td>
<td>1.1</td>
<td>0.14</td>
</tr>
<tr>
<td>Chi</td>
<td>0.60</td>
<td>.0169</td>
<td>.0994</td>
<td>10.1</td>
<td>.27</td>
<td>.83</td>
<td>0.61</td>
<td>0.11</td>
</tr>
<tr>
<td>Dal</td>
<td>0.25</td>
<td>.3025</td>
<td>.385</td>
<td>2.6</td>
<td>.07</td>
<td>.21</td>
<td>0.56</td>
<td>0.12</td>
</tr>
<tr>
<td>Hou</td>
<td>0.45</td>
<td>.160</td>
<td>.243</td>
<td>4.1</td>
<td>.11</td>
<td>.34</td>
<td>0.58</td>
<td>0.14</td>
</tr>
<tr>
<td>SD</td>
<td>1.0</td>
<td>.2025</td>
<td>.285</td>
<td>3.5</td>
<td>.09</td>
<td>.29</td>
<td>0.75</td>
<td>0.13</td>
</tr>
<tr>
<td>Overall</td>
<td>0.65</td>
<td>1/37.3= 0.027</td>
<td>37.3</td>
<td>1.00</td>
<td></td>
<td></td>
<td>0.65</td>
<td>0.16</td>
</tr>
</tbody>
</table>

---

### City-specific MLEs for Log Relative Risks

![City-specific MLEs for Log Relative Risks](image)
City-specific MLEs (Left) and Empirical Bayes Estimates (Right)

Shrinkage of Empirical Bayes Estimates

Maximum likelihood estimates

Empirical Bayes estimates
Key Ideas

- Better to use data for all cities to estimate the relative risk for a particular city
  - Reduce variance by adding some bias
  - Smooth compromise between city specific estimates and overall mean

- Empirical-Bayes estimates depend on measure of natural variation
  - Assess sensitivity to estimate of NV
Caveats

- Used simplistic methods to illustrate the key ideas:
  - Treated natural variance and overall estimate as known when calculating uncertainty in EB estimates
  - Assumed normal distribution or true relative risks
- Can do better using Markov Chain Monte Carlo methods – more to come