Discussion:
Random effects versus fixed effects

Sandrah P. Eckel, Ph.D.
eckel@usc.edu

Career Development Meeting  USC  Feb 26, 2009
Fixed effects vs. random effects

Focus on the simplest case:

- Response $Y$ for participant $i$ in town $j$: $Y_{ij}$
- **Aim**: account for town-specific differences in $Y$

**Fixed effect model: indicators for town**

$$y_{ij} = \beta_0 + \alpha_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

In practice: use indicator ("dummy") variables for each town

**Random effect model: random intercepts for town**

$$y_{ij} = \beta_0 + U_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad U_j \sim N(0, \tau^2)$$
Simulated clustered data
Simulated data: fixed effects model

\[ y_{ij} = \beta_0 + \alpha_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \]

Mean deviation of neighborhood 4 from overall mean \( \hat{\alpha}_4 \)

Overall mean \( \hat{\beta}_0 \)
Simulated data: random effects model

\[ y_{ij} = \beta_0 + U_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad U_j \sim N(0, \tau^2) \]
What’s different with random effect estimates?

- Shrinkage
- Weighted average of overall mean and cluster-specific mean
- Minimizes mean squared error (MSE = Variance + Bias²)
  - Trades variance for bias

\[ \hat{U}_j = \rho_j \bar{y}_j + (1 - \rho_j) \bar{y} \]
\[ \rho_j = \frac{\tau^2}{\tau^2 + \sigma^2 / n_j} \]
Example: Baseball batting averages  
- Efron and Morris, 1977

**Goal:** predict end of season batting average based on preliminary data

**Data:** 1970 season  
18 players  
first 45 times at bat

**Unbiased estimate:**  
batting average after 45 times at bat

**Shrinkage estimate:**  
Use information from other players, shrink estimates towards overall mean
Shrinkage estimates
16 of 18 closer to “true” season average
Example: Estimating radon levels
- Gelman & Hill, 2007

- Reduce variance by adding bias
- Borrow strength
- Compromise between overall mean and county-specific mean

Fixed effects

Random effects
Example: NMMAAPS

City-specific and regional estimates
Which should we use: random or fixed?

Common advice:

<table>
<thead>
<tr>
<th>Target of inference</th>
<th>FIXED</th>
<th>RANDOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters in this dataset</td>
<td>“Large” # within cluster</td>
<td>Population of clusters</td>
</tr>
<tr>
<td>“Large” # of clusters</td>
<td>More parameters</td>
<td>Fewer parameters</td>
</tr>
<tr>
<td>Unbiased estimation</td>
<td>Cluster-specific means</td>
<td>Prediction (smaller MSE)</td>
</tr>
<tr>
<td>Overall mean</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

General advice:

- It depends
- Explore by estimating group-specific models, if possible

Andrew Gelman’s advice:

“Always use multilevel modeling (‘random effects’)”
- Gelman & Hill (2007), p.246
What about covariates correlated with $U_j$?

**Assumption**: $\text{Cov}(U_j, x_{ij}) = 0$
- Random intercepts are uncorrelated with the covariates
- Random intercepts represent effects of unobserved covariates that may correlate with observed covariates
- If we omit covariate $w_j$ that’s correlated with $U_j$ and $x_{ij}$
  - Random intercept: estimate of coefficient on $x_{ij}$ may be biased
  - Fixed effects: estimate of coefficient on $x_{ij}$ still consistent
- Hausman specification test
  - Compares coefficients in the two models
  - If significantly different, suggests using the fixed effects model
- Can still use random intercept model and protect against bias in coefficient on $x_{ij}$ by including $\bar{x}_j$ as another covariate
Example: Simplified CHS model

\[ y_{ci} = \beta_0 + \beta_1 (x_{ci} - \bar{x}_c.) + \beta_2 \bar{x}_c. + U_c + \varepsilon_{ci} \]

\[ \varepsilon_{ci} \sim N(0, \sigma^2), \quad U_c \sim N(0, \tau^2) \]

- \( y_{ci} \): lung function in city \( c \), person \( i \)
- \( x_{ci} \): air pollution (e.g. O\(_3\))

- Omitted city-level covariate (e.g. elevation) may be correlated with air pollution and \( U_c \)
- Prevent bias in estimate of \( \beta_1 \) by including the city-level mean pollution
  - Allows simultaneous estimation of between and within city effects
Should we be using either? Does Y differ by cluster?

**Fixed:**  \[ y_{ij} = \beta_0 + \alpha_j + \varepsilon_{ij}, \; \varepsilon_{ij} \sim N(0, \sigma^2) \]

- Test: ANOVA, Likelihood ratio test
- Relative fit criteria: AIC, etc…
- Scientifically meaningful \( \hat{\alpha}_j \)

**Random:**  \[ y_{ij} = \beta_0 + U_j + \varepsilon_{ij}, \; \varepsilon_{ij} \sim N(0, \sigma^2), \; U_j \sim N(0, \tau^2) \]

- Test if variance \( \tau^2 \) different from 0
  - Problem: 0 on boundary of parameter space
- Assess confidence interval for \( \tau^2 \)
  - Is variance scientifically meaningful?