Combining Data by Statistical Matching, Imputation and Modeling

Purpose for combining data

• Improve coverage
  • Survey data from different frames (e.g. landline and cell phone)
• Increase sample size
  • Meta analysis
  • Combining probability sample with nonprobability sample (improves coverage as well)
• Bring together variables from different files
  • Neighborhood Air quality measurements
Data Linkage

Entity Level (Individual Records)
  - Record Linkage (Entity Resolution)
  - Statistical Matching (Data Fusion)

Area Level (Group Statistics)
  - Direct Area Match

Entity-to-Area (Multi-level)
  - Aggregated Entity Match with Area
  - Model-based Entity from Area Distribution
### Statistical Matching

- Record’s measurements are at the same level
- Little-to-no overlap of records across samples

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_{111} )</td>
<td>( x_{211} )</td>
</tr>
<tr>
<td>( y_{112} )</td>
<td>( x_{212} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( y_{11q} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( x_{111} )</td>
<td>( z_{211} )</td>
</tr>
<tr>
<td>( x_{121} )</td>
<td>( z_{212} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( x_{11p} )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample 2</th>
<th>Sample 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{221} )</td>
<td>( y_{1n11} )</td>
</tr>
<tr>
<td>( x_{222} )</td>
<td>( y_{1n12} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( x_{22p} )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( z_{221} )</td>
<td>( x_{1n11} )</td>
</tr>
<tr>
<td>( z_{222} )</td>
<td>( x_{1n12} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( z_{22r} )</td>
<td>( x_{1n1p} )</td>
</tr>
</tbody>
</table>
Combining Multiple Complex Surveys


**Start:** Multiple surveys where key variables are contained in many, but not all surveys

- Each survey used different designs and data collection methods, so the sampling and nonsampling error properties are different
- Cannot simply pool data for analysis

**Step 1:** For each survey

- Construct a model based on the sample design and the relationships in the data
- Generate synthetic populations using data from each survey

Each generated population inverts the sample design to create what is effectively a simple random sample.

**Step 2:** Pool data and use standard imputation approaches to fill in missing variables for the data from each survey.