What is ecological inference (EI)?

- **Goal**: infer individual level behavior from aggregate data.
- **Unit of analysis**: contingency table with observed marginals.

![Contingency Table](image)

\[
\begin{array}{c|ccc|c}
\text{row} & \text{col}_1 & \text{col}_2 & \text{col}_3 & \text{row} \\
\hline
\text{row}_1 & N_{11} & N_{12} & N_{13} & N_{1} \\
\text{row}_2 & N_{21} & N_{22} & N_{23} & N_{2} \\
\text{row}_3 & N_{31} & N_{32} & N_{33} & N_{3} \\
\end{array}
\]

- **eiPack** methods estimate unobserved internal cells (or functions thereof).
- Other packages focus on 2 × 2 inference (e.g., eco, MCMCpack).
- **eiPack**: R × C inference.
Other packages focus on $2 \times 2$ inference (e.g., eco, MCMCpack)
eiPack: $R \times C$ inference
eiPack methods:
- Method of bounds
- Ecological regression
- Multinomial-Dirichlet model

eiPack data: senc
- Individual level party affiliation
- Black, White, and Native American voters
- 8 counties (212 precincts) in SE North Carolina
- Cell counts known

The models implemented in eiPack share:
- A common input syntax of the form:
  \[ \text{cbind}(\text{col1, ..., colC}) \sim \text{cbind}(\text{row1, ...,rowR}) \]
- Functions to calculate proportions of some subset of columns
- Appropriate print, summary, and plot functions
Method of bounds

- Quantity of interest: proportion of row members in each column for each unit
- Observed row and column marginals determine upper and lower bounds
- Row thresholds implemented for extreme case analysis

Output:

<table>
<thead>
<tr>
<th></th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>0.519</td>
<td>0.559</td>
</tr>
<tr>
<td>25</td>
<td>0.450</td>
<td>0.469</td>
</tr>
<tr>
<td>28</td>
<td>0.392</td>
<td>0.487</td>
</tr>
</tbody>
</table>

Olivia Lau, Ryan T. Moore, Michael Kellermann

eiPack: R × C Ecological Inference and Data Management
Method of bounds

- Precincts at least 90% White
- Proportion Democratic: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0

Ecological regression

- Express data as proportions of row totals
- Regress each column on all row proportions (C regressions)
- Coefficients estimate cell proportions

Ecological regression

- Express data as proportions of row totals
- Regress each column on all row proportions (C regressions)
- Coefficients estimate cell proportions
- \texttt{eiPack}: freq. and Bayesian regression

\texttt{lambda} functions calculate shares of a subset of columns – e.g. “among Blacks, Dem. share of 2-party registration”
Ecological regression

Multinomial-Dirichlet (MD) model

Express data as counts

Fit hierarchical Bayesian model

- Level 1: column marginals $\sim$ Multinomial, $\perp$ across units
- Level 2: rows of cell fractions $\sim$ Dirichlet, $\perp$ across rows and units
- Level 3: Dirichlet parameters $\sim$ Gamma, i.i.d.

lambda and density.plot functions
Reasonable-sized problems produce unreasonable amounts of data

E.g., a model for voting in Ohio includes
- 11000 precincts
- 3 racial groups
- 4 party options

1000 iterations yields about $1.3 \times 10^8$ parameter draws

Draws occupy $\approx 1$ GB of RAM; probably not enough iterations

eiPack allows users to write chains to disk, or discard chains not of interest
Visit our poster for more!