`eiPack`: Tools for R × C Ecological Inference and Higher-Dimension Data Management

Olivia Lau   Ryan T. Moore   Michael Kellermann

Department of Government
Institute for Quantitative Social Science
Harvard University

Vienna, Austria
16 June 2006

What is ecological inference (EI)?

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

	col_1	col_2	col_3	
row_1	N_{11i}	N_{12i}	N_{13i}	$N_{1\cdot i}$
row_2	N_{21i}	N_{22i}	N_{23i}	$N_{2\cdot i}$
row_3	N_{31i}	N_{32i}	N_{33i}	$N_{3\cdot i}$
	$N_{\cdot 1i}$	$N_{\cdot 2i}$	$N_{\cdot 3i}$	N_i

What is ecological inference (EI)?

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

	col_1	col_2	col_3	
row_1	N_{11i}	N_{12i}	N_{13i}	$N_{1\cdot i}$
row_2	N_{21i}	N_{22i}	N_{23i}	$N_{2\cdot i}$
row_3	N_{31i}	N_{32i}	N_{33i}	$N_{3\cdot i}$
	$N_{\cdot 1i}$	$N_{\cdot 2i}$	$N_{\cdot 3i}$	N_i

- `eiPack` methods estimate unobserved internal cells (or functions thereof)

`eiPack`

- Other packages focus on 2 × 2 inference (e.g., `eco`, `MCMCpack`)
- `eiPack`: $R \times C$ inference

eiPack

- Other packages focus on 2 × 2 inference (e.g., `eco`, `MCMCpack`)
- `eiPack`: $R \times C$ inference
- `eiPack` methods:
  - Method of bounds
  - Ecological regression
  - Multinomial-Dirichlet model

eiPack

- Other packages focus on 2 × 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

eiPack

The models implemented in `eiPack` share:

eiPack

The models implemented in `eiPack` share:

- A common input syntax of the form:

  `cbind(col1, ..., colC)` ∼ `cbind(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

Method of bounds

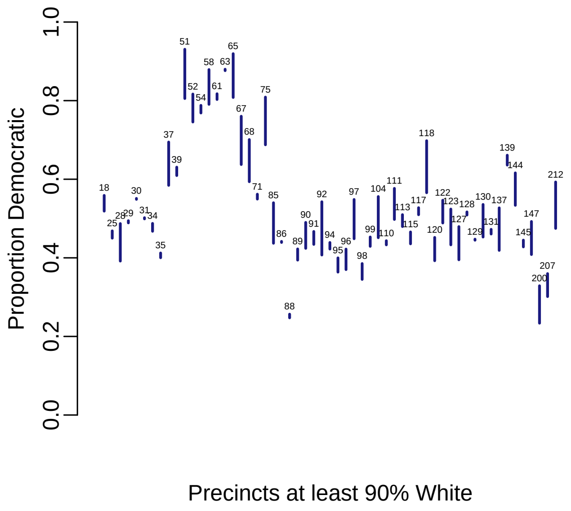
- Quantity of interest: proportion of row members in each column for each unit
- Observed row and column marginals determine upper and lower bounds
- Row thresholds implemented for *extreme case analysis*

Method of bounds

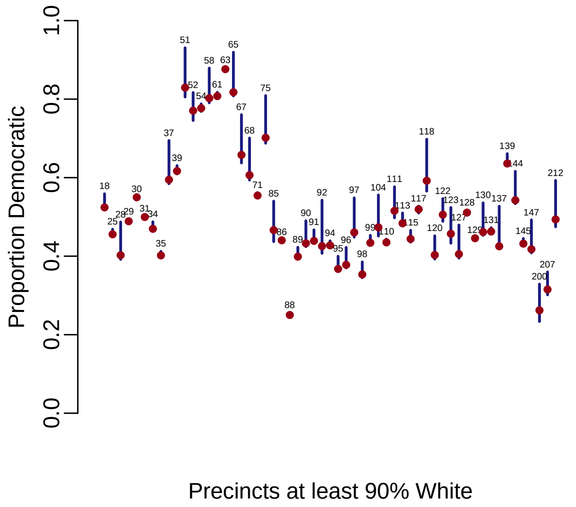
- Quantity of interest: proportion of row members in each column for each unit
- Observed row and column marginals determine upper and lower bounds
- Row thresholds implemented for *extreme case analysis*
- Output:

```
 $white.dem
     lower  upper
18  0.519  0.559
25  0.450  0.469
28  0.392  0.487
```

Method of bounds



Method of bounds



Ecological regression

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

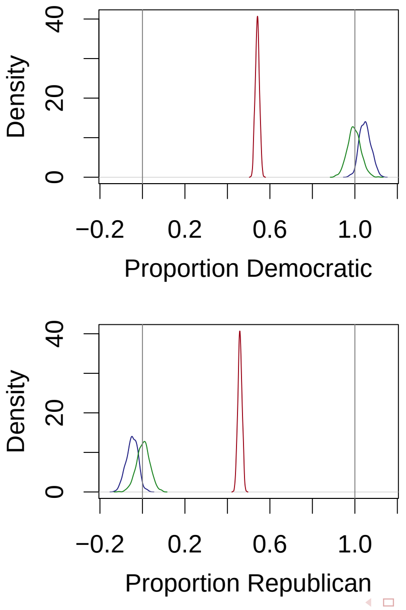
Ecological regression

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

Ecological regression

- Express data as proportions of row totals
- Regress each column on all row proportions (*C* regressions)
- Coefficients estimate cell proportions
- `eiPack`: freq. and Bayesian regression
- `lambda` functions calculate shares of a subset of columns – e.g. “among Blacks, Dem. share of 2-party registration”

Ecological regression



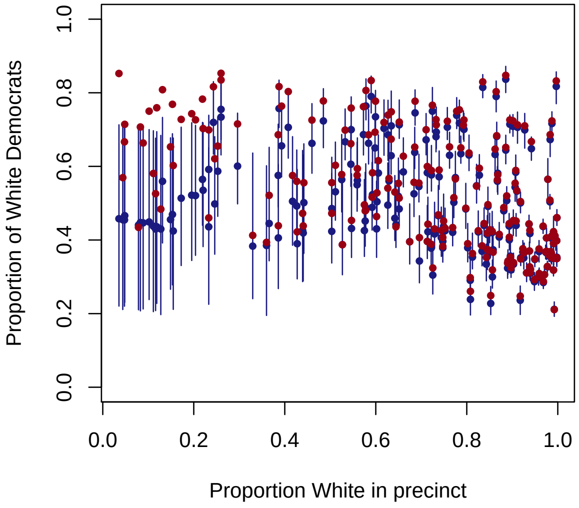
Multinomial-Dirichlet (MD) model

- Express data as counts
- Fit hierarchical Bayesian model
  - Level 1: column marginals $\sim$ *Multinomial*, $\perp\!\!\!\perp$ across units
  - Level 2: rows of cell fractions $\sim$ *Dirichlet*, $\perp\!\!\!\perp$ across rows and units
  - Level 3: Dirichlet parameters $\sim$ *Gamma*, i.i.d.

Multinomial-Dirichlet (MD) model

- Express data as counts
- Fit hierarchical Bayesian model
  - Level 1: column marginals $\sim$ *Multinomial*, $\perp\!\!\!\perp$ across units
  - Level 2: rows of cell fractions $\sim$ *Dirichlet*, $\perp\!\!\!\perp$ across rows and units
  - Level 3: Dirichlet parameters $\sim$ *Gamma*, i.i.d.
- `lambda` and `density.plot` functions

Multinomial-Dirichlet (MD) model



Data Management

- Reasonable-sized problems produce unreasonable amounts of data

Data Management

- 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

Data Management

- 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 1.3×10^8 parameter draws
- Draws occupy $\approx$ 1GB of RAM; probably not enough iterations

Data Management

- 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 1.3×10^8 parameter draws
- Draws occupy $\approx$ 1GB 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!