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

Olivia Lau
olau@fas.harvard.edu

Ryan T. Moore
rtmoore@fas.harvard.edu

Michael Kellermann
kellerm@fas.harvard.edu

Department of Government and Institute for Quantitative Social Science, Harvard University (Cambridge, Massachusetts)

Vienna, Austria
16 June 2006

What is ecological inference (EI)?

Goal: infer individual level behavior from aggregate units

- Units of analysis are contingency tables with observed row and column marginals.
- EI estimates unobserved internal cells for each unit.

Let

- $i = 1, \ldots, n$ index units
- $r = 1, \ldots, R$ index the row marginals for each unit i
- $c = 1, \ldots, C$ index the column marginals for each unit i

For $R = 3$ and $C = 3$, each unit looks like (red = unobserved):

	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

The data may also be defined as fractions. Let

- Cell fractions be $\beta_{rci} = \frac{N_{rci}}{N_{r\cdot i}}$
- Row marginals be $X_{ri} = \frac{N_{r\cdot i}}{N_i}$
- Column marginals be $T_{ci} = \frac{N_{\cdot ci}}{N_i}$

Ecological regression

Goodman (1953) observes the identities

$$T_{ci} = \sum_{r=1}^{R} \beta_{rci} X_{ri} \quad \text{for} \quad c = 1, \ldots, C \quad \text{and} \quad i = 1, \ldots, n$$

$$\sum_{c=1}^{C} \beta_{rci} = 1 \quad \text{for} \quad r = 1, \ldots, R \quad \text{and} \quad i = 1, \ldots, n$$

Define the population cell fractions β_{rc} such that $\sum_{c=1}^{C} \beta_{rc} = 1$ for every r. Assuming that $\beta_{rci} = \beta_{rc}$ for all i, estimating the regression equations

$$T_{ci} = \beta_{rc} X_{ri} + \varepsilon_{ci} \quad \text{for} \quad c = 1, \ldots, C$$

recovers the population parameters β_{rc} when the standard linear regression assumptions apply, including $E[\varepsilon_{ci}] = 0$ and $Var[\varepsilon_{ci}] = \sigma_c^2$ for all i.

Method of bounds

Duncan and Davis (1953) note that row and column marginals determine upper and lower bounds on the possible proportion of row members in each column.

- In a unit, let j be a given row and K be the set of all columns. K includes k', the column of interest, k'', the set of columns other than k' included, and $\tilde{k}$, the set of columns excluded.
- In the example below, consider the bounds on the proportion of Native American two-party registrants who are Democrats: j is Native American, k' is Democrat, k'' is Republican, and $\tilde{k}$ is No Party.
- The quantity of interest can be written

$$\frac{N_{jk'}}{N_{jk'} + \sum_{k \in k''} N_{jk}}$$

The lower and upper bounds on this proportion are given by

$$\text{Lower Bound:} \quad \frac{\max(0, N_j - \sum_{k \neq k'} N_k)}{\max(0, N_j - \sum_{k \neq k'} N_k) + \min(N_j, \sum_{k \in k''} N_k)}$$

$$\text{Upper Bound:} \quad \frac{\min(N_j, N_{k'})}{\min(N_j, N_{k'}) + \max(0, N_j - N_{k'} - \sum_{k \in \tilde{k}} N_k)}$$

- Bounds tend to become more informative as homogeneity increases within a unit; *extreme case analysis* centers on units satisfying row thresholds such as "60% Native American precincts" in the example below.

Multinomial-Dirichlet (MD) model

Rosen et al. (2001) assume that the observed vectors of column marginals are independent across units and the unobserved row vectors of cell fractions are independent across rows and units.

- Without covariate

$$(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) \overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri})$$

$$(\beta_{r1i}, \ldots, \beta_{rCi}) \overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}(\alpha_{r1}, \ldots, \alpha_{rC})$$

$$\alpha_{rc} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)$$

- With a covariate Z_i

$$(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) \overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri})$$

$$(\beta_{r1i}, \ldots, \beta_{rCi}) \overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}\left(d_r \exp(\gamma_{r1} + \delta_{r1} Z_i), \ldots, d_r \exp(\gamma_{r(C-1)} + \delta_{r(C-1)} Z_i),\ d_r\right)$$

$$d_r \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)$$

  - Priors on each γ_{rc} and δ_{rc} are improper uniform
  - Parameterization of the prior on each $(\beta_{r1i}, \ldots, \beta_{rCi})$ is linear with respect to the covariate Z_i in the log-odds ratio of the expected fractions

$$\log\left(\frac{E(\beta_{rci})}{E(\beta_{rCi})}\right) = \gamma_{rc} + \delta_{rc} Z_i$$

Input-output for higher-dimension data objects

- Since units are tables, and the MD model allows for inferences about the internal cells from the margins, each unit i has RC parameters, to yield a minimum of RCn unit level parameters, in addition to any population-level hyperparameters. For s simulations, the MD model yields a minimum of $RCns$ draws.
- In a state such as Ohio, for example, the MD model requires (11,000 precinct × 3 party × 4 group) parameter draws per simulation. For 1,000 simulations, this is 1.32×10^8 unit-level draws, which occupy nearly 1GB of RAM, in addition to population-level hyperparameters.
- Rather than storing unit-level parameters in the workspace, we provide the option of saving each chain as a `.tar.gz` file on disk:

```
ei.MD.bayes(..., ret.beta = "s")
```

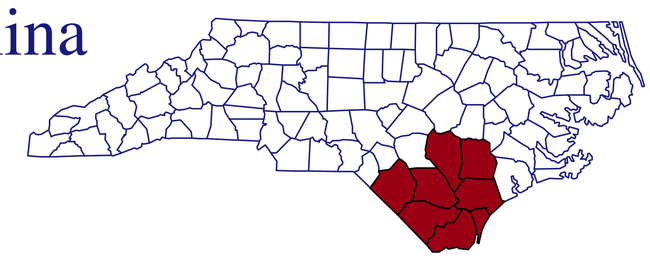
or discarding the unit-level draws entirely:

```
ei.MD.bayes(..., ret.beta = "d")
```

- To reconstruct the chains, users can select the row marginals, column marginals, and units of interest:

```
read.betas(rows = c("black", "white"),
   columns = c("dem"), units = 1:150, dir = getwd())
```

eiPack at work: Voter registration in southeastern North Carolina



The models implemented in `eiPack` share

- A common input syntax for each model of the form:

```
cbind(col1, ..., colC) ~ cbind(row1, ...,rowR)
```

- Functions to calculate shares for some subset of groups
- Print and summary functions for each model

To illustrate these methods, we use the `senc` data in `eiPack`:

- Data on party affiliation of Black, White, and Native American voters in eight counties (212 precincts) in SE North Carolina
- Interior cell counts allow for model benchmarking

	Democrat	Republican	No Party	Total
Black	*69999*	*3190*	*5057*	78246
White	*121941*	*97299*	*41295*	260535
Native American	*24480*	*1653*	*1376*	27509
Total	216420	102142	47728	366290

Ecological regression

- `ei.reg` performs the frequentist regression of column marginals on row marginals:

```
ei.reg(cbind(dem, rep, non) ~ cbind(black, white, natam),
   data = senc)
```

- `ei.reg.bayes` performs Bayesian regression of column marginals on row marginals using the improper prior $p(\beta, \sigma^2 | X) \propto \sigma^{-2}$, which is uniform on $(\beta, \log\sigma)$:

```
ei.reg.bayes(cbind(dem, rep, non) ~ cbind(black, white,
   natam), data = senc, sample = 1000, weights = NULL)
```

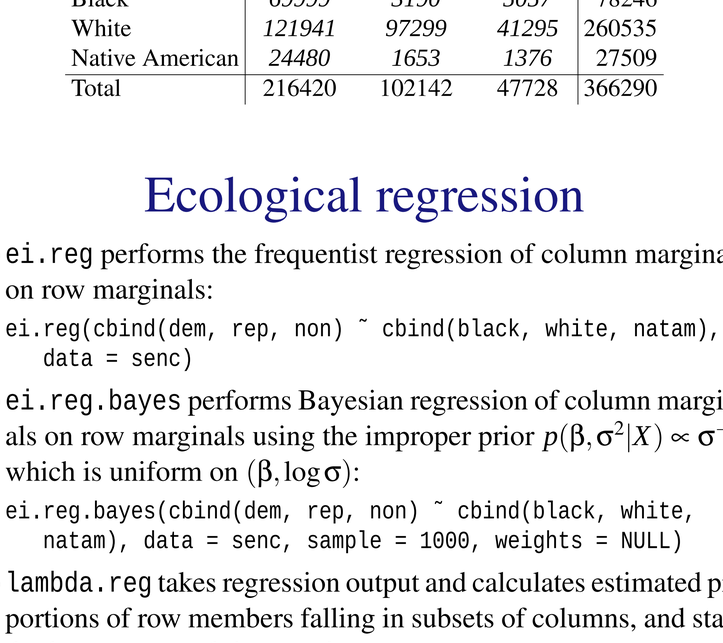
- `lambda.reg` takes regression output and calculates estimated proportions of row members falling in subsets of columns, and standard errors around those estimates.

```
lambda.reg(out.reg, columns = c("dem", "rep"))
```

- `density.plot` provides a graphical summary of `lambda` output:

```
density.plot(lreg)
```

- In the `senc` data, regression estimates of Black and Native American two-party registration shares exceed the possible limits.



Method of bounds

- `bounds` calculates unit-level bounds on the proportion of row members within a particular column:

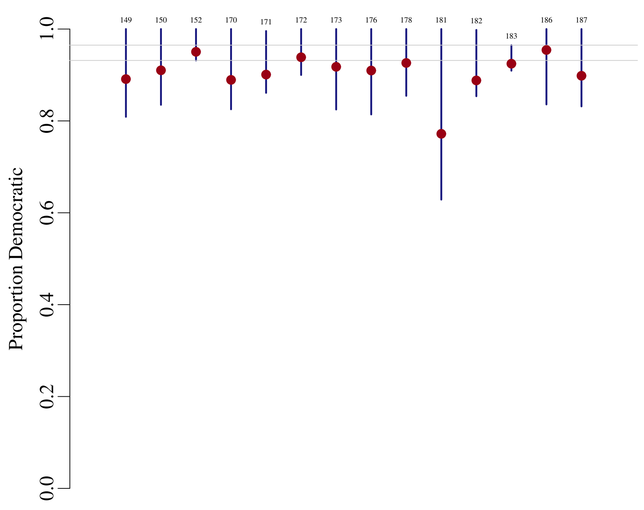
```
bounds(cbind(dem, rep, non) ~ cbind(black, white, natam),
   data = senc, rows = c("black", "natam"), column = "dem",
   excluded = NULL, threshold = 0.6, total = NULL)
```

The `excluded` argument allows calculation of bounds on proportions of a subset of columns. Using `excluded = "non"` above would calculate the bounds on the share of two-party registration.

- `plot` graphically displays the bounds for one row and highlights the intersection (if any) of the plotted bounds:

```
plot(out, row = "natam", column = "dem", intersection = T)
```

- In the `senc` data, the lower bound for Democratic registration among Native Americans is greater that 0.8 in most of the 60% Native American precincts. Values between 0.93 and 0.96 are possible in all of these precincts.



Multinomial-Dirichlet (MD) model

- `tuneMD` calibrates the tuning parameters used for Metropolis-Hastings sampling:

```
tuneMD(cbind(dem, rep, non) ~ cbind(black, white, natam),
   data = senc, ntunes = 10, sample = 1000,
   thin = 1000)
```

- `ei.MD.bayes` is the primary model-fitting function, calling `C` code to generate MCMC draws. Users can
  - return precinct parameters, discard, or save to disk
  - return parameters as `mcmc` objects or arrays
  - define a function to operate on each iteration within `C`

```
ei.MD.bayes(cbind(dem, rep, non) ~ cbind(black, white,
   natam), covariate = NULL, data = senc, lambda1 = 4,
   lambda2 = 2,  tune.list = tune.nocov, start.list = NULL,
   sample = 1000, thin = 5000, burnin = 1000000,
   ret.beta = 'r', ret.mcmc = TRUE, usrfun = NULL)
```

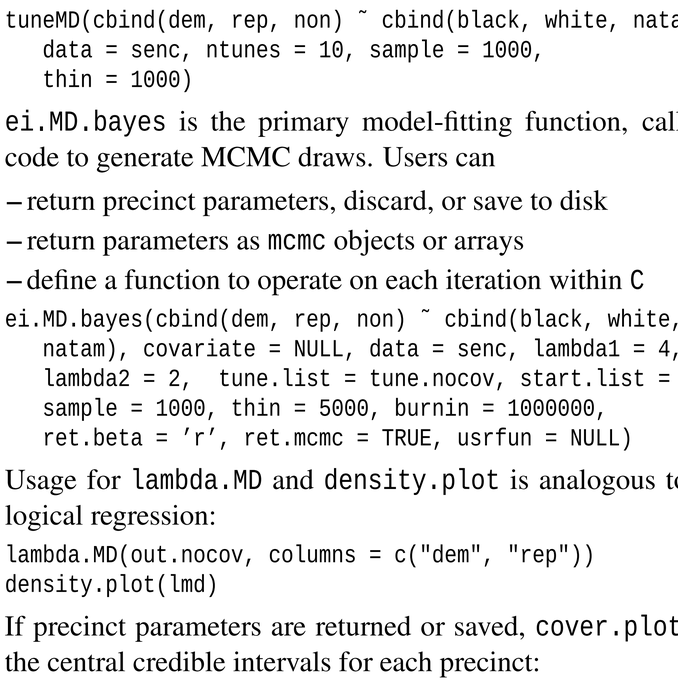
- Usage for `lambda.MD` and `density.plot` is analogous to ecological regression:

```
lambda.MD(out.nocov, columns = c("dem", "rep"))
density.plot(lmd)
```

- If precinct parameters are returned or saved, `cover.plot` plots the central credible intervals for each precinct:

```
cover.plot(out.nocov, row = "white", column = "dem")
```

- In the `senc` data, comparing the estimated precinct parameters to the true values shows that the percentage of White Democrats is underestimated in precincts with few White voters.



Acknowledgments

Thanks to Gary King, Kevin Quinn, and D. James Greiner for suggestions, Matt Cox and Bob Kinney for technical advice, and the Institute for Quantitative Social Science for travel funding.

References

Duncan, Otis Dudley and Beverley Davis. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18:665–666.

Goodman, Leo. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18:663–664.

Rosen, Ori, Wenxin Jiang, Gary King and Martin A. Tanner. 2001. "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case." *Statistica Neerlandica* 55(2):134–156.