`eiPack`: Tools for R $\times$ C Ecological Inference and Higher-Dimension Data Management

Olivia Lau Ryan T. Moore Michael Kellermann*

August 28, 2006

1 Introduction

Under certain conditions, ecological inference (EI) models allow researchers to infer individual-level behavior from aggregate data when individual-level data is unavailable. Table 1 shows a typical unit of ecological analysis: a contingency table with observed row and column marginals and unobserved interior cells.

	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

Table 1: A typical $R \times C$ unit in ecological inference; red quantities are typically unobserved.

Existing packages that implement EI methods, such as `eco` and `MCMCpack`, focus on 2×2 inference. `eiPack` offers methods for the more general case in which the ecological units are $R \times C$ tables.

2 Methods and Data in `eiPack`

The methods currently implemented in `eiPack` are the method of bounds (Duncan and Davis, 1953), ecological regression (Goodman, 1953), and the Multinomial-Dirichlet model (Rosen et al., 2001).

*`olau@fas.harvard.edu`, `rtmoore@fas.harvard.edu`, and `kellerm@fas.harvard.edu`. Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138

The functions that implement these models share several attributes. The ecological tables are defined using a common formula of the form `cbind(col1, ..., colC)` ~ `cbind(row1, ...,rowR)`. The row and column marginals can be expressed as either proportions or counts. Auxiliary functions renormalize the results for some subset of columns taken from the original ecological table, and appropriate `print`, `summary`, and `plot` functions conveniently summarize the model output.

In the following section, we demonstrate the features of `eiPack` using the (included) `senc` dataset, which contains individual-level party affiliation data for Black, White, and Native American voters in 8 counties in southeastern North Carolina. These counties include 212 precincts, which form the ecological units in this dataset. Because the data are observed at the individual level, the interior cell counts are known, allowing us to benchmark the estimates implied by each method.

2.1 Method of Bounds

The method of bounds (Duncan and Davis, 1953) uses the observed row and column marginals to calculate deterministic upper and lower bounds for functions of the interior cells of each ecological unit. As implemented in `eiPack`, it calculates for a specified column $k' \in k = \{1, \ldots, C\}$ the deterministic bounds on the proportion of individuals in each row who belong in that column. For each unit being considered, let j be the row of interest, k index columns, k' be the column of interest, k'' be the set of other columns considered, and $\tilde{k}$ be the set of columns excluded. For example, if we want 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 unit-level quantity of interest is

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

The lower and upper bounds on this quantity given by the observed marginals are, respectively:

$$\frac{\max(0, N_{ji} - \sum_{k \neq k'} N_{ki})}{\max(0, N_{ji} - \sum_{k \neq k'} N_{ki}) + \min(N_{ji}, \sum_{k \in k''} N_{ki})}$$

and

$$\frac{\min(N_{ji}, N_{k'i})}{\min(N_{ji}, N_{k'i}) + \max(0, N_{ji} - N_{k'i} - \sum_{k \in \tilde{k}} N_{ki})}$$

The method of bounds is not a statistical procedure in the traditional sense; the bounds implied by the row and column marginals are deterministic and there is no model of the data-generating process. One population-level quantity of interest calculated by `eiPack` is the interval defined by the intersection of the unit-level bounds (**?**). This interval, if it exists, represents the range of values that are consistent with the observed marginals in all of the ecological units.

Since the bounds become more informative as within-unit homogeneity increases, researchers and practitioners often restrict their attention to those units in which one group dominates. `eiPack`

allows users to set row thresholds to conduct this *extreme case analysis* (also known as *homogeneous precinct analysis* in the voting context). For example, suppose the user is interested in the proportion of two-party White registrants registered as Democrats in precincts that are at least 90% White. `eiPack` calculates the desired bounds:

```
> out <- bounds(cbind(dem, rep, non) ~ cbind(black, white, natam),
+   data = senc, rows = c("white"), column = "dem",
+   excluded = "non", threshold = 0.9, total = NULL)
```

These calculated bounds can then be represented graphically; in this example, there is no interval consistent with the bounds in each precinct:

```
> plot(out, row = "white", column = "dem")
```

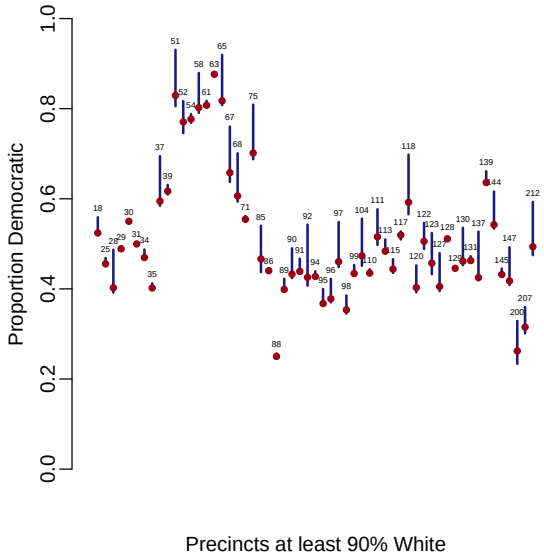

Figure 1: A plot of deterministic bounds.

2.2 Ecological Regression

In ecological regression (Goodman, 1953), observed row and column marginals are expressed as proportions and each column is regressed separately on the row proportions, thus performing C regressions. Regression coefficients then estimate the population internal cell proportions. For a given unit i, define

- X_{ri}, the proportion of individuals in row r,

- T_{ci}, the proportion of individuals in column c, and
- β_{rci}, the proportion of row r individuals in column c

The following identities hold:

$$T_{ci} \;=\; \sum_{r=1}^{R} \beta_{rci} X_{ri} \quad \text{and} \quad \sum_{c=1}^{C} \beta_{rci} \;=\; 1$$

Defining the population cell fractions β_{rc} such that $\sum_{c=1}^{C} \beta_{rc} = 1$ for every r, ecological regression assumes that $\beta_{rc} = \beta_{rci}$ for all i, and estimates the regression equations $T_{ci} = \beta_{rc} X_{ri} + \epsilon_{ci}$. Under the standard linear regression assumptions, including $E[\epsilon_{ci}] = 0$ and $Var[\epsilon_{ci}] = \sigma_c^2$ for all i, these regressions recover the population parameters β_{rc}. `eiPack` implements frequentist and Bayesian regression models (via `ei.reg` and `ei.reg.bayes`, respectively).

Output from ecological regression can be summarized numerically just as in `lm` or graphically using density plots. For the Bayesian model, densities represent functions of the posterior draws of the β_{rc}; for the frequentist model, densities reflect functions of regression point estimates and standard errors calculated using the δ-method.

We include functions to calculate estimates and standard errors of shares of a subset of columns in order to address questions such as, e.g., "among Blacks, what is the Democratic share of 2-party registration?"

```
> out.reg <- ei.reg(cbind(dem, rep, non) ~ cbind(black, white,
+   natam), data = senc)
> lreg <- lambda.reg(out.reg, columns = c("dem", "rep"))
> density.plot(lreg)
```

2.3 Multinomial-Dirichlet (MD) model

In the Multinomial-Dirichlet model (Rosen et al., 2001), the data is expressed as counts and a hierarchical Bayesian model is fit using a Metropolis-within-Gibbs algorithm implemented in `C`. Level 1 models the observed column marginals as Multinomial (and independent across units); Level 2 models the unobserved rows of cell fraction as Dirichlet (and independent across rows and units); Level 3 models the Dirichlet parameters as i.i.d. Gamma. More formally, without a covariate, the model is

$$\begin{aligned}
(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) &\overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri}) \\
(\beta_{r1i}, \ldots, \beta_{rCi}) &\overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}(\alpha_{r1}, \ldots, \alpha_{rC}) \\
\alpha_{rc} &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)
\end{aligned}$$

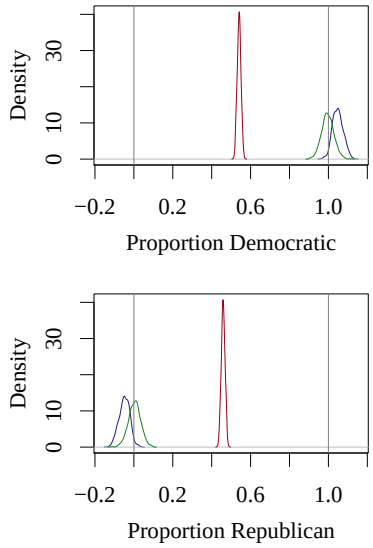

Figure 2: Density plots of Bayesian ecological regression output.

With a unit-level covariate Z_i in the second level, the model becomes

$$\begin{aligned}
(N_{\cdot 1i}, \ldots, N_{\cdot Ci}) &\overset{\perp\!\!\!\perp}{\sim} \text{Multinomial}(N_i, \sum_{r=1}^{R} \beta_{r1i} X_{ri}, \ldots, \sum_{r=1}^{R} \beta_{rCi} X_{ri}) \\
(\beta_{r1i}, \ldots, \beta_{rCi}) &\overset{\perp\!\!\!\perp}{\sim} \text{Dirichlet}\left(d_r \exp(\gamma_{rc} + \delta_{rc} Z_i), \ldots, \right. \\
&\qquad \left. d_r \exp(\gamma_{r(C-1)} + \delta_{r(C-1)} Z_i), d_r\right) \\
d_r &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(\lambda_1, \lambda_2)
\end{aligned}$$

Improper uniform priors are assumed for each γ_{rc} and δ_{rc}. The parameterization of the prior on each $(\beta_{r1i}, \ldots, \beta_{rCi})$ implies that the following log-odds ratio of expected fractions is linear with respect to the covariate Z_i:

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

Conducting an analysis using the MD model requires two steps. First, `tuneMD` calibrates the tuning parameters used for Metropolis-Hastings sampling:

```
> tune.nocov <- tuneMD(cbind(dem, rep, non) ~ cbind(black, white,
+   natam), data = senc, ntunes = 10, sample = 1000, thin = 1000)
```

Second, `ei.MD.bayes` fits the model by calling `C` code to generate MCMC draws:

```
> out.nocov <- ei.MD.bayes(cbind(dem, rep, non) ~ cbind(black, white,
+   natam), covariate = NULL, data = senc, lambda1 = 4, lambda2 = 2,
+   tune.list = tune.nocov, ...)
```

The output of this function can be returned as `mcmc` objects or arrays; in the former case, the standard diagnostic tools for `mcmc` objects can be applied directly. The MD implementation includes `lambda` and `density.plot` functions, usage for which is analogous to ecological regression:

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

If the precinct-level parameters are returned or saved, `cover.plot` plots the central credible intervals for each precinct. The segments represent the 95% central credible intervals and their medians for each unit (the true value for each precinct is the red dot, not included in the standard `cover.plot`).

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

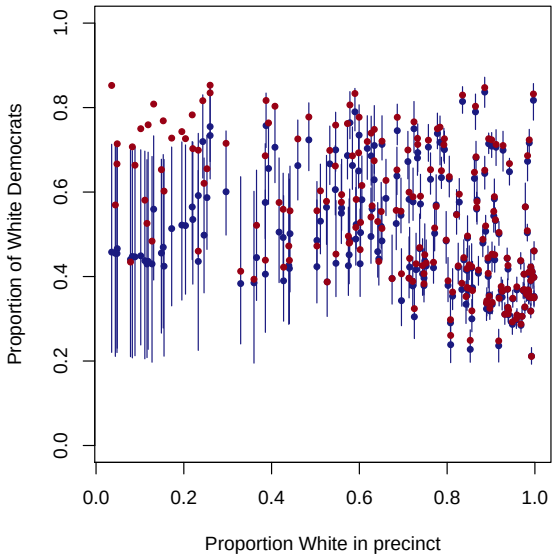

Figure 3: Coverage plot for MD model output.

3 Data Management

In the MD model, reasonable-sized problems produce unreasonable amounts of data. For example, a model for voting in Ohio includes 11000 precincts, 3 racial groups, and 4 parties. Implementing

1000 iterations yields about 130 million parameter draws. These draws occupy about 1GB of RAM, and this is almost certainly not enough iterations. We provide a few options to users in order to make this model tractable for large EI problems.

The unit-level parameters present the most significant data management problem. Rather than storing unit-level parameters in the workspace, users can save each chain as a `.tar.gz` file on disk using the option `ei.MD.bayes(..., ret.beta = "s")`, or discard the unit-level draws entirely using `ei.MD.bayes(..., ret.beta = "d")`. To reconstruct the chains, users can select the row marginals, column marginals, and units of interest, without reconstructing the entire matrix of unit-level draws:

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

If users are interested in some function of the unit-level parameters, the implementation of the MD model allows them to define a function in `R` that will be called from within the `C` sampling algorithm, in which case the unit-level parameters need not be saved for post-processing.

4 Conclusion

`eiPack` was developed with the support of the Institute for Quantitative Social Science at Harvard University. Thanks to Gary King, Kevin Quinn, and D. James Greiner for suggestions and Matt Cox and Bob Kinney for technical advice. For further information, see `http://www.people.fas.harvard.edu/~olau/software/eiPack.html`.

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* 18:663–666.

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. http://gking.harvard.edu/files/abs/rosen-abs.shtml.