Journal of Statistical Software

May 2011, Volume 42, Issue 5. *http://www.jstatsoft.org/*

eco: R Package for Ecological Inference in 2×2 Tables

Kosuke Imai
Princeton University

Ying Lu
New York University

Aaron Strauss
The Mellman Group

Abstract

eco is a publicly available R package that implements the Bayesian and likelihood methods proposed in Imai, Lu, and Strauss (2008b) for ecological inference in 2×2 tables as well as the method of bounds introduced by (Duncan and Davis 1953). The package fits both parametric and nonparametric models using either the Expectation-Maximization algorithms (for likelihood models) or the Markov chain Monte Carlo algorithms (for Bayesian models). For all models, the individual-level data can be directly incorporated into the estimation whenever such data are available. Along with in-sample and out-of-sample predictions, the package also provides a functionality which allows one to quantify the effect of data aggregation on parameter estimation and hypothesis testing under the parametric likelihood models. This paper illustrates the usage of **eco** with several real data examples that are also part of the package.

Keywords: aggregate data, Bayesian inference, bounds, likelihood inference, missing data, missing information.

1. Introduction

This paper illustrates how to use **eco**, a publicly available R package (R Development Core Team 2011), to implement the Bayesian and likelihood methods proposed in Imai, Lu, and Strauss (2008b) for ecological inference in 2×2 tables. The package also implements the method of bounds introduced by (Duncan and Davis 1953) for the analysis of general $R \times C$ tables. Ecological inference refers to the “inferences about individual behavior drawn from data about aggregates” (Freedman 1999, p. 4027). Such cross-level inferences are frequently conducted in epidemiology, political science, and sociology when only aggregate-level data are available (e.g., Greenland and Robins 1994; Achen and Shively 1995; King 1997; King, Rosen, and Tanner 2004). Yet, the difficulty of ecological inference is that the observed correlation at the aggregate level does not necessarily imply the same individual-level relationship. Using an example of literacy rates across different racial groups, Robinson (1950) powerfully illustrated this "ecological fallacy."

Since Robinson's seminal article, various methods have been proposed for ecological inference. Duncan and Davis (1953) showed how to derive the bounds on unknown quantities of interest from aggregate data. We generalize and implement this method for $R \times C$ tables in **eco**. Goodman (1953, 1959) developed the regression-based approach to ecological inference, which gained popularity among applied researchers in the next several decades (e.g., Freedman, Klein, Sacks, Smyth, and Everett 1991; Achen and Shively 1995; Gelman, Park, Ansolabehere, Price, and Minnite 2001) – this approach can be easily implemented via `lm()` command in R, and hence is not implemented in **eco**. Recent years have witnessed a growing number of new methods based on modern statistical techniques (e.g., King, Rosen, and Tanner 1999; Rosen, Jiang, King, and Tanner 2001; Imai and King 2004; Judge, Miller, and Cho 2004; Wakefield 2004) – some of these methods are available in R via **Zelig** (Imai, King, and Lau 2008a, 2009) and **MCMCpack** (Martin, Quinn, and Park 2011). At the same time, the appropriateness of the assumptions underlying some of these models is often disputed (e.g., Freedman, Ostland, Roberts, and Klein 1998; Cho 1998; King 1999; Cho and Gaines 2004).

In a recent paper, Imai, Lu, and Strauss (2008b) have proposed a theoretical framework for Bayesian and likelihood inference in 2 × 2 ecological tables. The framework is based on the theory of coarse data which is originally developed by Heitjan and Rubin (1991). We show that ecological inference can be formulated as a coarse data problem and that Bayesian and likelihood inference can be conducted within this coarse data framework. The main advantage of this framework is that it clarifies the modeling assumptions necessary for Bayesian and likelihood ecological inference. In particular, Imai, Lu, and Strauss (2008b) show that the ecological inference problem can be decomposed into three key factors: *distributional effects* which address the possible misspecification of parametric modeling assumptions about the unknown distribution of missing data, *contextual effects* which represent the possible correlation between missing data and observed variables, and *aggregation effects* which are directly related to the loss of information caused by data aggregation.

Furthermore, while Imai, Lu, and Strauss (2008b) propose statistical methods to address distributional and contextual effects, they also show that aggregation effects cannot be overcome by statistical adjustments. Instead, they demonstrate how to formally quantify the effect of data aggregation on parameter estimation and hypothesis testing. In this paper, we illustrate how to implement these proposed methods using an R package **eco**, which is freely available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=eco`. In Section 2, we start our discussion by describing and generalizing the method of bounds (Duncan and Davis 1953). We then outline the parametric and nonparametric models proposed by Imai, Lu, and Strauss (2008b), which respect the constraints imposed by the bounds. We also briefly review the method to formally quantify the aggregation effects. In Section 3, we illustrate the use of the **eco** package through the analysis of several real data examples.

2. The methodology

We use racial voting as a concrete example to describe ecological inference in 2×2 tables. Although this is a prominent example in political science, other problems in different disciplines may fit into the same framework. Table 1 presents a 2 × 2 ecological table of racial voting

	black voters	white voters	
Voted	W_{i1}	W_{i2}	Y_i
Not Voted	$1 - W_{i1}$	$1 - W_{i2}$	$1 - Y_i$
	X_i	$1 - X_i$	

Table 1: 2×2 ecological table for the racial voting example. X_i, Y_i, W_{i1}, and W_{i2} are proportions, and hence lie between 0 and 1. The unit of observation is typically a geographical unit and is denoted by i.

example. Suppose that from the census data we observe the fraction of registered white and black voters for each county, i.e., X_i and $1 - X_i$. The overall turnout rate Y_i can be obtained from the election returns for each county. However, the proportions of black and white voters who turned out, W_{i1} and W_{i2} respectively, are unknown.

The **eco** package implements the method of bounds and fits both parametric and nonparametric methods for such ecological data. The estimation is based on the Expectation-Maximization (EM) algorithms (Dempster, Laird, and Rubin 1977) for likelihood models and on the Markov chain Monte Carlo (MCMC) algorithms for Bayesian models. These algorithms are described in Imai, Lu, and Strauss (2008b). Below, we briefly summarize each method and model. Note that although we do not discuss the issue of convergence of Markov chains in detail, users of **eco** should follow standard advice and conduct convergence diagnostics, perhaps using the **coda** package (Plummer, Best, Cowles, and Vines 2006).

2.1. The method of bounds

Suppose that in a simple random sample of size n from a population, we observe the margins of Table 1 for each county i. The method of bounds is based on the following deterministic relationship,

$$Y_i \quad = \quad W_{i1} X_i + W_{i2}(1 - X_i), \quad \text{for} \quad i = 1, 2, \ldots, n \tag{1}$$

where $X_i, Y_i, W_{i1}, W_{i2} \in [0, 1]$. When Y_i is equal to either 0 or 1, W_{i1} and W_{i2} are completely known. If $X_i = 1$, then $W_{i1} = Y_i$ but W_{i2} does not exist. Similarly, if $X_i = 0$, then $W_{i2} = Y_i$ but W_{i1} does not exist. King (1997) called Equation 1 a tomography line. For every i, this tomography line defines a *deterministic* relationship between the missing data, $W_i = (W_{i1}, W_{i2})$ and the observed data, (Y_i, X_i). Duncan and Davis (1953) first recognized that with Equation 1, one can narrow the original bound of $[0, 1]$ for W_i to the following intervals,

$$W_{i1} \quad \in \quad \left[\max\left(0, \frac{X_i + Y_i - 1}{X_i}\right), \ \min\left(1, \frac{Y_i}{X_i}\right)\right], \tag{2}$$

$$W_{i2} \quad \in \quad \left[\max\left(0, \frac{Y_i - X_i}{1 - X_i}\right), \ \min\left(1, \frac{Y_i}{1 - X_i}\right)\right]. \tag{3}$$

Given these bounds for each i (e.g., a county), the analysis of larger units (e.g, a state) can be carried out by simply aggregating the upper and lower bounds with appropriate weights; $N_i X_i$ and $N_i(1 - X_i)$ for W_{i1} and W_{i2}, respectively, where N_i is the total number of voters in county i. When the resulting bounds are sufficiently narrow, researchers can draw reasonably informative conclusions about (in-sample) missing cells.

The bounds in Equations 2 and 3 can be easily generalized to the situation of $R \times C$ ecological tables where $R \geq 2$ and $C \geq 2$. These generalized bounds can also be computed via the **eco** package. Suppose that we denote the observed row and column margins by Y_{ir} and X_{ic} for $c = 1, \ldots, C$, and $r = 1, \ldots, R$ where $\sum_{c=1}^{C} X_i = 1$ and $\sum_{r=1}^{R} Y_{ir} = 1$ for all i. Then, the unobserved proportion in the rth row and cth column can be defined as W_{irc}. The results in the statistical literature on contingency tables (Bonferroni 1936; Fréchet 1940; Hoeffding 1940) imply that the bounds are given by,

$$\max\left\{0, \ \frac{X_{ic} + Y_{ir} - 1}{X_{ic}}\right\} \ \leq \ W_{irc} \ \leq \ \min\left\{1, \ \frac{Y_{ir}}{X_{ic}}\right\}. \tag{4}$$

Although applied researchers often find the bounds too wide for their purposes, the method of bounds shows the identifying power of the data without any statistical assumption. That is, the bounds imply the exact degree to which the data are informative about W_i. For this reason, statistical analysis that does not incorporate this deterministic relationship is likely to be sensitive to modeling assumptions. The **eco** package computes the bounds for the general $R \times C$ case as well as for the 2×2 case (see Section 3.1).

2.2. Parametric models

Next, we describe the parametric models implemented by the package **eco**. Imai, Lu, and Strauss (2008b) propose the parametric models based on three assumptions. The simplest parametric model is based on the assumption of *coarsened at random* (CAR) and defined by,

$$W_i^* \mid \mu, \Sigma \ \overset{\text{i.i.d.}}{\sim} \ \mathcal{N}(\mu, \Sigma),$$

where $W_i^* = (\text{logit}(W_{i1}), \text{logit}(W_{i2}))$, μ represents a 2×1 vector of population means, and Σ is a 2×2 positive-definite variance matrix. The model, which is similar to the ones proposed by King (1997) and Wakefield (2004), assumes the independence between W_i and X_i and thus no contextual effect. The maximum likelihood (ML) estimates of μ and Σ can be computed via the EM algorithm. The Bayesian analysis, on the other hand, is based on the following conjugate prior distribution,

$$\mu \mid \Sigma \ \sim \ \mathcal{N}(\mu_0, \Sigma/\tau_0^2), \quad \text{and} \quad \Sigma \ \sim \ \text{InvWish}\,(\nu_0, \ S_0^{-1}),$$

where μ_0 denotes a (2×1) vector of the prior mean, τ_0 is a scalar, ν_0 is the prior degrees of freedom parameter, and S_0 represents a (2×2) positive definite prior scale matrix. The posterior inference can then be conducted by the MCMC algorithm.

The assumption of no contextual effect under the CAR model is unrealistic in many situations. Imai, Lu, and Strauss (2008b) consider two modeling strategies in order to relax this assumption. First, one may collect additional covariates Z_i and assume no contextual effect after conditioning on Z_i. Such a strategy is often employed in the literature (e.g., King 1997; King *et al.* 1999). Thus, we can extend our CAR model to the following CCAR (*conditionally coarsened at random*) model,

$$W_i^* \mid \beta, \Sigma, Z_i \ \overset{\text{indep.}}{\sim} \ \mathcal{N}(Z_i^\top \beta, \Sigma),$$

where β represents a $(k \times 1)$ vector of coefficients, and Z_i is a $(k \times 2)$ matrix of covariates. The ML estimates of β and Σ can be obtained by the *ECM* algorithm and the Bayesian analysis can be conducted by placing the semi-conjugate prior distribution,

$$\beta \mid \Sigma \;\sim\; \mathcal{N}(\beta_0, A_0^{-1}), \quad \text{and} \quad \Sigma \;\sim\; \text{InvWish}\,(\nu_0,\; S_0^{-1}),$$

where β_0 is a $(k \times 1)$ vector of prior means, and A_0 is a $(k \times 2)$ matrix of prior precision. The MCMC algorithm can be used to sample from the posterior distribution.

Finally, Imai, Lu, and Strauss (2008b) suggest an alternative approach where the contextual effects are directly modeled without additional covariates. This NCAR (*not coarsened at random*) model is formally defined as,

$$(W_i^*, X_i^*) \mid \eta, \Phi \;\overset{\text{i.i.d.}}{\sim}\; \mathcal{N}(\eta, \Phi),$$

where $X_i^* = \text{logit} X_i$, η is a (3×1) vector of population means, and Φ is a (3×3) matrix of covariance. The ML estimates of η and Φ can be obtained by the EM algorithm, whereas the Bayesian analysis of the NCAR model can be conducted in the same way as under the CAR model except that the NCAR model relies upon the trivariate normal distribution rather than the bivariate normal distribution. An advantage of the NCAR model over the CCAR model is that the former does not require the availability of additional covariates to model the contextual effects. Indeed, under the NCAR model, one needs not specify the conditional expectation function of W_i^* given Z_i. The **eco** package implements all three models within the Bayesian or maximum likelihood framework (see Section 3.2).

2.3. Nonparametric models

To address the *distributional effects*, Imai, Lu, and Strauss (2008b) propose Bayesian nonparametric models based on a Dirichlet process prior (e.g., Dey, Müller, and Sinha 1998). This model generalizes the CAR and NCAR parametric models to the case of the unknown distribution of W_i^*. For the CAR assumption, the Bayesian nonparametric model can be written as follows,

$$\begin{aligned} W_i^* \mid \mu_i, \Sigma_i &\;\sim\; \mathcal{N}\,(\mu_i,\; \Sigma_i), \\ \mu_i, \Sigma_i \mid G &\;\sim\; G, \\ G \mid \alpha &\;\sim\; \mathcal{D}(G_0,\; \alpha), \\ \alpha &\;\sim\; \text{Gamma}(a_0,\; b_0), \end{aligned}$$

where $\mathcal{D}(G_0, \alpha)$ represents the Dirichlet process prior with the base prior distribution G_0 and the scalar concentration parameter α. Under G_0, (μ_i, Σ_i) is distributed as,

$$\mu_i \mid \Sigma_i \;\sim\; \mathcal{N}\left(\mu_0,\; \frac{\Sigma_i}{\tau_0^2}\right), \quad \text{and} \quad \Sigma_i \;\sim\; \text{InvWish}\,(\nu_0,\; S_0^{-1}).$$

The MCMC algorithm summarized in Imai, Lu, and Strauss (2008b) can be used to sample from the posterior distribution of this model. Furthermore, the nonparametric NCAR model can be formulated in the same manner by using the parametric NCAR model as the base model and specifying the Dirichlet process prior distribution on (η_i, Φ_i), where η and Φ are now indexed by i. The package **eco** implements this Bayesian nonparametric model under both the CAR and NCAR assumptions (see Section 3.4).

2.4. Formal assessment of aggregation effects

The fourth method we implement via the **eco** package is the formal assessment of aggregation effects under the parametric models. Imai, Lu, and Strauss (2008b) propose to measure the effect of data aggregation on parameter estimation and hypothesis testing by calculating the fraction of missing information. The idea is to quantify the amount of information the observed aggregate-level data provide in comparison with the information one would obtain if the individual-level data were available. In the context of parameter estimation, the fraction of missing information is defined as,

$$F_\theta \equiv \text{diag}\left(I - \mathcal{I}_{obs}(\hat{\theta})\mathcal{I}_{com}(\hat{\theta})^{-1}\right), \tag{5}$$

where $\mathcal{I}_{obs}$ is the observed Fisher information matrix and $\mathcal{I}_{com}$ represents the expected information matrix based on the complete-data log-likelihood function. Then, each element of the vector F_θ represents the fraction of missing information for each parameter. In the **eco** package, we use the Supplemented EM (SEM) algorithm (Meng and Rubin 1991) and compute the fraction of missing information for the parametric CAR and NCAR models (see Section 3.3).

For the hypothesis testing, we follow the approach proposed by Kong, Meng, and Nicolae (2008) and compute the fraction of missing information against the null hypothesis $H_0 : \theta = \theta_0$, which is defined by,

$$F_H \equiv 1 - \frac{l_{obs}(\hat{\theta} \mid Y, X) - l_{obs}(\theta_0 \mid Y, X)}{E[\, l_{com}(\hat{\theta} \mid W, X) - l_{com}(\theta_0 \mid W, X) \mid Y, X; \hat{\theta}]}, \tag{6}$$

where $l_{obs}(\theta \mid Y, X)$ and $l_{com}(\theta \mid W, X)$ represent the observed-data log-likelihood and the complete-data log-likelihood functions, respectively. Moreover, $\hat{\theta}$ is the ML estimate of θ and the expectation is taken over the conditional distribution of W given (Y, X). Then, F_H equals one minus the logarithm of *the observed likelihood ratio statistic* divided by the logarithm of *the expected likelihood ratio statistic*. In the **eco** package, we use the SEM algorithm and compute F_H with the null hypothesis of the equal marginal means, i.e., $H_0 : E(W_1) = E(W_2)$, under the parametric CAR and NCAR models (see Section 3.3).

2.5. Additional individual-level data

When bounds are not informative, ecological inference is difficult. The parametric inference will be sensitive to modeling assumptions, and the nonparametric model will not be able to recover the underlying distribution. Therefore, incorporating individual-level data may be helpful whenever such additional information is available. For example, one might conduct a survey in randomly selected counties to obtain such information. Sometimes, a small scale survey can be conducted to get rough estimates of W_i for some counties, and incorporating such auxiliary information can also be helpful (Wakefield 2004). In the **eco** package, it is straightforward to incorporate such information into the estimation of both parametric and nonparametric models (see Section 3.5).

3. Illustrative examples

In this section, we illustrate how to implement the methods described in Section 2 via the package **eco** using some example data sets which also is a part of the package. The detailed references for the commands and data sets we use appear in the help files of the **eco** package.

3.1. Computing the bounds

We first consider the computation of the bounds described in Section 2.1 using the function `ecoBD()`. We illustrate the use of this function with the voter registration data from 275 counties of four Southern states in the United States: Florida, Louisiana, North Carolina, and South Carolina. The data set is taken from King (1997) and is available as `reg` as a part of the **eco** package. To load this data set, type at the R prompt (after loading the package via the `library("eco")` command),

```
R> library("eco")
R> data("reg")
```

which stores the data frame as `reg` in the workspace. The data set can be summarized as,

```
R> summary(reg)
```

```
       X                  Y                  N                  W1
 Min.   :0.00826   Min.   :0.297   Min.   :  1800   Min.   :0.000
 1st Qu.:0.13061   1st Qu.:0.678   1st Qu.:  9000   1st Qu.:0.459
 Median :0.24286   Median :0.783   Median : 15500   Median :0.571
 Mean   :0.25725   Mean   :0.777   Mean   : 32448   Mean   :0.562
 3rd Qu.:0.37143   3rd Qu.:0.910   3rd Qu.: 31350   3rd Qu.:0.692
 Max.   :0.73899   Max.   :1.000   Max.   :613000   Max.   :1.000
       W2
 Min.   :0.321
 1st Qu.:0.776
 Median :0.888
 Mean   :0.855
 3rd Qu.:1.000
 Max.   :1.000
```

where `X` is the fraction of black voters in each county, `Y` represents the fraction of registered voters, and `N` is the total number of voters in each county. In this data set, the registration rates are observed separately for blacks and whites, which are given by `W1` and `W2`, respectively.

To compute the bounds using the `reg` data set, we simply use the following syntax,

```
R> res.BD <- ecoBD(Y ~ X, data = reg)
R> print(resBD)

Call:
ecoBD(formula = Y ~ X, data = reg)
```

```
Aggregate Lower Bounds (Proportions):
       X       not X
Y      0.3426  0.7047
not Y  0.0154  0.0729

Aggregate Upper Bounds (Proportions):
       X      not X
Y      0.985  0.927
not Y  0.657  0.295
```

which prints out the aggregate lower and upper bounds. For example, the registration rate for blacks lies between 0.34 and 0.99, while that for whites is between 0.70 and 0.93. The actual registration rates for blacks and whites (which are usually unknown but in this case can be estimated using the sample means of `W1` and `W2` in the dataset `reg`) are 0.56 and 0.86, respectively.

The county-level bounds are also stored in the output object from `ecoBD()`. For example, the bound for the first county can be obtained by the following commands,

```
R> res.BD$Wmin[1, , ]

             X    not X
Y     0.545455 0.850498
not Y 0.000000 0.000000

R> res.BD$Wmax[1, , ]

             X    not X
Y     1.000000 1.000000
not Y 0.454545 0.149502
```

It is also possible to incorporate the information about the total number of eligible voters (`N` in the data set). The following commands accomplish this,

```
R> res.BD1 <- ecoBD(Y ~ X, N = N, data = reg)
R> print(res.BD1)

Call:
ecoBD(formula = Y ~ X, data = reg, N = N)

Aggregate Lower Bounds (Proportions):
       X       not X
Y      0.2168  0.7055
not Y  0.0244  0.0792

Aggregate Upper Bounds (Proportions):
       X       not X
Y      0.976  0.921
not Y  0.783  0.294

Aggregate Lower Bounds (Counts):
       X        not X
Y       427500  4904400
not Y    48200   550500

Aggregate Upper Bounds (Counts):
       X        not X
Y      1923800  6400700
not Y  1544500  2046800
```

The county-level bounds can be obtained from the output object. They are stored as `Nmin` and `Nmax`.

Finally, `ecoBD()` also computes the bounds for $R \times C$ ecological tables. The syntax is very similar to the 2×2 case. For example, `ecoBD(cbind(Y1, Y2, Y3) ~ X1 + X2 + X3 + X4, data = data)` specifies 3×4 ecological tables.

3.2. Fitting the parametric models

In this section, we illustrate how to use the **eco** package to fit the parametric ecological inference models. First, we review the maximum likelihood (ML) estimation of the parametric models described in Section 2.2 using the function `ecoML()`. We then demonstrate the fitting of the Bayesian parametric model using the `eco()` function. The dataset used to illustrate these functions is the race and literacy dataset first collected by Robinson (1950) on the state level, and then refined to the county level by King (1997). For 1,040 counties, the marginal percentage of blacks (`X`), the marginal literacy rate (`Y`), and the population (`N`) is available, as well as the true cell-level values for literacy among blacks (`W1`) and whites (`W2`). As before, type the following in an `R` prompt to view summary statistics of the dataset,

```
R> data("census")
R> summary(census)
```

```
       Y                X                N                  W1
 Min.   :0.4055   Min.   :0.0508   Min.   :    798   Min.   :0.2012
 1st Qu.:0.7790   1st Qu.:0.1412   1st Qu.:   9900   1st Qu.:0.6251
 Median :0.8401   Median :0.3108   Median :  14428   Median :0.6897
 Mean   :0.8258   Mean   :0.3377   Mean   :  21710   Mean   :0.6845
 3rd Qu.:0.8912   3rd Qu.:0.4939   3rd Qu.:  20940   3rd Qu.:0.7513
 Max.   :0.9908   Max.   :0.9393   Max.   :1261132   Max.   :0.9665
       W2
 Min.   :0.5563
 1st Qu.:0.8936
 Median :0.9302
 Mean   :0.9189
 3rd Qu.:0.9599
 Max.   :0.9940
```

The maximum likelihood estimation. We first demonstrate the ML estimation of the CAR model, which assumes no contextual effect, via the EM algorithm. Using the default values provided by the function (see the help file of `ecoML()` in the **eco** package), with the exception of suppressing the output, the following command fits the model and stores its output as `res.ML`,

```
R> res.ML <- ecoML(Y ~ X, data = census, verbose = FALSE)
```

As this fitting process via the EM algorithm may take a long time on some computers, setting `verbose` to the value of `TRUE` (its default value) tracks the progress of the program,

```
R> res.ML <- ecoML(Y ~ X, data = census, verbose = TRUE)

OPTIONS (flag: 4) Ncar: No; Fixed Rho: No; SEM: First run
cycle 1/1000: 0.000 0.000 1.000 1.000 0.000
cycle 2/1000: 1.223 1.886 0.698 1.057 0.010
cycle 3/1000: 1.234 2.151 0.602 0.877 -0.051 Prev LL: -1299.52
cycle 4/1000: 1.161 2.249 0.560 0.823 -0.070 Prev LL: -1230.91
cycle 5/1000: 1.090 2.320 0.524 0.807 -0.070 Prev LL: -1208.75
[output truncated for presentation purposes]
cycle 349/1000: 0.654 2.785 0.236 0.916 0.271 Prev LL: -1126.77
Final Theta: 0.654 2.785 0.236 0.916 0.271 Final LL: -1126.77
```

The `OPTIONS` output line is for verification purposes, and informs the user that there are no contextual effect to be modeled, the correlation parameter (ρ) is not fixed by the user, and that the fraction of missing information will be calculated via the SEM algorithm. The next lines of output display the iteration number of the EM loop, the parameter values for that iteration, and the observed log-likelihood for the previous set of parameter values. As showed by the `cycle 1` output line, the default starting parameter values for CAR are $\mu_0 = 1$, $\mu_2 = 0$, $\sigma_1^2 = 0$, $\sigma_1^2 = 1$, $\rho = 0$.

A few remarks about the convergence are worth mentioning although we refer readers to the relevant literature for the details (e.g., McLaughlan and Krishnan 1997). When the absolute value difference between each of the parameter values from the previous iteration falls below the convergence threshold, `epsilon`, the algorithm stops. The last line of the output displays the final, converged parameter estimates. The default value for `epsilon` is 10^{-10}; this threshold can be changed by the user. The number of maximum iterations to cycle through before halting (`maxit`) can also be adjusted; the default value for `maxit` is $1,000$. The failure to set these inputs to appropriate values may result in inaccurate estimates.

Once the EM algorithm is completed, the SEM algorithm will begin automatically (as long as the `SEM` parameter is not set to `FALSE`). See Section 3.3 for details on executing the SEM algorithm. If the fraction of missing information is not desired, the user should set `SEM` to `FALSE` in the interest of computational time.

Summary statistics for the fitted model can be displayed simply by using the `summary()` function (a shorter summary is available via the `print()` function),

```
R> summary(res.ML)
```

```
Call:  ecoML(formula = Y ~ X, data = census, verbose = TRUE)

*** Parameter Estimates ***

Original Model Parameters:
                  mu1    mu2  sigma1 sigma2   rho
ML est.       0.65354 2.7847 0.23574 0.9159 0.271
std. err.     0.03259 0.0644 0.02029 0.1026 0.093
frac. missing 0.62566 0.5669 0.66159 0.6429 0.772

*** Insample Predictions ***

Unweighted:
      mean std.dev   2.5 % 97.5 %
W1 0.65007 0.07564 0.48492  0.802
W2 0.91973 0.05908 0.79422  0.986

Weighted:
      mean std.dev   2.5 % 97.5 %
W1 0.64645 0.07891 0.49178  0.801
W2 0.92407 0.06796 0.79087  1.057


Log-likelihood: -1126.773
Number of Observations: 1040
Number of EM iterations: 350
Number of SEM iterations: 95
Convergence threshold for EM: 1e-10
```

where the "weighted" insample predictions are computed based on the weights proportional to the group-specific population size while assuming the overall population size is the same across counties (when the information about overall population size is available, this information can be used via the option `N` as shown in the next example below).

Under the CAR assumption, the point estimate for the unweighted, mean black literacy rate is 66% and for the unweighted, mean white literacy, 92%. The estimated unweighted standard deviations for the county-level black and white literacy rates are 7.6% and 5.9% respectively. The correlation between logit-transformed county literacy rates is estimated to be 0.271 (`rho`). In addition to the aggregate-level in-sample predictions, county-level estimates are available in the $n \times 2$ matrix `res.ML$W`. Figure 1 shows that these country-level predictions are farily close to their observed values. To display all available elements of the output object returned by the `ecoML()` function, one can use the `names()` command.

From the output we can see that on the average the insample predictions of W_1 and W_2 are close to their observed values. Compared with the observed means of 0.6845 and 0.9189, the means of the predicted values are 0.6501 and 0.9197, respectively.

Bayesian estimation. Next, we illustrate how to fit the Bayesian parametric models via the MCMC algorithms using **eco**. Here, we use the NCAR model, which unlike the CAR

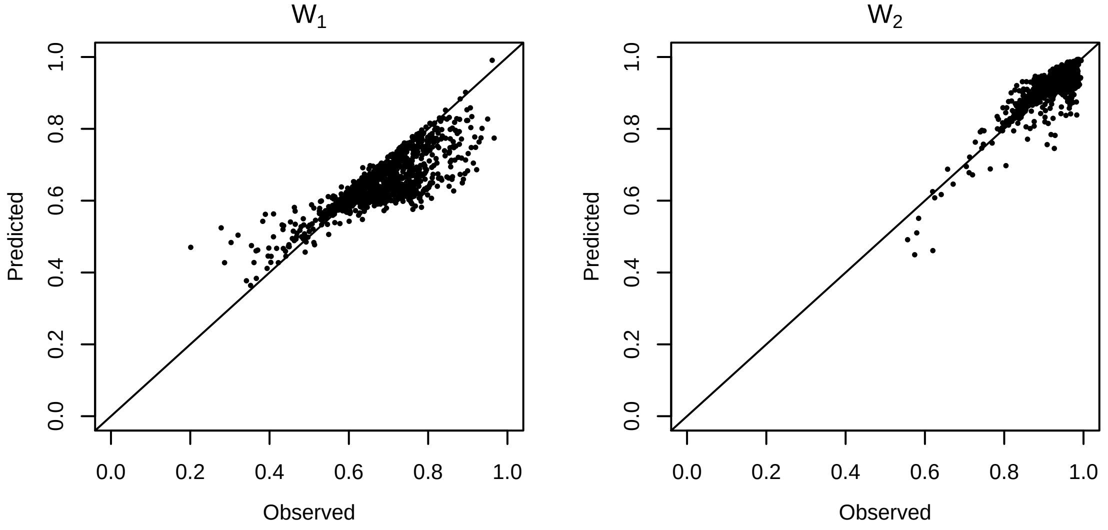

Figure 1: In-sample predictive performance of `ecoML()`.

model assumes the existence of contextual effect. First, the model can be fitted by the Gibbs sampler using the following syntax,

```
R> res <- eco(Y ~ X, N = N, data = census, context = TRUE, parameter = TRUE,
+     verbose = TRUE)

Starting Gibbs Sampler...
 10 percent done.
 20 percent done.
 30 percent done.
[output truncated for presentation purposes]
100 percent done.
```

where `context=TRUE` indicates that the NCAR ecological inference model is to be fitted, `parameter = TRUE` means that the posterior draws of the parameters will be saved in the output `res` in order to make out-of-sample predictions based on the fitted model, and `N` specifies the total number of individual-level observations within each aggregate unit so that both weighted and unweighted estimates can be obtained. The progress of the Gibbs sampler is printed to the screen if `verbose` is set to be `TRUE`. Moreover, one can specify the prior distribution for the parameters of the multivariate normal distribution in `eco()` (see the help file of `eco()` in the **eco** package for details). The default values of the prior parameters are specified as $\mu_0 = (0, 0), \tau_0 = 2, \nu_0 = 4$ and $S_0 = 10I_2$ where I_2 is a 2×2 identity matrix. This specification leads to an approximately uniform prior distribution of W_1 and W_2.

In this example, the other parameters are all set to be the default values. In particular, only $5,000$ Gibbs draws are taken, with no initial burn-in draws. In practice, to ensure proper convergence, the MCMC should be run for a longer period and the initial draws should also be discarded so that inference will be made based on draws from the target posterior distribution.

Multiple MCMC chains should also be run and analzyed using, for example, the **coda** package. We refer the readers to the standard texts on Bayesian data analysis for general advice on convergence (e.g., Gelman, Carlin, Stern, and Rubin 2004). Here, we implement the following syntax, which fits the same NCAR model as above but discards the initial 20,000 draws from a total of 50,000 draws while saving every 10th draw (thus, using the thinning interval of 10),

```
R> res <- eco(Y ~ X, N = N, data = census, context = TRUE, parameter = TRUE,
+    n.draws = 50000, burnin = 20000, thin = 9, verbose = TRUE)
```

The summary of the fitted model can be viewed using the `summary()` function,

```
R> summary(res)

Call: eco(formula = Y ~ X, data = census, N = N, context = TRUE,
parameter = TRUE, n.draws = 50000, burnin = 20000, thin = 9, verbose
= TRUE)

Parameter Estimates:
            mean std.dev    2.5 %  97.5 %
mu1      0.87546 0.12326  0.66900  1.1917
mu2      2.59947 0.16895  2.23791  2.9337
mu3     -0.86143 0.03624 -0.93457 -0.7895
Sigma11  0.30265 0.04654  0.23426  0.4212
Sigma12  0.03403 0.04248 -0.05348  0.1110
Sigma13 -0.29545 0.06961 -0.45102 -0.1737
Sigma22  0.90142 0.10404  0.72886  1.1367
Sigma23 -0.07646 0.11389 -0.30184  0.1542
Sigma33  1.37113 0.05947  1.25403  1.4928

*** Insample Predictions ***

Unweighted:
     mean  std.dev  2.5 % 97.5 %
W1 0.6631 0.017866 0.6318 0.7069
W2 0.9088 0.009108 0.8865 0.9248

Weighted:
     mean  std.dev  2.5 % 97.5 %
W1 0.6814 0.017236 0.6512 0.7236
W2 0.9318 0.007159 0.9143 0.9444

Number of Units: 1040

Number of Monte Carlo Draws: 3000
```

where the first part of the output summarizes the posterior distribution of the parameters of the trivariate normal distribution for the NCAR model – the ordinates of the distribution are (W_1^*, W_2^*, X^*), i.e., the logit-transformed black literacy rate, white literacy rate and black composition, respectively. Since N is specified in eco(), both "weighted" in-sample predictions aggregate estimates according to their actual group-specific population size. After controlling for the possible correlation between racial composition and literacy rate (i.e., contextual effect), the in-sample estimates, especially those of W1, are slightly improved compared to the CAR model (see the ML estimation of the CAR model earlier in this section).

In addition, the predict() function allows one to obtain the out-of-sample predictions of W_1 and W_2 based on their posterior predictive distributions using the posterior distribution of the model parameters. In the case of the NCAR model, the predictions will be based on the values of X_i which can be taken from another data set using the option newdata (the default, which we use here, is the data set used to fit the model). As before, the summary() function will summarize the results,

```
R> out <- predict(res, verbose = TRUE)

 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
 100 percent done.

R> summary(out)

Out-of-sample Prediction:
     mean std.dev   2.5 % 97.5 %
W1 0.6897 0.11497 0.44367 0.8793
W2 0.9071 0.08146 0.68846 0.9892
X  0.3362 0.21845 0.03964 0.8143

Number of Monte Carlo Draws: 3000
```

The default number of Monte Carlo draws is the same as the number of MCMC draws stored in the object res, but this number can be changed by users via the newdraw option.

3.3. Quantifying the aggregation effects

We revisit the function ecoML() to outline the computation of the fraction of missing information calculation described in Section 2.4. As in Section 3.2, the census dataset is used. If the SEM option is left at its default value of TRUE, the ecoML() function proceeds from the SEM algorithm, once the SEM algorithm is completed. The transition is shown as follows.

```
R> res.ML <- ecoML(Y ~ X, data = census, verbose = TRUE)

[output truncated for presentation purposes]
cycle 349/1000: 0.654 2.785 0.236 0.916 0.271 Prev LL: -1126.77
Final Theta: 0.654 2.785 0.236 0.916 0.271 Final LL: -1126.77
OPTIONS (flag: 4)   Ncar: No; Fixed Rho: No; SEM: Second run
cycle 1/1000: 0.000 0.000 1.000 1.000 0.000
```

```
R Matrix row 1 (Not done):   0.60  -0.75   0.17  -0.11  -0.16
R Matrix row 2 (Not done):  -0.20   0.40   0.01   0.10  -0.03
R Matrix row 3 (Not done):   0.01  -0.01   0.63  -0.14  -0.15
R Matrix row 4 (Not done):   0.00   0.10  -0.18   0.59  -0.20
R Matrix row 5 (Not done):  -0.07  -0.07  -0.26  -0.32   0.67

cycle 2/1000: 1.223 1.886 0.698 1.057 0.010
[output truncated for presentation purposes]
```

Note the final, converged parameter values (e.g., $\hat{\mu}_1 = 0.654$). The SEM algorithm then begins to calculate the DM matrix, which is necessary for the computation of the asymptotic variance matrix. Each row converges independently of the other rows, and the program output tracks this progress.

Once the SEM algorithm has converged, the final estimate for the DM matrix is displayed:

```
R> res.ML <- ecoML(Y ~ X, data = census, verbose = TRUE)

[output truncated for presentation purposes]
cycle 94/1000: 0.654 2.785 0.236 0.917 0.270 Prev LL: -1126.77

R Matrix row 1 (    Done):   0.52  -0.73  -0.08  -0.13  -0.25
R Matrix row 2 (    Done):  -0.20   0.46   0.01   0.19  -0.02
R Matrix row 3 (    Done):  -0.00   0.01   0.61  -0.18  -0.16
R Matrix row 4 (    Done):   0.01   0.09  -0.19   0.58  -0.20
R Matrix row 5 (    Done):  -0.06  -0.08  -0.28  -0.35   0.69

Final Theta: 0.654 2.785 0.236 0.917 0.270 Final LL: -1126.77
```

To obtain an estimate of the fraction of missing data simply type (see Equation 5),

```
R> res.ML$Fmis

[1] 0.6256639 0.5668982 0.6615895 0.6428587 0.7721757
```

For instance, the fraction of missing information for the calculation of black literacy is about 62.6%. The analogous quantity for white literacy is lower because in some counties, whites make up a very large proportion of the population, thus resulting in tighter bounds.

In addition to computing the fraction of missing data on parameter estimation, the **eco** package includes limited functionality for hypothesis testing (see Section 2.4). Currently, setting the `hyptest` parameter to `TRUE` for the `ecoML()` function, will calculate quantities of interest with parameters constrained to the null hypothesis: $\mu_1 = \mu_2$. Continuing with the census example, this null hypothesis would be that the mean black literacy rate is equal to the white literacy rate. To restrict the parameter space to this hypothesis, simply type,

```
R> res.ML.HT <- ecoML(Y ~ X, data = census, verbose = FALSE, hyptest = TRUE)
```

With the results of the algorithm for both the constrained and unconstrained problem, the calculation of the fraction of missing information under this hypothesis is straightforward. First, calculate the observed log-likelihood ratio test statistic, then the complete log-likelihood ratio test statistic, and then subtract the ratio of those two quantities from 1 (see Equation 6).

```
R> n <- dim(census)[1]
R> obs.loglik.stat <- 2 * (res.ML$loglik -  res.ML.HT$loglik)
R> com.loglik.stat <- Qfun(res.ML$theta.em, res.ML$suff.stat, n) -
+    Qfun(res.ML.HT$theta.em, res.ML.HT$suff.stat, n)
R> frac.miss.data <- 1 - (obs.loglik.stat / com.loglik.stat)
R> obs.loglik.stat

[1] 462.8226

R> frac.miss.data

[1] 3.575604
```

3.4. Fitting the nonparametric models

To avoid the distributional assumptions that are common to parametric models, the **eco** package also fits the nonparametric Bayesian models described in Section 2.3. Here, we use the voter registration dataset (see Section 2.1) to illustrate the use of the `ecoNP()` function, which fits the Bayesian nonparametric models via the MCMC algorithms. The following command fits the nonparametric model under the CAR assumption,

```
R> res <- ecoNP(Y ~ X, data = reg, N = N, parameter = TRUE, n.draws = 50000,
+    burnin = 20000, thin = 9, verbose = TRUE)

Starting Gibbs Sampler...
 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
100 percent done.
```

where the default prior specification places a diffuse distribution on the concentration parameter of the Dirichlet process prior. This default specification can be changed by the user (see the help file of `ecoNP()` in the **eco** package). Many of the inputs of `ecoNP()` are the same as those of `eco()`. For example, the nonparametric NCAR model can be fitted by the `context = TRUE` option.

Like the Bayesian parametric models, one can summarize the in-sample estimates of W using `summary()`. Unlike the parametric models, however, no parameter estimates are presented since the distribution of W is estimated nonparametrically based on a mixture of normal distributions,

```
R> summary(res)

Call: ecoNP(formula = Y ~ X, data = reg, N = N, parameter = TRUE,
n.draws = 50000, burnin = 20000, thin = 9,verbose = TRUE)

*** Insample Predictions ***

Unweighted:
     mean std.dev  2.5 % 97.5 %
W1 0.5824 0.03349 0.5159 0.6479
W2 0.8441 0.01160 0.8214 0.8671


Weighted:
     mean std.dev  2.5 % 97.5 %
W1 0.5375 0.04776 0.4437 0.6328
W2 0.8298 0.01355 0.8028 0.8564

Number of Units: 275
Number of Monte Carlo Draws: 3000
```

Finally, out-of sample predictions can be made in the same way as done for the parametric Bayesian model. That is, we use the generic `predict()` and `summary()` functions,

```
R> out <- predict(res, verbose = TRUE)

 10 percent done.
 20 percent done.
 30 percent done.
 [output truncated for presentation purposes]
100 percent done.

R> summary(out)

Out-of-sample Prediction:
     mean std.dev   2.5 % 97.5 %
W1 0.5915  0.2558 0.04671  0.997
W2 0.8291  0.2072 0.26665  1.000

Number of Monte Carlo Draws: 825000
```

Since the distribution is estimated nonparametrically, the out-of-sample prediction is generated for each observation. Hence, the total number of Monte Carlo draws is $825,000 = 275 \times 3,000$.

3.5. Incorporating the individual-level data

If survey or other supplemental, individual-level data is available, this information can easily be incorporated into the models. Adding this data will alleviate the adverse aggregation effects endemic to ecological inference. In **eco** package, the three main functions, `eco()`, `ecoML()`, `ecoNP()`, all take supplemental individual level data. In this section we illustrate an example using `ecoML()` and `eco()`.

Continuing with the county literacy rates example, assume that the actual, within-county literacy for whites and blacks is collected for the first 100 counties (and only these counties). Simply, create an $n \times 2$ matrix of the supplemental data with the first column `W1` and the second column `W2`,

```
R> survey.records <- 1:100
R> survey.data <- census[survey.records, c("W1", "W2")]
```

Next, execute the `ecoML()` set the `supplement` parameter to the matrix of survey data as follows,

```
R> res <- ecoML(Y ~ X, data = census[-survey.records, ],
+    supplement = survey.data, verbose = FALSE)
R> res$theta.em

       u1        u2        s1        s2       r12
0.6907367 2.6884871 0.2355874 0.7316871 0.4019745

R> res$Fmis

[1] 0.6194664 0.5708665 0.6396823 0.6209216 0.7701529
```

The estimated mean black literacy rate is about 65%, with the fraction of missing data being 61.9%, 0.6 percentage points less than without the supplemental data. Thus, adding true values for about 10% of the data points did not reduce information loss by much in this case.

Fitting the NCAR model is similar to the operation of the CAR model above. Since, no additional covariates are needed, simply adjust the `context` parameter,

```
R> survey.records<-1:100
R> survey.data <- census[survey.records, c("W1", "W2", "X")]
R> res.ML.NCAR <- ecoML(Y ~ X, data = census[-survey.records, ],
+    supplement = survey.data, context = TRUE, verbose = FALSE)
R> res.ML.NCAR$theta.em

        ux         u1         u2         sx         s1         s2        r1x
-0.4322371  0.6441360  2.5218557  1.3668375  0.1856183  0.3419140 -0.3184235
       r2x        r12
 0.5613339  0.1730306

R> res.ML.NCAR$Fmis
```

```
        ux         u1         u2         sx         s1         s2        r1x 
-0.2433929  0.3832901  0.3348818  0.7496725  0.6578700  0.6262803  0.5487799 
       r2x        r12 
 0.6249806  0.5197837 
```

Under the NCAR model, the mean black literacy rate is estimated to be 66%, which is slightly higher than estimated under the CAR model with the same survey data provided. The average white literacy rate is adjusted slightly downward in this more general model, from 94% to 93%. The change in the fraction of missing data is not monotonic when switching models. Information loss due to aggregation is larger for estimating the mean black literacy rate under NCAR, but is smaller when estimating mean white literacy.

Similarly, the **eco** package can fit Bayesian models with supplemented individual level data. Below, we fit the parametric CAR and NCAR models via `eco()` using the same census data with the first 100 counties's data revealed.

```
R> survey.records <- 1:100
R> survey.data <- census[survey.records, c("W1", "W2")]
R> res.CAR <- eco(Y ~ X, data = census[-survey.records, ], N = N,
+    supplement = survey.data, parameter = TRUE, n.draws = 50000,
+    burnin = 20000, thin = 9, verbose = FALSE)
R> survey.data <- census[survey.records, c("W1", "W2", "X")]
R> res.NCAR <- eco(Y ~ X, data = census[-survey.records, ], N = N,
+    supplement = survey.data, context = TRUE, parameter = TRUE,
+    n.draws = 50000, burnin = 20000, thin = 9, verbose = FALSE)
```

Then, the posterior means of the parameters can be computed directly from the output objects (or via the `summary()` function).

```
R> colMeans(res.CAR$mu)

      mu1       mu2 
0.6199912 2.8158310 

R> colMeans(res.CAR$Sigma)

  Sigma11   Sigma12   Sigma22 
0.2306994 0.1427673 0.9081327 

R> colMeans(res.NCAR$mu)

       mu1        mu2        mu3 
 0.7821447  2.6818828 -0.8604318 

R> colMeans(res.NCAR$sigma)

    Sigma11     Sigma12     Sigma13     Sigma22     Sigma23     Sigma33 
 0.25244727  0.13161135 -0.26913956  0.73974062  0.03141844  1.37279106 
```

Under the Bayesian CAR model, the mean black literacy rate is estimated to be 64.3% and white literacy rate 92.4%. Under the NCAR model, these two estimates are 68% and 91.8%, respectively. Comparing to the sample estimates (68.4% and 91.9%), `ecoNP()` performs very well in this particular data set with the aid of 100 individual-level observations.

4. What's new?

This section summarizes the history of all prior changes that are made to the **eco** package.

Version	Date	Changes
3.1-4	2009-07-13	Minor documentation fixes, final version for JSS publication.
3.1-3	2009-07-05	Minor documentation fixes.
3.1-2	2009-01-29	Minor documentation fixes.
3.1-1	2007-06-27	Some minor improvements.
3.0-2	2007-01-11	Made it comparable with the Windows; a bug fix in `summary.ecoML()`.
3.0-1	2006-12-27	A major revision; added ML estimation, calculation of fraction of missing information, stable release for R 2.4.1.
2.2-2	2006-09-23	Changed due to updates in R.
2.2-1	2005-09-28	Nonparametric model with contextual effects added.
2.1-1	2005-07-06	A major revision; added bounds and prediction; added/updated other functionalities.
1.1-1	2005-06-15	Add the Metropolis algorithm to sample W.
1.0-1	2004-12-21	First official version; submitted to CRAN.
0.9-1	2004-09-07	First beta version.

Acknowledgments

We thank an anonymous referee for helpful suggestions. Financial support from the National Science Foundation (SES–0550873) and Princeton University Committee on Research in the Humanities and Social Sciences is acknowledged.

References

Achen CH, Shively WP (1995). *Cross-Level Inference*. University of Chicago Press, Chicago.

Bonferroni CE (1936). "Teoria Statistica delle Classi e Calcolo delle Probabilità." *Publicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Fienze*, **8**, 3–62.

Cho WKT (1998). "Iff the Assumptions Fits. . . : A Comment on the King Ecological Inference Solution." *Political Analysis*, **7**, 143–163.

Cho WKT, Gaines BJ (2004). "The Limits of Ecological Inference: The Case of Split-Ticket Voting." *American Journal of Political Science*, **48**(1), 152–171.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**, 1–37.

Dey D, Müller P, Sinha D (eds.) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Springer-Verlag, New York.

Duncan OD, Davis B (1953). "An Alternative to Ecological Correlation." *American Sociological Review*, **18**(6), 665–666.

Fréchet M (1940). *Les Probabilitiés, Associées a un Système d'Événments Compatibles et Dépendants*, volume Premiere Partie. Hermann & Cie, Paris.

Freedman DA (1999). "Ecological Inference and the Ecological Fallacy." In N Smelser, P Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, volume 6, pp. 4027–4030. Elsevier.

Freedman DA, Klein SP, Sacks J, Smyth CA, Everett CG (1991). "Ecological Regression and Voting Rights." *Evaluation Review*, **15**, 673–816.

Freedman DA, Ostland M, Roberts MR, Klein SP (1998). "Review of 'A Solution to the Ecological Inference Problem'." *Journal of the American Statistical Association*, **93**, 1518–1522.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis.* 2nd edition. Chapman & Hall, London.

Gelman A, Park DK, Ansolabehere S, Price PN, Minnite LC (2001). "Models, Assumptions and Model Checking in Ecological Regressions." *Journal of the Royal Statistical Society A*, **164**, 101–118.

Goodman L (1953). "Ecological Regressions and Behavior of Individuals." *American Sociological Review*, **18**, 663–666.

Goodman LA (1959). "Some Alternatives to Ecological Correlation." *The American Journal of Sociology*, **64**, 610–624.

Greenland S, Robins JM (1994). "Ecologic Studies: Biases, Misconceptions, and Counterexamples." *American Journal of Epidemiology*, **139**, 747–760.

Heitjan DF, Rubin DB (1991). "Ignorability and Coarse Data." *The Annals of Statistics*, **19**, 2244–2253.

Hoeffding W (1940). "Masstabinvariate Korrelationstheorie." *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik de Universität Berlin*, **5**, 179–233. Reprinted as Scale-Invariant Correlation Theory. in Fisher NI and Sen, PK (ed.) (1994). *The Collected Works of Wassily Hoeffding*, pp. 57–107. Springer-Verlag, New York.

Imai K, King G (2004). "Did Illegal Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?" *Perspectives on Politics*, **2**(3), 537–549.

Imai K, King G, Lau O (2008a). "Toward A Common Framework of Statistical Analysis and Development." *Journal of Computational and Graphical Statistics*, **17**(4), 892–913.

Imai K, King G, Lau O (2009). ***Zelig**: Everyone's Statistical Software.* R package version 3.4-5, URL http://CRAN.R-project.org/package=Zelig.

Imai K, Lu Y, Strauss A (2008b). "Bayesian and Likelihood Inference for 2 × 2 Ecological Tables: An Incomplete Data Approach." *Political Analysis*, **16**(1), 41–69.

Judge GG, Miller DJ, Cho WKT (2004). "An Information Theoretic Approach to Ecological Estimation and Inference." In G King, O Rosen, M Tanner (eds.), *Ecological Inference: New Methodological Strategies*, pp. 162–187. Cambridge University Press, Cambridge.

King G (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press, Princeton, NJ.

King G (1999). "Comment on "Review of 'A Solution to the Ecological Inference Problem' "." *Journal of the American Statistical Association*, **94**, 352–355.

King G, Rosen O, Tanner MA (1999). "Binomial-Beta Hierarchical Models for Ecological Inference." *Sociological Methods & Research*, **28**, 61–90.

King G, Rosen O, Tanner MA (eds.) (2004). *Ecological Inferece: New Methodological Strategies.* Cambridge University Press.

Kong A, Meng XL, Nicolae DL (2008). "Quantifying the Fraction of Missing Information for Hypothesis Testing in Statistical and Genetic Studies." *Statistical Science*, **23**(3), 287–312.

Martin AD, Quinn KM, Park JH (2011). "**MCMCpack**: Markov Chain Monte Carlo in R." *Journal of Statistical Software*, **42**(9), 1–21. URL http://www.jstatsoft.org/v42/i09/.

McLaughlan GJ, Krishnan T (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

Meng XL, Rubin DB (1991). "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm." *Journal of the American Statistical Association*, **86**, 899–909.

Plummer M, Best N, Cowles K, Vines K (2006). "**coda**: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Robinson WS (1950). "Ecological Correlations and the Behavior of Individuals." *American Sociological Review*, **15**(3), 351–357.

Rosen O, Jiang W, King G, Tanner MA (2001). "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case." *Statistica Neerlandica*, **55**(2), 134–156.

Wakefield J (2004). "Ecological Inference for 2 × 2 Tables." *Journal of the Royal Statistical Society A*, **167**, 385–445.

Affiliation:

Kosuke Imai
Department of Politics
Princeton University
Princeton, NJ 08544, United States of America
Telephone: +1-609-258-6601
Fax: +1-609-258-1110
E-mail: kimai@Princeton.Edu
URL: http://imai.princeton.edu/

Ying Lu
Department of Humanities and Social Sciences
Steinhardt School of Education, Cultural and Human Development
New York University
246 Greene Street
New York, NY 10012, United States of America
E-mail: yl46@nyu.edu

Aaron Strauss
The Mellman Group
1023 31st St NW
5th Floor
Washington, DC 20007, United States of America
Telephone: +1-202-625-0370
E-mail: astrauss@mellmangroup.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 42, Issue 5 *Submitted:* 2007-02-06
May 2011 *Accepted:* 2007-06-18