

## 0.1 **ei.hier: Hierarchical Ecological Inference Model for $2 \times 2$ Tables**

Given contingency tables with observed marginals, ecological inference (EI) models estimate each internal cell value for each table. The hierarchical EI model estimates a Bayesian model for  $2 \times 2$  tables. The model is implemented using a Markov Chain Monte Carlo algorithm (via a combination of slice and Gibbs sampling). For a Bayesian implementation of EI that accounts for temporal dependence, see Quinn's dynamic EI model (Section ??). For contingency tables larger than 2 rows by 2 columns, see  $R \times C$  EI (Section ??).

### Syntax

```
> z.out <- zelig(cbind(t0, t1) ~ x0 + x1, N = NULL,
                 model = "MCMCEi.hier", data = mydata)
> x.out <- setx(z.out, fn = NULL, cond = TRUE)
> s.out <- sim(z.out, x = x.out)
```

### Inputs

- **t0, t1**: numeric vectors (either counts or proportions) containing the column margins of the units to be analyzed.
- **x0, x1**: numeric vectors (either counts or proportions) containing the row margins of the units to be analyzed.
- **N**: total counts per contingency table (unit). If **t0, t1, x0** and **x1** are proportions, you must specify **N**.

### Additional Inputs

In addition, **zelig()** accepts the following additional inputs for **ei.hier** to monitor the convergence of the Markov chain:

- **burnin**: number of the initial MCMC iterations to be discarded (defaults to 5,000).
- **mcmc**: number of the MCMC iterations after burnin (defaults to 50,000).
- **thin**: thinning interval for the Markov chain. Only every **thin**-th draw from the Markov chain is kept. The value of **mcmc** must be divisible by this value. The default value is 1.
- **verbose**: defaults to **FALSE**. If **TRUE**, the progress of the sampler (every 10%) is printed to the screen.
- **seed**: seed for the random number generator. The default is **NA** which corresponds to a random seed of 12345.

The model also accepts the following additional arguments to specify prior parameters used in the model:

- **m0**: prior mean of  $\mu_0$  (defaults to 0).
- **M0**: prior variance of  $\mu_0$  (defaults to 2.287656).
- **m1**: prior mean of  $\mu_1$  (defaults to 0).
- **M1**: prior variance of  $\mu_1$  (defaults to 2.287656).
- **a0**:  $a_0/2$  is the shape parameter for the Inverse Gamma prior on  $\sigma_0^2$  (defaults to 0.825).
- **b0**:  $b_0/2$  is the scale parameter for the Inverse Gamma prior on  $\sigma_0^2$  (defaults to 0.0105).
- **a1**:  $a_1/2$  is the shape parameter for the Inverse Gamma prior on  $\sigma_1^2$  (defaults to 0.825).
- **b1**:  $b_1/2$  is the scale parameter for the Inverse Gamma prior on  $\sigma_1^2$  (defaults to 0.0105).

Users may wish to refer to `help(MCMChierEI)` for more information.

## Convergence

Users should verify that the Markov Chain converges to its stationary distribution. After running the `zelig()` function but before performing `setx()`, users may conduct the following convergence diagnostics tests:

- `geweke.diag(z.out$coefficients)`: The Geweke diagnostic tests the null hypothesis that the Markov chain is in the stationary distribution and produces z-statistics for each estimated parameter.
- `heidel.diag(z.out$coefficients)`: The Heidelberger-Welch diagnostic first tests the null hypothesis that the Markov Chain is in the stationary distribution and produces p-values for each estimated parameter. Calling `heidel.diag()` also produces output that indicates whether the mean of a marginal posterior distribution can be estimated with sufficient precision, assuming that the Markov Chain is in the stationary distribution.
- `raftery.diag(z.out$coefficients)`: The Raftery diagnostic indicates how long the Markov Chain should run before considering draws from the marginal posterior distributions sufficiently representative of the stationary distribution.

If there is evidence of non-convergence, adjust the values for `burnin` and `mcmc` and rerun `zelig()`.

Advanced users may wish to refer to `help(geweke.diag)`, `help(heidel.diag)`, and `help(raftery.diag)` for more information about these diagnostics.

## Examples

### 1. Basic examples

Attaching the example dataset:

```
> data(eidat)
> eidat
```

Estimating the model using `ei.hier`:

```
> z.out <- zelig(cbind(t0, t1) ~ x0 + x1, model = "ei.hier", data = eidat,
+               mcmc = 40000, thin = 10, burnin = 10000, verbose = TRUE)
> summary(z.out)
```

Setting values for in-sample simulations given marginal values of `x0`, `x1`, `t0`, and `t1`:

```
> x.out <- setx(z.out, fn = NULL, cond = TRUE)
```

In-sample simulations from the posterior distribution:

```
> s.out <- sim(z.out, x = x.out)
```

Summarizing in-sample simulations at aggregate level weighted by the count in each unit:

```
> summary(s.out)
```

Summarizing in-sample simulations at unit level for the first 5 units:

```
> summary(s.out, subset = 1:5)
```

## Model

Consider the following  $2 \times 2$  contingency table for the racial voting example. For each geographical unit  $i = 1, \dots, p$ , the marginals  $t_i^0$ ,  $t_i^1$ ,  $x_i^0$ , and  $x_i^1$  are known, and we would like to estimate  $n_i^{00}$ ,  $n_i^{01}$ ,  $n_i^{10}$ , and  $n_i^{11}$ .

	No Vote	Vote	
Black	$n_i^{00}$	$n_i^{01}$	$x_i^0$
White	$n_i^{10}$	$n_i^{11}$	$x_i^1$
	$t_i^0$	$t_i^1$	$N_i$

The marginal values  $x_i^0$ ,  $x_i^1$ ,  $t_i^0$ ,  $t_i^1$  are observed as either counts or fractions. If fractions, the counts can be obtained by multiplying by the total counts per table  $N_i = n_i^{00} + n_i^{01} + n_i^{10} + n_i^{11}$  and rounding to the nearest integer. Although there are four internal cells, only two unknowns are modeled since  $n_i^{01} = x_i^0 - n_i^{00}$  and  $n_i^{11} = x_i^1 - n_i^{10}$ .

The hierarchical Bayesian model for ecological inference in  $2 \times 2$  is illustrated as following:

- The *stochastic component* of the model assumes that

$$\begin{aligned} n_i^{00} | x_i^0, \beta_i^b &\sim \text{Binomial}(x_i^0, \beta_i^b), \\ n_i^{10} | x_i^1, \beta_i^w &\sim \text{Binomial}(x_i^1, \beta_i^w) \end{aligned}$$

where  $\beta_i^b$  is the fraction of the black voters who vote and  $\beta_i^w$  is the fraction of the white voters who vote.  $\beta_i^b$  and  $\beta_i^w$  as well as their aggregate level summaries are the focus of inference.

- The *systematic component* is

$$\begin{aligned} \beta_i^b &= \frac{\exp \theta_i^0}{1 + \exp \theta_i^0} \\ \beta_i^w &= \frac{\exp \theta_i^1}{1 + \exp \theta_i^1} \end{aligned}$$

The logit transformations of  $\beta_i^b$  and  $\beta_i^w$ ,  $\theta_i^0$ , and  $\theta_i^1$  now take value on the real line. (Future versions may allow  $\beta_i^b$  and  $\beta_i^w$  to be functions of observed covariates.)

- The *priors* for  $\theta_i^0$  and  $\theta_i^1$  are given by

$$\begin{aligned} \theta_i^0 | \mu_0, \sigma_0^2 &\sim \text{Normal}(\mu_0, \sigma_0^2), \\ \theta_i^1 | \mu_1, \sigma_1^2 &\sim \text{Normal}(\mu_1, \sigma_1^2) \end{aligned}$$

where  $\mu_0$  and  $\mu_1$  are the means, and  $\sigma_0^2$  and  $\sigma_1^2$  are the variances of the two corresponding (independent) normal distributions.

- The *hyperpriors* for  $\mu_0$  and  $\mu_1$  are given by

$$\begin{aligned}\mu_0 &\sim \text{Normal}(m_0, M_0), \\ \mu_1 &\sim \text{Normal}(m_1, M_1),\end{aligned}$$

where  $m_0$  and  $m_1$  are the means of the (independent) normal distributions while  $M_0$  and  $M_1$  are the variances.

- The *hyperpriors* for  $\sigma_0^2$  and  $\sigma_1^2$  are given by

$$\begin{aligned}\sigma_0^2 &\sim \text{Inverse Gamma}\left(\frac{a_0}{2}, \frac{b_0}{2}\right), \\ \sigma_1^2 &\sim \text{Inverse Gamma}\left(\frac{a_1}{2}, \frac{b_1}{2}\right),\end{aligned}$$

where  $a_0/2$  and  $a_1/2$  are the shape parameters of the (independent) Gamma distributions while  $b_0/2$  and  $b_1/2$  are the scale parameters.

The default hyperpriors for  $\mu_0$ ,  $\mu_1$ ,  $\sigma_0^2$ , and  $\sigma_1^2$  are chosen such that the prior distributions of  $\beta^b$  and  $\beta^w$  are flat.

## Output Values

The output of each Zelig command contains useful information which you may view. For example, if you run

```
> z.out <- (cbind(t0, t1) ~ x0 + x1, N = NULL,
            model = "ei.hier", data = mydata)
```

then you may examine the available information in `z.out` by using `names(z.out)`, see the draws from the posterior distribution of the quantities of interest by using `z.out$coefficients`, and a default summary of information through `summary(z.out)`. Other elements available through the `$` operator are listed below.

- From the `zelig()` output object `z.out`, you may extract:
  - `coefficients`: draws from the posterior distributions of the parameters.
  - `zelig.data`: the input data frame if `save.data = TRUE`.
  - `N`: the total counts when the inputs are fractions.
  - `seed`: the random seed used in the model.
- From `summary(z.out)`, you may extract:

- **summary**: a matrix containing the summary information of the posterior estimation of  $\beta_i^b$  and  $\beta_i^w$  for each unit and the parameters  $\mu_0$ ,  $\mu_1$ ,  $\sigma_1$  and  $\sigma_2$  based on the posterior distribution. The first  $p$  rows correspond to  $\beta_i^b$ ,  $i = 1, \dots, p$ , the row names are in the form of `p0tablei`. The  $(p + 1)$ -th to the  $2p$ -th rows correspond to  $\beta_i^w$ ,  $i = 1, \dots, p$ . The row names are in the form of `p1tablei`. The last four rows contain information about  $\mu_0$ ,  $\mu_1$ ,  $\sigma_0^2$  and  $\sigma_1^2$ , the prior means and variances of  $\theta_0$  and  $\theta_1$ .
- From the `sim()` output object `s.out`, you may extract quantities of interest arranged as arrays indexed by simulation  $\times$  column  $\times$  row  $\times$  observation, where column and row refer to the column dimension and the row dimension of the contingency table, respectively. In this model, only  $2 \times 2$  contingency tables are analyzed, hence column= 2 and row= 2 in all cases. Available quantities are:
  - `qi$ev`: the simulated expected values of each internal cell given the observed marginals.
  - `qi$pr`: the simulated expected values of each internal cell given the observed marginals.

## How to Cite

To cite the *ei.hier* Zelig model use:

Ben Goodrich and Ying Lu. 2007. “ei.hier: Hierarchical Ecological Inference Model for 2x2 tables” in Kosuke Imai, Gary King, and Olivia Lau, “Zelig: Everyone’s Statistical Software,” <http://gking.harvard.edu/zelig>.

To cite Zelig as a whole, please reference these two sources:

Kosuke Imai, Gary King, and Olivia Lau. 2007. “Zelig: Everyone’s Statistical Software,” <http://GKing.harvard.edu/zelig>.

Kosuke Imai, Gary King, and Olivia Lau. 2007. “Toward A Common Framework for Statistical Analysis and Development,” <http://gking.harvard.edu/files/abs/z-abs.shtml>.

## See also

*ei.hier* function is part of the MCMCpack library by Andrew D. Martin and Kevin M. Quinn (Martin and Quinn 2005). The convergence diagnostics are part of the CODA library by Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines (Plummer et al. 2005). Sample data are adapted from Martin and Quinn (2005).

# Bibliography

Martin, A. D. and Quinn, K. M. (2005), *MCMCpack: Markov chain Monte Carlo (MCMC) Package*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2005), *coda: Output analysis and diagnostics for MCMC*.