eco: R Package for Fitting Bayesian Models of Ecological Inference in 2×2 Tables*

Kosuke Imai† Ying Lu‡

Version 2.2-1
September 28, 2005

Abstract

`eco` is a publicly available R package that fits parametric and nonparametric Bayesian models for ecological inference in 2×2 tables. The models are fit using the Markov chain Monte Carlo algorithms that are described in Imai and Lu (2004).

1 Installation

To use `eco`, you must install the statistical software R (if it is not already installed) as well as the `eco` package.

1.1 Windows systems

1. **Installing the latest version of R.** You may skip this step if the latest version of R is already installed on your system. If R is not installed on your system, go to the Comprehensive R Archive Network (CRAN) website (http://cran.r-project.org) and download the latest R installer for Windows. Double-click on the `.exe` file to launch the installer. We recommend that you accept the default installation options.

2. **Installing `eco`.** Start R and type at the prompt:

*Research support was provided in part by the Committee on Research in the Humanities and Social Sciences at Princeton University

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6610, Fax: 973–556–1929, Email: kimai@Princeton.Edu, URL: http://imai.princeton.edu

‡Postdoc Researcher, Institute for Quantitative Social Sciences, Harvard University, Cambridge MA 02138. Phone: 617–496–2031, Email: ylu@Latte.Harvard.Edu

```
install.packages("eco")
```

1.2 Unix/Linux systems

1. **Installing the latest version of R.** You may skip this step if the latest version of R is already installed on your system. If R is not installed on your system, it may either be installed locally (e.g., in an individual user's `bin` directory) or globally (e.g., in the `/bin` directory). The latter requires administrative privileges. In either case, the latest release of R may be downloaded from the CRAN website (http://cran.r-project.org).

2. **Installing** `eco`.

   (a) Create a local library directory if it does not exist already. Here, we use `~/.R/library` but you can specify a different directory. This directory can be created by typing the following command at the command prompt,

   ```
   mkdir ~/.R ~/.R/library
   ```

   (b) Open the `~/.Renviron` file in your home directory (or create it if it does not exist) and add the following line,

   ```
   R_LIBS="~/.R/library"
   ```

   Alternatively, one can define the environmental variable. For example, add the following line to your Bourne shell startup file (e.g., `.bashrc` file if you are using a bash shell),

   ```
   export R_LIBS="$HOME/.R/library"
   ```

   (c) Start R and type at the prompt:

   ```
   install.packages("eco", lib="~/.R/library/")
   ```

1.3 MacOS X systems

1. **Installing the latest version of R.** You may skip this step if the latest version of R is already installed on your system. If R is not installed on your system, you may download it from the CRAN website (http://cran.r-project.org).

2. **Installing** `eco`**.** If you are using RAqua, typing the following command at the prompt,

   ```
   install.packages("eco")
   ```

will install `eco` into the default local library directory, `~/Library/R/library`. If you are using the command line R, then the installation of the `eco` package can be done exactly in the same way as in Unix/Linux systems. You might want to set `R_LIBS` to `~/Library/R/library` so that the command line R and RAqua can share the same local library directory.

2 Command Overview

Only two commands are available now; `eco()` produces the in-sample and out-of-sample predictions for ecological inference problem in 2×2 tables, and `summary()` will summarize the results. To run an example script, start R and run the following commands:

```
library(eco)     # loads eco library
example(eco)     # runs the example script for the parametric model
example(ecoNP)   # runs the example script for the nonparametric model
```

For details of the commands and example scripts, see Section 3.

3 Command References

`eco` *Fitting the Parametric Bayesian Model of Ecological Inference in 2x2 Tables*

Description

`eco` is used to fit the parametric Bayesian model (based on a Normal/Inverse-Wishart prior) for ecological inference in 2×2 tables via Markov chain Monte Carlo. It gives the in-sample predictions as well as the estimates of the model parameters. The model and algorithm are described in Imai and Lu (2004). The contextual effect can also be modeled by following the strategy described in Imai and Lu (2005).

Usage

```
eco(formula, data = parent.frame(), N = NULL, supplement = NULL,
    context = FALSE, mu0 = 0, tau0 = 2, nu0 = 4, S0 = 10,
    mu.start = 0, Sigma.start = 10, parameter = TRUE,
    grid = FALSE, n.draws = 5000, burnin = 0, thin = 0,
    verbose = FALSE)
```

Arguments

`formula`	A symbolic description of the model to be fit, specifying the column and row margins of 2×2 ecological tables. `Y ~ X` specifies `Y` as the column margin and `X` as the row margin. Details and specific examples are given below.
`data`	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `eco` is called.
`N`	An optional variable representing the size of the unit; e.g., the total number of voters.
`supplement`	An optional matrix of supplemental data. The matrix has two columns, which contain additional individual-level data such as survey data for W_1 and W_2, respectively. If `NULL`, no additional individual-level data are included in the model. The default is `NULL`.
`context`	Logical. If `TRUE`, the contextual effect is also modeled. See Imai and Lu (2005) for details. The default is `FALSE`.
`mu0`	A scalar or a numeric vector that specifies the prior mean for the mean parameter μ. If it is a scalar, then its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=TRUE`, the length of μ is 3, otherwise it is 2. The default is `0`.
`tau0`	A positive integer representing the prior scale for the mean parameter μ. The default is `2`.
`nu0`	A positive integer representing the prior degrees of freedom of the variance matrix Σ. the default is `4`.
`S0`	A postive scalar or a positive definite matrix that specifies the prior scale matrix for the variance matrix Σ. If it is a scalar, then the prior scale matrix will be a digonal matrix with the same dimensions as Σ and the diagonal elements all take value of `S0`, otherwise `S0` needs to have same dimensions as Σ. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.
`mu.start`	A scalar or a numeric vector that specifies the starting values of the mean parameter μ. If it is a scalar, then its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=FALSE`, the length of μ is 2, otherwise it is 3. The default is `0`.
`Sigma.start`	A scalar or a positive definite matrix that specified the starting value of the variance matrix Σ. If it is a scalar, then the prior scale matrix will be a digonal matrix with the same dimensions as Σ and the diagonal elements all take value of `S0`,

otherwise `S0` needs to have same dimensions as Σ. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.

`parameter` Logical. If `TRUE`, the Gibbs draws of the population parameters, μ and Σ, are returned in addition to the in-sample predictions of the missing internal cells, W. The default is `TRUE`.

`grid` Logical. If `TRUE`, the grid method is used to sample W in the Gibbs sampler. If `FALSE`, the Metropolis algorithm is used where candidate draws are sampled from the uniform distribution on the tomography line for each unit. Note that the grid method is significantly slower than the Metropolis algorithm.

`n.draws` A positive integer. The number of MCMC draws. The default is `5000`.

`burnin` A positive integer. The burnin interval for the Markov chain; i.e. the number of initial draws that should not be stored. The default is `0`.

`thin` A positive integer. The thinning interval for the Markov chain; i.e. the number of Gibbs draws between the recorded values that are skipped. The default is `0`.

`verbose` Logical. If `TRUE`, the progress of the Gibbs sampler is printed to the screen. The default is `FALSE`.

Details

An example of 2×2 ecological table for racial voting is given below:

	black voters	white voters	
Voted	W_{1i}	W_{2i}	Y_i
Not voted	$1 - W_{1i}$	$1 - W_{2i}$	$1 - Y_i$
	X_i	$1 - X_i$	

where Y_i and X_i represent the observed margins, and W_1 and W_2 are unknown variables. All variables are proportions and hence bounded between 0 and 1. For each i, the following deterministic relationship holds, $Y_i = X_i W_{1i} + (1 - X_i) W_{2i}$.

Value

An object of class `eco` containing the following elements:

`call` The matched call.

`X` The row margin, X.

`Y` The column margin, Y.

`N`	The size of each table, N.
`burnin`	The number of initial burnin draws.
`thin`	The thinning interval.
`nu0`	The prior degrees of freedom.
`tau0`	The prior scale parameter.
`mu0`	The prior mean.
`S0`	The prior scale matrix.
`W`	A three dimensional array storing the posterior in-sample predictions of W. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.
`Wmin`	A numeric matrix storing the lower bounds of W.
`Wmax`	A numeric matrix storing the upper bounds of W.
`mu`	The posterior draws of the population mean parameter, μ.
`Sigma`	The posterior draws of the population variance matrix, Σ.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://imai.princeton.edu`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

References

Imai, Kosuke and Ying Lu. (2004) " Parametric and Nonparametric Bayesian Models for Ecological Inference in 2 × 2 Tables." Working Paper, Princeton University, available at `http://imai.princeton.edu/research/einonpar.html`

Imai, Kosuke and Ying Lu. (2005) "An Incomplete Data Approach to Ecological Inference." Working Paper, Princeton University, available at `http://imai.princeton.edu/research/coarse.html`

See Also

`ecoNP`, `predict.eco`, `summary.eco`

Examples

```
## load the registration data
data(reg)

## NOTE: convergence has not been properly assessed for the following
## examples. See Imai and Lu (2004, 2005) for more complete analyses.

## fit the parametric model with the default prior specification
res <- eco(Y ~ X, data = reg, verbose = TRUE)
## summarize the results
summary(res)

## obtain out-of-sample prediction
out <- predict(res, verbose = TRUE)
## summarize the results
summary(out)

## load the Robinson's census data
data(census)

## fit the parametric model with contextual effects and N
## using the default prior specification
res1 <- eco(Y ~ X, N = N, context = TRUE, data = census, verbose = TRUE)
## summarize the results
summary(res1)

## obtain out-of-sample prediction
out1 <- predict(res1, verbose = TRUE)
## summarize the results
summary(out1)
```

---

ecoNP *Fitting the Nonparametric Bayesian Model of Ecological Inference in 2x2 Tables*

Description

ecoNP is used to fit the nonparametric Bayesian model (based on a Dirichlet process prior) for ecological inference in 2×2 tables via Markov chain Monte Carlo. It gives the in-sample predictions as well as out-of-sample predictions for population inference. The model and algorithm are described in Imai and Lu (2004). The contextual effect can also be modeled by following the strategy described in Imai and Lu (2005).

Usage

```
ecoNP(formula, data = parent.frame(), N = NULL, supplement = NULL,
      context = FALSE, mu0 = 0, tau0 = 2, nu0 = 4, S0 = 10,
      alpha = NULL, a0 = 1, b0 = 0.1, parameter = FALSE,
      grid = FALSE, n.draws = 5000, burnin = 0, thin = 0,
      verbose = FALSE)
```

Arguments

formula	A symbolic description of the model to be fit, specifying the column and row margins of 2×2 ecological tables. `Y ~ X` specifies `Y` as the column margin and `X` as the row margin. Details and specific examples are given below.
data	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `ecoNP` is called.
N	An optional variable representing the size of the unit; e.g., the total number of voters.
supplement	An optional matrix of supplemental data. The matrix has two columns, which contain additional individual-level data such as survey data for W_1 and W_2, respectively. If `NULL`, no additional individual-level data are included in the model. The default is `NULL`.
context	Logical. If `TRUE`, the contextual effect is also modeled. See Imai and Lu (2005) for details. The default is `FALSE`.
mu0	A scalar or a numeric vector that specifies the prior mean for the mean parameter μ. If it is a scalar, then its value will be repeated to yield a vector of the length of μ, otherwise, it needs to be a vector of same length as μ. When `context=TRUE` , the length of μ is 3, otherwise it is 2. The default is `0`.
tau0	A positive integer representing the prior scale for the mean parameter μ. The default is 2.

`nu0` A positive integer representing the prior degrees of freedom of the variance matrix Σ. the default is `4`.

`S0` A postive scalar or a positive definite matrix that specifies the prior scale matrix for the variance matrix Σ. If it is a scalar, then the prior scale matrix will be a digonal matrix with the same dimensions as Σ and the diagonal elements all take value of `S0`, otherwise `S0` needs to have same dimensions as Σ. When `context=TRUE`, Σ is a 3×3 matrix, otherwise, it is 2×2. The default is `10`.

`alpha` A positive scalar representing a user-specified fixed value of the concentration parameter, α. If `NULL`, α will be updated at each Gibbs draw, and its prior parameters `a0` and `b0` need to be specified. The default is `NULL`.

`a0` A positive integer representing the value of shape parameter of the gamma prior distribution for α. The default is `1`.

`b0` A positive integer representing the value of the scale parameter of the gamma prior distribution for α. The default is `0.1`.

`parameter` Logical. If `TRUE`, the Gibbs draws of the population parameters, μ and Σ, are returned in addition to the in-sample predictions of the missing internal cells, W. The default is `FALSE`. This needs to be set to `TRUE` if one wishes to make population inferences through `predict.eco`. See an example below.

`grid` Logical. If `TRUE`, the grid method is used to sample W in the Gibbs sampler. If `FALSE`, the Metropolis algorithm is used where candidate draws are sampled from the uniform distribution on the tomography line for each unit. Note that the grid method is significantly slower than the Metropolis algorithm.

`n.draws` A positive integer. The number of MCMC draws. The default is `5000`.

`burnin` A positive integer. The burnin interval for the Markov chain; i.e. the number of initial draws that should not be stored. The default is `0`.

`thin` A positive integer. The thinning interval for the Markov chain; i.e. the number of Gibbs draws between the recorded values that are skipped. The default is `0`.

`verbose` Logical. If `TRUE`, the progress of the gibbs sampler is printed to the screen. The default is `FALSE`.

Details

An example of 2×2 ecological table for racial voting is given below:

	black voters	white voters

Voted	W_{1i}	W_{2i}	Y_i
Not voted	$1 - W_{1i}$	$1 - W_{2i}$	$1 - Y_i$
	X_i	$1 - X_i$	

where Y_i and X_i represent the observed margins, and W_1 and W_2 are unknown variables. All variables are proportions and hence bounded between 0 and 1. For each i, the following deterministic relationship holds, $Y_i = XW_{1i} + (1 - X_i)W_{2i}$.

Value

An object of class `ecoNP` containing the following elements:

`call`	The matched call.
`X`	The row margin, X.
`Y`	The column margin, Y.
`burnin`	The number of initial burnin draws.
`thin`	The thinning interval.
`nu0`	The prior degrees of freedom.
`tau0`	The prior scale parameter.
`mu0`	The prior mean.
`S0`	The prior scale matrix.
`a0`	The prior shape parameter.
`b0`	The prior scale parameter.
`W`	A three dimensional array storing the posterior in-sample predictions of W. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.
`Wmin`	A numeric matrix storing the lower bounds of W.
`Wmax`	A numeric matrix storing the upper bounds of W.
`mu`	A three dimensional array storing the posterior draws of the population mean parameter, μ. The first dimension indexes the Monte Carlo draws, the second dimension indexes the columns of the table, and the third dimension represents the observations.

`Sigma` A three dimensional array storing the posterior draws of the population variance matrix, Σ. The first dimension indexes the Monte Carlo draws, the second dimension indexes the parameters, and the third dimension represents the observations.

`alpha` The posterior draws of α.

`nstar` The number of clusters at each Gibbs draw.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://www.princeton.edu/~kimai`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

References

Imai, Kosuke and Ying Lu. (2004) " Parametric and Nonparametric Bayesian Models for Ecological Inference in 2×2 Tables." Proceedings of the American Statistical Association. `http://www.princeton.edu/~kimai/research/einonpar.html`

Imai, Kosuke and Ying Lu. (2005) "An Incomplete Data Approach to Ecological Inference." Working Paper, Princeton University, available at `http://www.princeton.edu/~kimai/research/einonpar.html`

See Also

`eco`, `predict.eco`, `summary.ecoNP`

Examples

```
## load the registration data
data(reg)

## NOTE: We set the number of MCMC draws to be a very small number in
## the following examples; i.e., convergence has not been properly
## assessed. See Imai and Lu (2004, 2005) for more complete examples.

## fit the nonparametric model to give in-sample predictions
## store the parameters to make population inference later
res <- ecoNP(Y ~ X, data = reg, n.draws = 50, param = TRUE, verbose = TRUE)
##summarize the results
```

```
summary(res)

## obtain out-of-sample prediction
out <- predict(res, verbose = TRUE)
## summarize the results
summary(out)

## density plots of the out-of-sample predictions
par(mfrow=c(2,1))
plot(density(out[,1]), main = "W1")
plot(density(out[,2]), main = "W2")

## load the Robinson's census data
data(census)

## fit the parametric model with contextual effects and N
## using the default prior specification
res1 <- ecoNP(Y ~ X, N = N, context = TRUE, param = TRUE, data = census,
              n.draws = 25, verbose = TRUE)
## summarize the results
summary(res1)

## out-of sample prediction
pres1 <- predict(res1)
summary(pres1)
```

---

ecoBD *Calculating the Bounds for Ecological Inference in RxC Tables*

Description

ecoBD is used to calculate the bounds for missing internal cells of $R \times C$ ecological table. The data can be entered either in the form of counts or proportions.

Usage

```
ecoBD(formula, data = parent.frame(), N = NULL)
```

Arguments

`formula`	A symbolic description of ecological table to be used, specifying the column and row margins of $R \times C$ ecological tables. Details and specific examples are given below.
`data`	An optional data frame in which to interpret the variables in `formula`. The default is the environment in which `ecoBD` is called.
`N`	An optional variable representing the size of the unit; e.g., the total number of voters. If `formula` is entered as counts and the last row and/or column is omitted, this input is necessary.

Details

The data may be entered either in the form of counts or proportions. If proportions are used, `formula` may omit the last row and/or column of tables, which can be calculated from the remaining margins. For example, `Y ~ X` specifies `Y` as the first column margin and `X` as the first row margin in 2×2 tables. If counts are used, `formula` may omit the last row and/or column margin of the table only if `N` is supplied. For larger tables, one can use `cbind()` and `+`. For example, `cbind(Y1, Y2, Y3) ~ X1 + X2 + X3 + X4)` specifies 3×4 tables.

An $R \times C$ ecological table in the form of counts:

$$\begin{array}{ccccc} n_{i11} & n_{i12} & \dots & n_{i1C} & n_{i1.} \\ n_{i21} & n_{i22} & \dots & n_{i2C} & n_{i2.} \\ \dots & \dots & \dots & \dots & \dots \\ n_{iR1} & n_{iR2} & \dots & n_{iRC} & n_{iR.} \\ n_{i.1} & n_{i.2} & \dots & n_{i.C} & N_i \end{array}$$

where $n_{nr.}$ and $n_{i.c}$ represent the observed margins, N_i represents the size of the table, and n_{irc} are unknown variables. Note that for each i, the following deterministic relationships hold; $n_{ir.} = \sum_{c=1}^{C} n_{irc}$ for $r = 1, \dots, R$, and $n_{i.c} = \sum_{r=1}^{R} n_{irc}$ for $c = 1, \dots, C$. Then, each of the unknown inner cells can be bounded in the following manner,

$$\max(0, n_{ir.} + n_{i.c} - N_i) \leq n_{irc} \leq \min(n_{ir.}, n_{i.c}).$$

If the size of tables, `N`, is provided,

An $R \times C$ ecological table in the form of proportions:

$$\begin{array}{cccc} W_{i11} & W_{i12} & \dots & W_{i1C} & Y_{i1} \end{array}$$

$$\begin{array}{ccccc} W_{i21} & W_{i22} & \ldots & W_{i2C} & Y_{i2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ W_{iR1} & W_{iR2} & \ldots & W_{iRC} & Y_{iR} \\ X_{i1} & X_{i2} & \ldots & X_{iC} & \end{array}$$

where Y_{ir} and X_{ic} represent the observed margins, and W_{irc} are unknown variables. Note that for each i, the following deterministic relationships hold; $Y_{ir} = \sum_{c=1}^{C} X_{ic} W_{irc}$ for $r = 1, \ldots, R$, and $\sum_{r=1}^{R} W_{irc} = 1$ for $c = 1, \ldots, C$. Then, each of the inner cells of the table can be bounded in the following manner,

$$\max(0, (X_{ic} + Y_{ir} - 1)/X_{ic}) \leq W_{irc} \leq \min(1, Y_{ir}/X_{ir}).$$

Value

An object of class `ecoBD` containing the following elements (When three dimensional arrays are used, the first dimension indexes the observations, the second dimension indexes the row numbers, and the third dimension indexes the column numbers):

`call`	The matched call.
`X`	A matrix of the observed row margin, X.
`Y`	A matrix of the observed column margin, Y.
`N`	A vector of the size of ecological tables, N.
`aggWmin`	A three dimensional array of aggregate lower bounds for proportions.
`aggWmax`	A three dimensional array of aggregate upper bounds for proportions.
`Wmin`	A three dimensional array of lower bounds for proportions.
`Wmax`	A three dimensional array of upper bounds for proportions.
`Nmin`	A three dimensional array of lower bounds for counts.
`Nmax`	A three dimensional array of upper bounds for counts.

The object can be printed through `print.ecoBD`.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://www.princeton.edu/~kimai`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

References

Imai, Kosuke. (2005) "Ecological Inference in $R \times C$ Tables" Working Paper, Princeton University.

See Also

`eco`, `ecoNP`

Examples

```
## load the registration data
data(reg)

## calculate the bounds
res <- ecoBD(Y ~ X, N = N, data = reg)
## print the results
print(res)
```

`predict.eco` *Out-of-Sample Posterior Prediction under the Parametric Bayesian Model for Ecological Inference in 2x2 Tables*

Description

Obtains out-of-sample posterior predictions under the fitted parametric Bayesian model for ecological inference. `predict` method for class `eco` and `ecoX`.

Usage

```
## S3 method for class 'eco':
predict(object, newdraw = NULL, subset = NULL, verbose = FALSE, ...)
## S3 method for class 'ecoX':
predict(object, newdraw = NULL, subset = NULL, newdata = NULL, cond = FALSE, verbos
```

Arguments

`object`	An output object from `eco` or `ecoNP`.
`newdraw`	An optional list containing two matrices (or three dimensional arrays for the nonparametric model) of MCMC draws of μ and Σ. Those elements should be named as `mu` and `Sigma`, respectively. The default is the original MCMC draws stored in `object`.
`newdata`	An optional data frame containing a new data set for which posterior predictions will be made. The new data set must have the same variable names as those in the original data.
`subset`	A scalar or numerical vector specifying the row number(s) of `mu` and `Sigma` in the output object from `eco`. If specified, the posterior draws of parameters for those rows are used for posterior prediction. The default is `NULL` where all the posterior draws are used.
`cond`	logical. If `TRUE`, then the conditional prediction will made for the parametric model with contextual effects. The default is `FALSE`.
`verbose`	logical. If `TRUE`, helpful messages along with a progress report on the Monte Carlo sampling from the posterior predictive distributions are printed on the screen. The default is `FALSE`.
`...`	further arguments passed to or from other methods.

Details

The posterior predictive values are computed using the Monte Carlo sample stored in the `eco` output (or other sample if `newdraw` is specified). Given each Monte Carlo sample of the parameters, we sample the vector-valued latent variable from the appropriate multivariate Normal distribution. Then, we apply the inverse logit transformation to obtain the predictive values of proportions, W. The computation may be slow (especially for the nonparametric model) if a large Monte Carlo sample of the model parameters is used. In either case, setting `verbose = TRUE` may be helpful in monitoring the progress of the code.

Value

`predict.eco` yields a matrix of class `predict.eco` containing the Monte Carlo sample from the posterior predictive distribution of inner cells of ecological tables. `summary.predict.eco` will summarize the output, and `print.summary.predict.eco` will print the summary.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

See Also

`eco`, `predict.ecoNP`

---

predict.eco — *Out-of-Sample Posterior Prediction under the Nonparametric Bayesian Model for Ecological Inference in 2x2 Tables*

Description

Obtains out-of-sample posterior predictions under the fitted nonparametric Bayesian model for ecological inference. `predict` method for class `ecoNP` and `ecoNPX`.

Usage

```
## S3 method for class 'ecoNP':
predict(object, newdraw = NULL, subset = NULL, obs = NULL, verbose = FALSE, ...)
## S3 method for class 'ecoNPX':
predict(object, newdraw = NULL, subset = NULL, obs = NULL, cond = FALSE, verbose =
```

Arguments

`object`	An output object from `ecoNP`.
`newdraw`	An optional list containing two matrices (or three dimensional arrays for the nonparametric model) of MCMC draws of μ and Σ. Those elements should be named as `mu` and `Sigma`, respectively. The default is the original MCMC draws stored in `object`.
`subset`	A scalar or numerical vector specifying the row number(s) of `mu` and `Sigma` in the output object from `eco`. If specified, the posterior draws of parameters for those rows are used for posterior prediction. The default is `NULL` where all the posterior draws are used.

`obs`	An integer or vector of integers specifying the observation number(s) whose posterior draws will be used for predictions. The default is `NULL` where all the observations in the data set are selected.
`cond`	logical. If `TRUE`, then the conditional prediction will made for the parametric model with contextual effects. The default is `FALSE`.
`verbose`	logical. If `TRUE`, helpful messages along with a progress report on the Monte Carlo sampling from the posterior predictive distributions are printed on the screen. The default is `FALSE`.
`...`	further arguments passed to or from other methods.

Details

The posterior predictive values are computed using the Monte Carlo sample stored in the `eco` or `ecoNP` output (or other sample if `newdraw` is specified). Given each Monte Carlo sample of the parameters, we sample the vector-valued latent variable from the appropriate multivariate Normal distribution. Then, we apply the inverse logit transformation to obtain the predictive values of proportions, W. The computation may be slow (especially for the nonparametric model) if a large Monte Carlo sample of the model parameters is used. In either case, setting `verbose = TRUE` may be helpful in monitoring the progress of the code.

Value

`predict.eco` yields a matrix of class `predict.eco` containing the Monte Carlo sample from the posterior predictive distribution of inner cells of ecological tables. `summary.predict.eco` will summarize the output, and `print.summary.predict.eco` will print the summary.

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

See Also

`eco`, `ecoNP`, `summary.eco`, `summary.ecoNP`

`summary.eco` *Summarizing the Results for the Bayesian Parametric Model for Ecological Inference in 2x2 Tables*

Description

`summary` method for class `eco`.

Usage

```
## S3 method for class 'eco':
summary(object, CI = c(2.5, 97.5), param = TRUE,
        units = FALSE, subset = NULL, ...)

## S3 method for class 'summary.eco':
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`object`	An output object from `eco`.
`CI`	A vector of lower and upper bounds for the Bayesian credible intervals used to summarize the results. The default is the equal tail 95 percent credible interval.
`x`	An object of class `summary.eco`.
`digits`	the number of significant digits to use when printing.
`param`	Logical. If `TRUE`, the posterior estimates of the population parameters will be provided. The default value is `TRUE`.
`units`	Logical. If `TRUE`, the in-sample predictions for each unit or for a subset of units will be provided. The default value is `FALSE`.
`subset`	A numeric vector indicating the subset of the units whose in-sample predications to be provided when `units` is `TRUE`. The default value is `NULL` where the in-sample predictions for each unit will be provided.
`...`	further arguments passed to or from other methods.

Value

`summary.eco` yields an object of class `summary.eco` containing the following elements:

`call`	The call from `eco`.
`n.obs`	The number of units.
`n.draws`	The number of Monte Carlo samples.
`agg.table`	Aggregate posterior estimates of the marginal means of W_1 and W_2 using X and N as weights.
`param.table`	Posterior estimates of model parameters: population mean estimates of W_1 and W_2 and their logit transformations.
`W1.table`	Unit-level posterior estimates for W_1.
`W2.table`	Unit-level posterior estimates for W_2.

This object can be printed by `print.summary.eco`

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://www.princeton.edu/~kimai`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

See Also

`eco`, `predict.eco`

`summary.ecoNP` *Summarizing the Results for the Bayesian Nonparametric Model for Ecological Inference in 2x2 Tables*

Description

`summary` method for class `ecoNP`.

Usage

```
## S3 method for class 'ecoNP':
summary(object, CI = c(2.5, 97.5), param = FALSE,
        units = FALSE, subset = NULL, ...)

## S3 method for class 'summary.ecoNP':
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`object`	An output object from `ecoNP`.
`CI`	A vector of lower and upper bounds for the Bayesian credible intervals used to summarize the results. The default is the equal tail 95 percent credible interval.
`x`	An object of class `summary.ecoNP`.
`digits`	the number of significant digits to use when printing.
`param`	Logical. If `TRUE`, the posterior estimates of the population parameters will be provided. The default value is `FALSE`.
`units`	Logical. If `TRUE`, the in-sample predictions for each unit or for a subset of units will be provided. The default value is `FALSE`.
`subset`	A numeric vector indicating the subset of the units whose in-sample predications to be provided when `units` is `TRUE`. The default value is `NULL` where the in-sample predictions for each unit will be provided.
`...`	further arguments passed to or from other methods.

Value

`summary.ecoNP` yields an object of class `summary.ecoNP` containing the following elements:

`call`	The call from `ecoNP`.
`n.obs`	The number of units.
`n.draws`	The number of Monte Carlo samples.
`agg.table`	Aggregate posterior estimates of the marginal means of W_1 and W_2 using X and N as weights.
`param.table`	Posterior estimates of model parameters: population mean estimates of W_1 and W_2. If `subset` is specified, only a subset of the population parameters are included.

`W1.table` Unit-level posterior estimates for W_1.

`W2.table` Unit-level posterior estimates for W_2.

This object can be printed by `print.summary.ecoNP`

Author(s)

Kosuke Imai, Department of Politics, Princeton University ⟨kimai@Princeton.Edu⟩, `http://www.princeton.edu/~kimai`; Ying Lu, Institute for Quantitative Social Sciences, Harvard University ⟨ylu@Latte.Harvard.Edu⟩

See Also

`ecoNP`, `predict.eco`

`reg` *Voter Registration in US Southern States*

Description

This data set contains the racial composition, the registration rate, the number of eligible voters as well as the actual observed racial registration rates for every county in four US southern states: Florida, Louisiana, North Carolina, and South Carolina.

Usage

```
data(reg)
```

Format

A data frame containing 5 variables and 275 observations

X	numeric	the fraction of Black voters
Y	numeric	the fraction of voters who registered themselves
N	numeric	the total number of voters in each county
W1	numeric	the actual fraction of Black voters who registered themselves
W2	numeric	the actual fraction of White voters who registered themselves

References

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

`census` *Black Illiteracy Rates in 1910 US Census*

Description

This data set contains the proportion of the residents who are black, the proportion of those who can read, the total population as well as the actual black literacy rate and white literacy rate for 1040 counties in the US. The dataset was originally analyzed by Robison (1950) at the state level. King (1997) recoded the 1910 census at county level. The data set only includes those who are older than 10 years of age.

Usage

```
data(census)
```

Format

A data frame containing 5 variables and 1040 observations

X	numeric	the proportion of Black residents in each county
Y	numeric	the overall literacy rates in each county
N	numeric	the total number of residents in each county
W1	numeric	the actual Black literacy rate
W2	numeric	the actual White literacy rate

References

Robinson, W.S. (1950). "Ecological Correlations and the Behavior of Individuals." *American Sociological Review*, vol. 15, pp.351-357.

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

4 What's New?

version	date	changes
$2.2-1$	09.28.05	nonparametric model with contextual effects added
$2.1-1$	07.06.05	a major revision; added bounds and prediction; added/updated other functionalities
$1.1-1$	06.15.05	add the Metropolis algorithm to sample W
$1.0-1$	12.21.04	first official version; submitted to CRAN
$0.9-1$	09.07.04	first beta version

References

Imai, K. and Lu, Y. (2004). Parametric and nonparametric Bayesian models for ecological inference in 2×2 tables. *Unpublished Manuscript* available at `http://imai.princeton.edu/research/einonpar.html`.