EI: A(n R) Program for Ecological Inference

Gary King [1] Margaret Roberts [2]

October 18, 2010

[1]Albert J.' Weatherhead III University Professor, Harvard University (Institute for Quantitative Social Sciences, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.Harvard.Edu, King@Harvard.Edu, 617-500-7570.

[2]Department of Government, 1737 Cambridge Street, Harvard Univeristy, Cambridge MA 02138

Contents

1 Introduction: Ecological Inference

This program provides a method of inferring individual behavior from aggregate data. It implements the statistical procedures, diagnostics, and graphics from the book, **A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data** (Princeton: Princeton University Press, 1997), by Gary King. Please read the book prior to trying this program (a sample chapter and other related information is available at King's website). Except where indicated, all references to page, section, chapter, table, and figure numbers in this document refer to the book.

Ecological inference, as traditionally defined, is the process of using aggregate (i.e., "ecological") data to infer discrete individual-level relationships of interest when individual-level data are not available. As existing methods usually lead to inaccurate conclusions about the empirical world, the ecological inference problem had been to develop a method that gives accurate answers. Ecological inferences are required in political science research when individual-level surveys are unavailable (e.g., local or comparative electoral politics), unreliable (racial politics), insufficient (political geography), or infeasible (political history). They are also required in numerous areas of major significance in public policy (e.g., for applying the Voting Rights Act) and other academic disciplines ranging from epidemiology and marketing to sociology and quantitative history. Most researchers using aggregate data have encountered some form of the ecological inference problem.

Because the ecological inference problem is caused by the lack of individual-level information, no method of ecological inference, including that introduced in this book and estimated by this program, will produce precisely accurate results in every instance. However, potential difficulties are minimized here by models that include more available information, diagnostics to evaluate when assumptions need to be modified, easy methods of modifying the assumptions, and uncertainty estimates for all quantities of interest. We recommend reviewing Chapter 16 while using this program for actual research.

2 Overview: R-commands to run the application

In this section we describe the basic commands in `EI`. For this purpose, and without loss of generality, we use the running example from the book portrayed in Table 2.3 (page 31). This example uses the fraction of the voting age population who are black (X_i), the fraction turning out to vote (T_i), and the number of voting age people (N_i) in each precinct ($i = 1, \ldots, p$) to infer the fraction of blacks who vote (β_i^b) and the fraction of whites who vote (β_i^w), also in each precinct. (For an extended example of `EI`, refer to the user's guide below.) As only four commands are required to use `EI`, the program can be easily run interactively, or in batch mode as a regular R program. An example of a sequence of these commands are as follows:

```
> dbuf <- ei(t, x, n, 1, 1)
> summary(ei)
> eiread(ei, "betab", "betaw")
> plot(ei, "tomog", "betab", "betaw", "xtfit")
```

In most applications, `plot` and `eiread` would likely be run multiple times with different options chosen, and other commands would be included with these four to read in the data (`t`, `x`, and `n`).

- `ei`: To run the main procedure

  Use the format, `dbuf = ei(t,x,n,1,1)`, which takes three $p \times 1$ vectors as inputs: `t` (e.g. the fraction of the voting age population turning out to vote); `x` (e.g. the fraction of the voting age population who are black); and `n` (e.g. the total number of people in the voting age population). The remaining two inputs are for optional covariates. For the basic model, set them each to 1 for no covariates. The output of this procedure is the list `dbuf` (i.e. a data

buffer). After running `ei`, it is a good idea to save `dbuf` on disk for further analysis. The output data buffer from `ei` includes a large variety of different results useful for understanding the results of the analysis. A minimal set of nonrepetitive information is stored in this list (or data buffer), and a large variety of other information can be easily computed from it. Fortunately, you do not need to know whether the information you request is stored or computed as both are treated the same.

To extract information from the data buffer, three procedures are available:

- `summary`: To obtain a general summary of district-level information

  `summary(dbuf)` will produce aggregate information about the `ei` results. It also includes information on the values specified for the prior.

- `plot`: To graph relevant information

  For graphics, use `plot(dbuf, "name");`, where `dbuf` is the list that is the output of `ei`, and name can be any number of a long list of ready-made graphs. For example, `plot(dbuf,"tomog")`s will plot a tomography graph, and `plot(dbuf,"xt")` will display a scattercross graph. Any number of graphs can be selected in plot for output. For example, `plot(dbuf,"tomog","xt")` will produce both a tomography graph and a scattercross graph.

- `eiread`: To obtain relevant information and numerical results

  For numerical information, use `v <- eiread(dbuf, "name")`, where `v` is the item extracted, `dbuf` is the data buffer output from `ei`, and name can be any of a list of output possibilities. For example, `eiread(dbuf, "betab")` will give a vector of point estimates of β_i^b, `eiread(dbuf, "ci80w")` will give 80% confidence intervals for β_i^w. Any number of items can be selected for output. For example, `eiread(dbuf, "betab", "ci80b")` will produce a list of the point estimates and 80% confidence intervals for β_i^b.

3 User's Guide: An Example

We now show how to use `EI` through a simple example. We use data on voter registration and racial background of people from 268 counties in four Southern U.S. States: Florida, Louisiana, and South Carolina. These are the same data used in Chapter 10 of **A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data** by Gary King. The data include the total voting age population in each county (`n`), the proportion of the population in each county who are black (`x`), and the proportion of the population in each county who are registered to vote (`t`). The proportion of whites registered can be computed by (`1-x`). You can load the data into R using the command

```
> library(ei)
> data(matproii)
> attach(matproii)
```

The statistical goal is to estimate the fraction of blacks registered (β_i^b) and the fraction of whites registered (β_i^w) in each county. These quantities of interest are generally unknown. However, `EI` also includes an option that allows the user to assess the reliability of the `EI` algorithm when the true quantities of interest are known. To illustrate the use of this "truth" option, the data also include the true fractions of registered blacks (`tb`) and the true fraction of registered whites (`tw`).

3.1 The Basic EI Algorithm

To begin, we perform ecological inference by calling the function `ei`. `ei` first computes and maximizes a likelihood distribution based on the inputed data. Then it estimates county-level quantities of interest based on the possible values, or "bounds", for each county and the overall likelihood distribution.

To run this algorithm, at minimum we need to have three vectors. Two vectors, `t` and `x`, contain the aggregate known information about the counties of interest. In this example, `t` is the proportion of voters registered in each county, and `x` is the proportion of blacks in each county. These two vectors contain aggregate information, since we are interested in the proportion of voters in each country who are black. The last vector we need is `n`, the number of people of interest in each county. In this example, `n` is the number of people of voting age.

We proceed by performing ecological inference without covariates on the data:

```
> dbuf = ei(t, x, n, 1, 1)
```

The `1,1` in the `ei` call specify that no covariates should be included in estimation. To include a covariate on β_i^b simply replace the first `1` with a covariate vector for β_i^b. Similarly, to include a cpvariate on β_i^w, replace the second `1` with a covariate vector for β_i^w.

Next, we use `summary(dbuf)` to obtain general information about the estimation.

```
> summary(dbuf)

Erho = 0.5
Esigma = 0.5
Ebeta = 0.5
N = 268
Resamp = 2

Maximum likelihood results in scale of estimation (and se's)
      Bb0       Bw0       sigB       sigW       rho Zb Zw
 1.267095 1.9348579 -1.1150909 -1.3271242 1.6051356  0  0
 0.278658 0.2764883  0.2096715  0.1616891 0.3096247  0  0

Untruncated psi's
        BB       BW        SB        SW       RHO
 0.9798303 1.142635 0.3374653 0.2727997 0.9215691

Truncated psi's (ultimate scale)
        BB        BW        SB        SW       RHO
 0.6169952 0.8271868 0.1995737 0.1412539 0.7798519

Aggregate Bounds
          betab     betaw
lower 0.2125903 0.7025925
upper 0.9754242 0.9200036

Estimates of Aggregate Quantities of Interest
        mean          sd
Bb 0.5655017 0.017029042
Bw 0.8194223 0.004853354
```

The `summary` function provides basic information about the ecological inference estimation, the maximum likelihood estimates on three scales, and an overall summary of the quantities of interest. First, it reports the values of the priors used in ecological inference (`Erho, Esigma, Ebeta`). It also reports the number of counties in the dataset (`N`), as well as the number of importance sampling iterations required to produce estimates of the quantities of interest (`resamp`).

`summary` also produces information about the maximum likelihood estimates. First, it provides the maximum likelihood estimates in the scale of estimation. Next, it provides the values of the

MLEs when they are transformed into an untruncated bivariate normal distribution. These estimates provide information about the location of the highest density of estimates for the proportion of blacks who are registered to vote and the proportion of whites who are registered to vote. Last, it provides the values of the MLEs on the ultimate truncated bivariate normal scale. These estimates are an unweighted average of the fraction of blacks and whites registered to vote over all the counties in the sample. In this example, the `EI` algorithm predicts that the unweighted average of the proportion of blacks registered to vote across counties is 0.62 and the unweighted average of the proportion of whites registered to vote is 0.83.

Finally, `summary` produces information on aggregrate quantities of interest. The aggregate bounds are the mean of the bounds on the proportions of black and white voters, weighted by the population in each county. The aggregate quantities of interest are the weighted mean of the proportion of registered blacks and the proportion of registered whites in each county. In this example the weighted average proportion of blacks who are registered is 0.57, and the weighted average proportion of whites who are registered is 0.82.

3.2 Extracting Quantities of Interest

`eiread` extracts quantities of interest in a list format from the object `dbuf`. For example,

```
> bb.out <- eiread(dbuf, "betab", "sbetab")
```

extracts the point estimates and estimated standard deviations for β_i^b, the estimates of the proportion of registered blacks in each county. The user can then use `bb.out$betab` to obtain a vector of the point estimates for β_i^b, and `bb.out$sbetab` to obtain a vector of the standard devations for β_i^b.

`eiread()` takes any number of arguments to extract any number of quantities of interest from the data buffer. `?eiread` can be used to find a list of quantities of interest that are available in `EI`. Among the most useful arguments are `"betab"` and `"betaw"`, which report the point estimates; `"sbetab"` and `"sbetaw"`, which report the standard deviations of the point estimates; and `"CI80b` and `"CI80w"`, which report the confidence intervals of the point estimates.

3.3 Plotting in EI

Plotting `EI` output is extremely useful for understanding the results of the ecological inference algorithm and diagnosing problems with estimation. First, we graph a tomography plot of the data (Figure 1).

```
> plot(dbuf, "tomog")
```

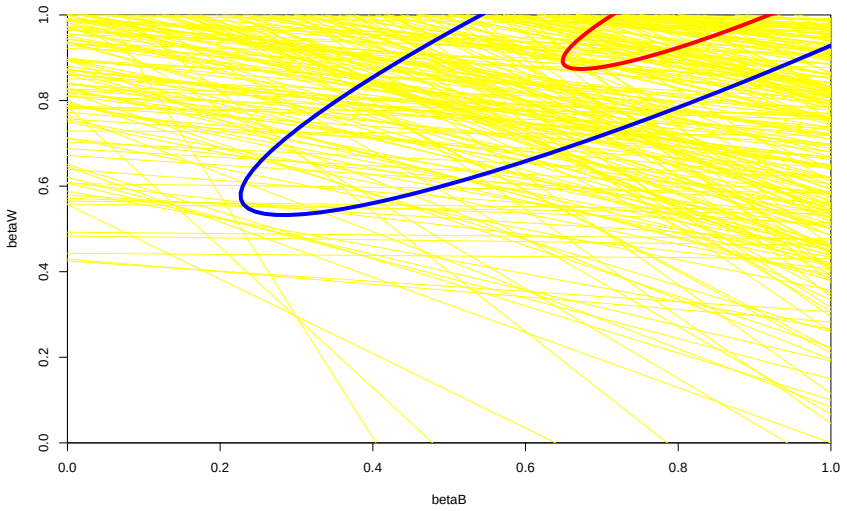

Figure 1: Tomography Plot with ML Contours

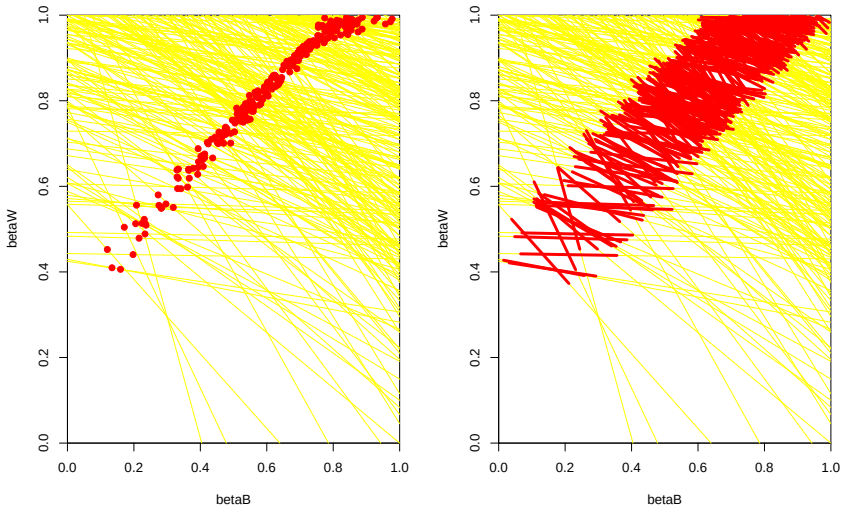

Figure 2: Tomography Plots: Point Estimates and Confidence Intervals

Each line on the map represents the possible values for β_i^b and β_i^w for one county. The contour lines identify the portion of the lines that have the highest probability of containing the true estimates of β_i^b and β_i^w. These contour lines provide information about the overall pattern of registration. Note that the area with highest probability is in the upper right-hand corner, where the proportion of whites registered is between 0.75 and 1 and the proportion of blacks registered is between 0.5 and 1. Further, we see that the lines are clustered in the upper half of the figure, indicating that the possible values of β_i^w will have a lower variance than the possible values of β_i^b.

Figure 2 is a double plot of the point estimates and their confidence intervals. To compute this, we call plot to generate two graphs: a tomography plot with the point estimates generated from the algorithm, and a tomography plot with 80% confidence intervals on the point estimate.

```
> plot(dbuf, "tomogE", "tomogCI")
```

This plot is useful to visualize the actual estimates and confidence intervals for each county. We can see that the point estimates and confidence intervals are clustered in the same area as the contours from the previous plot. Further, the point estimates and confidence intervals only fall on the lines that indicate the possible values of β_i^b and β_i^w.

Figure 3 shows plots that indicate the distribution of β_i^b and β_i^w. To produce these plots, run:

```
> plot(dbuf, "betab", "betaw")
```

Density plots are useful to visualize the location and uncertainty of the point estimates. The green line represents the density of the simulated point estimates, and the black tick marks are a rug plot of the point estimates,s β_i^b and β_i^w. You can see that the variance of the point estimates from β_i^b is much higher than the variance of point estimates from β_i^w.

Figure 4 portrays the results of the `EI` algorithm by plotting the proportion of blacks in each country by the proportion of registered voters in each county. To produce this plot, use:

```
> plot(dbuf, "xtfit")
```

The circles around each of the points in this plot are proportional to the population of each county. The graph represents the likelihood estimates by plotting the expected number of registered voters given the proportion of blacks in each county, represented by the yellow line. The red lines in the plot are the 80% confidence interval around the regression line. The higher uncertainty in the estimates of black registration can be seen by the absence of points on the right hand side of the graph and the larger confidence interval on the right hand side of the graph, where the proportion of blacks in the county is relatively large.

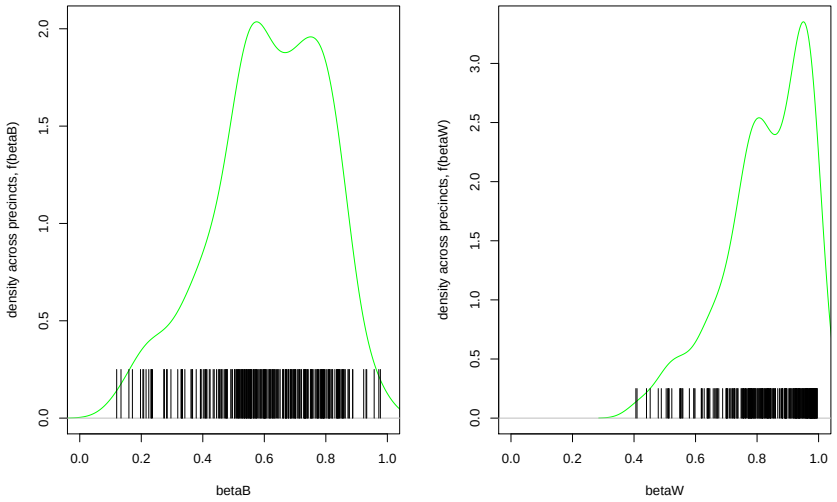

Figure 3: Plots with Distribution of Point Estimates for β_i^b and β_i^w

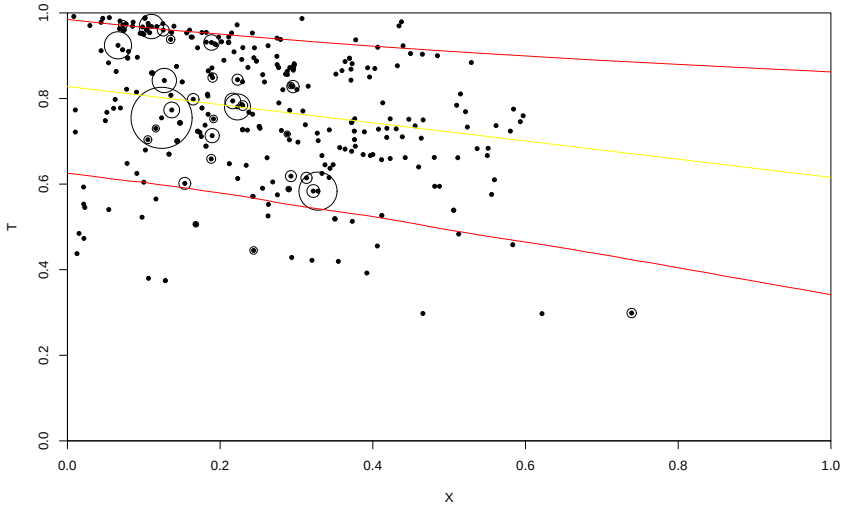

Figure 4: X and T Scatterplot with E(T|X) and 80% CIs

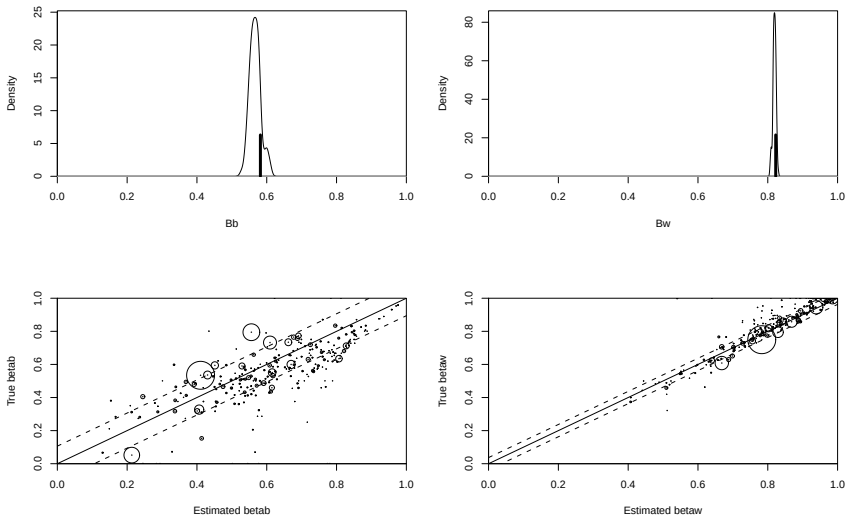

Figure 5: Comparing Estimates to the Truth at the County Level

Finally, if we have data on the true proportions of registered blacks and registered whites in each county, as we do in this dataset, we can use plots in `EI` to assess how well the algorithm works on the given data. To do this, rerun the `ei` algorithm, adding the truth vectors, `tb` and `tw` as the truth argument.

```
> truth = cbind(tb, tw)
> dbuf = ei(t, x, n, 1, 1, truth = truth)
```

Then use plot to compare the estimates of the `EI` algorithm to the true proportions of white and black registered voters in each county.

```
> plot(dbuf, "truth")
```

The "truth" plot (Figure 5) has four components. The top two figures have the posterior densities of the aggregate quantities of interest B^b and B^w. These densities tell us in what range the algorithm predicts that the point estimates lie. For example, the density of B^b is wider than the density of B^w, indicating that our estimates of B^w lie in a smaller range than than our estimates of B^b. The true aggregate values, computed from the `truth` data we inputed, are indicated by the vertical bars. The fact that the true B^b and B^w are within the densities we computed confirms that the model is did a good job at predicting the true proportions of registered white and black voters.

The bottom two figures in Figure 5 plot the estimated values of β_i^b and β_i^w against the true values. The size of each of the circles plotted is proportional to the number of blacks in the county in the first graph and the number of whites in each county in the second graph. If the estimated values were exactly equal to the true values, all of the points would be on the 45° line. Because the points fall quite close to the 45° and do not deviate from the line in a systematic way, we can see that the `EI` algorithm predicts the point estimates quite well.

4 Reference to EI Functions

4.1 `ei`: Ecological Inference Estimation

Description

`ei` is the main command in the package _ei_. It gives observation-level estimates (and various related statistics) of β_i^b and β_i^w given variables T_i and X_i $(i = 1, ..., n)$ in this accounting identity: $T_i = \beta_i^b * X_i + \beta_i^w * (1 - X_i)$. Results are stored in and `ei` object, that can be read with `summary()` or `eiread()` and graphed in `plot()`.

Usage

```
ei(t, x, n, Zb, Zw, erho=.5, esigma=.5, ebeta=0, ealphab=NA,
 ealphaw=NA, truth=NA)
```

Arguments

`t`	p x 1 vector
`x`	p x 1 vector
`n`	p x 1 vector
`Zb`	vector of 1's for no covariates or a p x k^b matrix of covariates
`Zw`	vector of 1's for no covariates or a p x k^w matrix of covariates
`erho`	The standard deviation of the normal prior on ϕ_5 for the correlation. Default=.05.
`esigma`	The standard deviation of an underlying normal distribution, from which a half normal is constructed as a prior for both $\breve{\sigma}_b$ and $\breve{\sigma}_w$.
`ebeta`	Standard deviation of the "flat normal" prior on $\breve{B}^b$ and $\breve{B}^w$. The flat normal prior is uniform within the unit square and dropping outside the square according to the normal distribution. Set to zero for no prior (default). Setting to positive values probabilitically keeps the estimated mode within the unit square. 0.25 is a resonable value to experiment with first.
`ealphab`	cols(Zb) x 2 matrix of means (in the first column) and standard deviations (in the second) of an independent normal prior distribution on elements of α^b. If you specify Zb, you should probably specify a prior, at least with mean zero and some variance (default is no prior). (See Equation 9.2, page 170, to interpret α^b).
`ealphaw`	cols(Zw) x 2 matrix of means (in the first column) and standard deviations (in the second) of an independent normal prior distribution on elements of α^w. If you specify Zw, you should probably specify a prior, at least with mean zero and some variance (default is no prior). (See Equation 9.2, page 170, to interpret α^w).

Details

The EI algorithm is run using the `ei` command. A summary of the results can be seen graphically using `plot(eiobject)` or numerically using `summary(eiobject)`. Quantities of interest can be calculated using `eiread(eiobject)`.

References

Gary King (1997). A Solution to the Ecological Inference Problem. Princeton: Princeton University Press.

Examples

```
data(sample)
attach(sample)
dbuf <- ei(t,x,n,1,1)
summary(dbuf)
```

4.2 `eiread`: Quantities of Interest from Ecological Inference Estimation

Description

`eiread` is the command that pulls quantities of interest from the `ei` object. The command returns a list of quantities of interest requested by the user.

Usage

```
eiread(ei.object, ...)
```

Arguments

`ei.object`	An `ei` object from the function `ei`.
`...`	A list of quantities of interest for `eiread()` to return. See values below.

Value

`betab`	p x 1 point estimate of β_i^b based on its mean posterior. See section 8.2
`betaw`	p x 1 point estimate of β_i^w based on its mean posterior. See section 8.2
`sbetab`	p x 1 standard error for the stimate of β_i^b, based on the standard deviation of its posterior. See section 8.2
`sbetaw`	p x 1 standard error for the stimate of β_i^w, based on the standard deviation of its posterior. See section 8.2
`phi`	Maximum posterior estimates of the CML
`psisims`	Matrix of random simulations of ψ. See section 8.2
`bounds`	p x 4: bounds on β_i^b and β_i^w, lowerB ~ upperB ~ lowerW ~ upperW. See Chapter 5.
`abounds`	2 x 2: aggregate bounds rows:lower, upper; columns: betab, betaw. See Chapter 5.
`aggs`	Simulations of district-level quantities of interest $\hat{B^b}$ and $\hat{B^w}$. See Section 8.3.
`maggs`	Point estimate of 2 district-level parameters, $\hat{B^b}$ and $\hat{B^w}$ based on the mean of aggs. See Section 8.3.
`VCaggs`	Variance matrix of 2 district-level parameters, $\hat{B^b}$ and $\hat{B^w}$. See Section 8.3.
`CI80b`	p x 2: lower~upper 80% confidence intervals for β_i^b. See section 8.2.
`CI80w`	p x 2: lower~upper 80% confidence intervals for β_i^w. See section 8.2.
`eaggbias`	Regressions of estimated β_i^b and β_i^w on a constant term and X_i.
`goodman`	Goodman's Regression. See Section 3.1

References

Gary King (1997). A Solution to the Ecological Inference Problem. Princeton: Princeton University Press.

Examples

```
data(sample)
attach(sample)
dbuf <- ei(t,x,n,1,1)
eiread(dbuf, "betab", "betaw")
```

4.3 `plot.ei`: Plotting Ecological Inference Estimates

Description

'plot' method for the class '"ei"'.

Usage

```
plot.ei(ei.object, ...)
```

Arguments

`ei.object`	An ei.object from the function ei.
`...`	A list of options to return in graphs. See values below.

Value

`tomogD`	Tomography plot with the data only. See Figure 5.1, page 81.
`tomog`	Tomography plot with ML contours. See Figure 10.2, page 204.
`tomogCI`	Tomography plot with 80% confidence intervals. Confidence intervals appear on the screen in red with the remainder of the tomography line in yellow. The confidence interval portion is also printed thicker than the rest of the line. See Figure 9.5, page 179.
`tomogCI95`	Tomography plot with 95% confidence intervals. Confidence intervals appear on the screen in red with the remainder of the tomography line in yellow. The confidence interval portion is also printed thicker than the rest of the line. See Figure 9.5, page 179.
`tomogE`	Tomography plot with estimated mean posterior β_i^b and β_i^w points.
`tomogP`	Tomography plot with mean posterior contours.
`betab`	Density estimate (i.e., a smooth version of a histogram) of point estimates of β_i^b's with whiskers.
`betaw`	Density estimate (i.e., a smooth version of a histogram) of point estimates of β_i^w's with whiskers.
`xt`	Basic X_i by T_i scatterplot.
`xtc`	Basic X_i by T_i scatterplot with circles sized proportional to N_i.
`xtfit`	X_i by T_i plot with estimated $E(T_i\|X_i)$ and conditional 80% confidence intervals. See Figure 10.3, page 206.
`xtfitg`	xtfit with Goodman's regression line superimposed.
`estsims`	All the simulated β_i^b's by all the simulated β_i^w's. The simulations should take roughly the same shape of the mean posterior contours, except for those sampled from outlier tomography lines.
`boundXb`	X_i by the bounds on β_i^b (each precinct appears as one vertical line), see the lines in the left graph in Figure 13.2, page 238.
`boundXw`	X_i by the bounds on β_i^w (each precinct appears as one vertical line), see the lines in the right graph in Figure 13.2, page 238.
`truth`	Compares truth to estimates at the district and precinct-level. Requires truth in the eiobject. See Figures 10.4 (page 208) and 10.5 (page 210).

References

Gary King (1997). A Solution to the Ecological Inference Problem. Princeton: Princeton University Press.

Examples

```
data(sample)
attach(sample)
dbuf <- ei(t,x,n,1,1)
plot(dbuf, "tomog", "betab", "betaw", "xtfit")
```

References

[1] Gary King, *A solution to the Ecological Inference Problem: Reconstructing Individual Behaviour from Aggregate Data*, Princeton University Press (1997).

[2] Useful documentation also avalaible at `http://GKing.Harvard.Edu`.

[3] For R language see `http://www.r-project.org`.

[4] Venables, W.N.,and Ripley, B.D., *Statistics and Computing*, Springer (2002).