Ecological Inference: Reconstructing Individual behavior from Aggregate Data; software available at `http://gking.harvard.edu/ei/`, free of charge and open-source (under the terms of the GNU GPL, v. 2).

Gary King [1]

August 10, 2010

[1]David Florence Professor of Government, Harvard University (Institute for Quantitative Social Sciences, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; `http://GKing.Harvard.Edu`, `King@Harvard.Edu`, (617) 495-2027).

Contents

1 Acknowlegments

The code, which was originally written in Gauss, has been translated into R by Elena Villalon who also generated the plots that are presented in this document.

2 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. I recommend reviewing Chapter 16 while using this program for actual research.

3 Overview: R-commands to run the application

As only four commands are required to use $\mathfrak{E}I$ the program can be easily run interactively, or in batch mode as a regular R program. Each command may also be used with many optional globals and subcommands. An example of these commands are as follows:

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).

These four commands described in this section through a very simple use of $\mathfrak{E}I$. (Refer to the reference section below for further details and more sophisticated uses. That section also includes sample data you can run with simple examples.) For this purpose, and without loss of generality, I 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.

- 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., data buffer). After running ei, it is a good idea to save dbuf on disk for futher 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, two procedures are available:

- eigraph or plot: To graph relevant information

  For graphics, use plot(dbuf, "name") or eigraph(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"); to print a tomography graph, or eigraph(dbuf,"xt"); to display a scattercross graph.

- analysis: To obtain relevant information and numerical results

  For numerical information, use v <- eiread(dbuf, "name"), where v is the item extracted, dbuf is the list that is the output of ei, and name can be any of a long list of output possibilities. For example, use betab for a vector of point estimates of β_i^b, ci80w for 80% confidence intervals for β_i^w. summary() can be used to print a summary of district-level estimates and information.

4 Reference to EI's 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: Princton University Press.

Examples

```
print("!!!!!!!")
```

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: Princton University Press.

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: Princton University Press.

References

[1] Gary King, *A solution to the Echological 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).