Ecological inference with R: the *ecoreg* package

Version 0.2

Christopher Jackson
MRC Biostatistics Unit, Cambridge
chris.jackson@mrc-bsu.cam.ac.uk *

March 2006

Abstract

In typical small-area studies of health and environment we wish to make inference on the relationship between individual-level quantities using aggregate, or ecological, data. Such ecological inference is often subject to bias and imprecision, due to the lack of individual-level information in the data. Simple regressions of area-level mean outcomes on area-level mean exposures are usually biased. To alleviate bias, the within-area distribution of exposures should be accounted for. The *ecoreg* package can be used to fit this class of models for ecological inference from aggregate data. In addition, full outcome and covariate information from a survey of individuals within the areas can be used to improve bias and precision. *ecoreg* can be used in this way to analyse ecological and individual data simultaneously, using *hierarchical related regression*.

1 Ecological inference

Ecological studies analyse data defined at a group level, but aim to make inferences about the individuals within the groups. To make reliable individual-level inferences from these studies, a number of problems must be overcome. One crucial difficulty is that the group-level exposure-response relationship may not reflect the individual-level relationship, a problem known as *ecological bias*, or the *ecological fallacy*. See, for example, [1, 2, 3, 4] for discussion of these issues.

Denote the outcome count in area i, with population N_i, by y_i. To model y_i in terms of exposures measured as aggregate-level summaries, the usual model is a simple binomial or Poisson regression on the area-level covariate means $\overline{z}_i$. With binary covariates, $\overline{z}_i$ is the proportion exposed over the area. However, this only models the relationship between the *aggregate* exposures and outcomes. It is only justified as a method of estimating individual-level relationships if all individuals in the area have the same covariate value, or there is the same exposure-response relationship at the individual and aggregate levels, which is generally only true for linear models. With non-linear models, such as Binomial or Poisson models, using the same model form at both levels will lead to ecological bias [1].

*Written while a member of the BIAS project at the Department of Epidemiology and Public Health, Imperial College London

Despite these problems, aggregate data *can* provide information about individual-level relationships. As the ratio of the between-area to the within-area variability of the exposure increases, the aggregate data summarise the true distribution of the exposure more accurately, and contain more information about the true individual-level exposure-outcome relationship. With sufficient exposure information, and with *correctly-specified* models for the mean outcome [1, 5, 6], ecological bias can be reduced to negligible levels.

In general, successful ecological inference requires samples of individual-level data within areas. Individual exposure data are usually required to reduce ecological bias by accounting for the within-area variability of exposures. But further improvements can also be made by using samples of exposures and *outcomes* for selected individuals, as discussed by Wakefield [7] for 2×2 tables, and more generally by Jackson *et al.*[8].

2 Models for aggregate and individual data

The models described here are also described in the papers by Jackson *et al.*[8] [9].

2.1 Individual data

We begin by specifying the form of the relationship between the individual-level risk of the binary outcome and the covariates. If individual-level exposure and outcome data are available, this is used to model them. It will also be used as the basis for an equivalent model for the aggregate data, as described in the next section. The risk p_{ij} of the individual-level outcome y_{ij} for the jth individual in area i is assumed to be a logit-linear function of the covariates. The most general model we consider is

$$\text{logit}(p_{ij}) = \mu_i + \sum_r \alpha_r x_{ir} + \sum_r \beta_r z_{ijr} + \gamma_{s_{ij}} \tag{1}$$

where x_{ir} are group-level covariates, and z_{ijr} are individual-level covariates. The group-level covariates may include descriptions of the socio-economic status of the area, or the health service provision in the area. Individual-level covariates might comprise individual behaviours such as smoking, demographics such as ethnicity, or individual indicators of wealth and social class. Individuals may be influenced by the overall average exposure in the area, in addition to their own, so that the group-level variables may include the means of certain individual-level variables. γ_s represents an additional contribution to the baseline risk for an individual occupying one of several strata s, usually defined by age and sex.

The baseline risk μ_i may be fixed at μ or considered as a random effect with some distribution across areas. This can account for any remaining overdispersion and heterogeneity between areas, after adjusting for observed area-level variables. A random effect also allows the borrowing of information across areas and can stabilise estimation from areas with small populations [10].

2.2 Aggregate data

2.2.1 Marginal model

Suppose the area-level exposures have been estimated from a survey. For example, in the UK census, aggregate data on social class and education are calculated using a 10% sample to maintain confidentiality. The proportion of smokers in the area might also be estimated from sales figures instead of

a census. Then, individual outcomes can be assumed to be independent and identically distributed, with risk equal to some marginal "group-level risk" p_i. We assume

$$y_i \sim Bin(N_i, p_i), \tag{2}$$

where p_i is determined by integrating the individual-level model over the joint within-area distribution of covariates [1, 5]. Thus, p_i is the average risk for an individual in group i.

$$p_i = \int p_{ij}(\mathbf{x}) f_i(\mathbf{x}) d\mathbf{x} = E_{\mathbf{x}}(p_{ij}(\mathbf{x})|i) \tag{3}$$

For a *single binary covariate*, observing the proportion exposed gives us enough information to estimate a binomial within-area distribution. For *continuous covariates* the mean is not sufficient to estimate the within-area distribution, and we generally need samples of individual covariate data to be able to estimate the within-area variability. For *multiple binary covariates* the joint distribution is estimated by the cross-classification of individuals between covariates. Typical census data do not usually have this cross-classification, and we need individual data to estimate it.

2.2.2 Conditional model

An alternative to the marginal model was proposed by Wakefield [7]. The binomial model (2) is based on the assumption that each individual in group i has an identical marginal probability p_i of outcome, integrating over their unknown exposures. If the binary exposure is known for *all* the individuals in the area from a full population census, but not necessarily coupled with exposures for the same individuals, then these should be conditioned on, leading to a likelihood based on a convolution of binomial distributions. The exact likelihood is related to the extended (or non-central) hypergeometric distribution, however, Wakefield describes a more computationally convenient normal approximation.

2.2.3 Binary covariates

For clarity we demonstrate the marginal model for 3 binary covariates, however, the framework extends immediately to any number of binary covariates. The integral to obtain the group-level risk is equivalent to a sum. Each individual falls into one of $S \times 2^3$ categories, defined by the distinct combinations of the 3 covariates and the S age-sex strata, and indexed by k. Let ϕ_{ik} be the probability that an individual occupies category k. Let q_{ik} be the probability of the individual outcome conditionally on occupying category k. Rewriting the index k as $\{k_1, k_2, k_3, s\}$, where k_r (0 or 1) indicates the presence or absence of covariate $r = 1, \ldots, 3$.

$$p_i = \sum_k \phi_{ik} q_{ik} = \sum_{k_1,k_2,k_3,s} \phi_{\{i,k_1,k_2,k_3,s\}} q_{\{i,k_1,k_2,k_3,s\}}. \tag{4}$$

The outcome model conditionally on the unobserved category is

$$\text{logit}(q_{\{i,k_1,k_2,k_3,s\}}) = \mu_i + \text{logit}(e_s) + \sum_r \alpha_r x_{ir} + \sum_r k_r \beta_r \tag{5}$$

The effect β_r of the binary individual-level covariate r only enters into this equation when the covariate r is present. This is a generalisation of the model presented for two covariates by Lasserre *et*

al.[11], except that the outcome is assumed to be non-rare and binomial instead of Poisson. $\text{logit}(e_s)$ is a fixed offset, where e_s is the risk of the outcome in stratum s, estimated from national population data. This is similar in spirit to "indirect standardisation", see, for example, [12]. If population and outcome totals are available for strata within areas, then we could generalise this to a separate binomial model for each stratum within area, with coefficients for the other covariates shared between the models, assuming no stratum-covariate interactions. It is also important to check whether *exposures* are correlated with strata. If not accounted for, this can lead to "mutual standardisation bias" [13].

We normally replace ϕ_{ik} in (4) by an estimate $\hat{\phi}_{ik}$. Ideally this *probability* would be estimated by the *proportion* of individuals in area i occupying category k. But typical census data are not sufficient to know the complete cross-classification of individuals between all of these covariate and strata categories. Usually, our only information is the marginal proportions of single covariates, for example, the proportion $\phi_i^{(1)}$ of individuals who are economically inactive. If we assume that the covariates are independent, then ϕ_{ik} can be estimated by the product of the proportions of individuals occupying each marginal category defining k. But generally, socioeconomic indicators, such as unemployment and social class, are highly correlated. [11] demonstrated that, in a typical case, bias is negligible when the joint covariate distribution is estimated by the product of the two marginal distributions, even when the covariates are correlated. However, we may wish to study more than two covariates.

To estimate ϕ_{ik} we can often use a combination of marginal proportions $\overline{z}_{ir}$ and individual covariate data z_{ijr}. In the context of the UK census, for example, these correspond to district-level aggregate data and the Samples of Anonymised Records. Let C_{ik} be the number of individuals in area i in this individual-level dataset occupying category k, computed from the z_{ijr}. We use the following two principles. Firstly, the estimated $\hat{\phi}_{ik}$ corresponding to categories k in which binary covariate r is 1 must sum to z_{ir}. Secondly, the ratio of estimates $\hat{\phi}_{i_k}/\hat{\phi}_{i_l}$ must be the same as the ratio C_{ik}/C_{il}. This gives, for example, where R is the set of categories in which covariate r is 1,

$$\hat{\phi}_{ik} = C_{ik}\overline{z}_{ir} / \sum_{l \in R} C_{il} \tag{6}$$

2.2.4 Continuous covariates

Suppose now the individual-level model (1) depends only on an intercept and one continuous individual-level covariate x_{ij}: $\text{logit}(p_{ij}) = \mu_i + \beta x_{ij}$. The ecological data consist of the within-area mean m_i of x_{ij}. In some cases, as well as the within-area mean, we may also have an estimate s_i^2 of the within-area variance of x_{ij}, for example, from geographical modelling of an environmental exposure surface (e.g. Best *et al.*[14]). Then we suppose that these exposures are normally distributed, with $x_{ij} \sim N(m_i, s_i^2)$. If an exposure is not naturally normally distributed, it can often be transformed to normality. We can then calculate the area-specific risks (3) by integrating over f_i, here the density function of the normal distribution.

$$p_i \quad = \quad \int \text{expit}(\mu_i + \beta x) f_i(x) dx \tag{7}$$

Our assumed underlying model for $p_{ij}(x)$ is logit-linear. In this case, the integral is not available in closed form. However, if we approximate the logit by a probit link function, then (7) evaluates to

$$p_i \quad = \quad \text{expit}\left\{(1 + c^2\beta^2 s_i^2)^{-1/2}(\mu_i + \beta m_i)\right\} \tag{8}$$

where $c = 16\sqrt{3}/(15\pi)$ (Salway and Wakefield [15]).

If, instead, we were using a Poisson model for a rare outcome, $y_i \sim \text{Pois}(N_i p_i)$, and a log-linear individual-level model $\log(p_{ij}) = \mu_i + \beta x_{ij}$, then the integrals can be evaluated explicitly without an approximation. Instead of (8), we would have (Richardson *et al.*[1])

$$p_i = \exp\left\{\mu_i + \beta m_i + \frac{\beta^2 s_i^2}{2}\right\} \tag{9}$$

These methods generalise to multiple jointly normally-distributed covariates.

2.2.5 Semi-parametric approach

We have described a fully parametric model for ecological inference. Prentice and Sheppard [6] described an alternative semi-parametric approach based on estimating functions. This is often simply called the *aggregate data* method. Suppose a sample of covariates (binary or continuous) are available from a subset of n_i individuals, but not the corresponding outcomes on the same individuals. Broadly, the mean and variance of the total disease count y_i are calculated in terms of an *aggregate* risk $\frac{1}{n_i}\sum_j p_{ij}$. p_{ij} is the risk for individual j in area i, conditionally on their covariate values. This method is not implemented in the *ecoreg* R package. This approach does not require a within-area distribution to be specified for the covariates. It requires samples of covariate data as an explicit part of the model. On the other hand, the parametric approaches described above, and implemented in *ecoreg*, require individual covariate data implicitly to estimate an appropriate within-area distribution.

2.3 Combining aggregate and individual data

To summarise the model for the ecological data, we have a binomial model (2) for the area-level outcome y_i. The corresponding area-level risk p_i is calculated explicitly in terms of the transformed group baseline risk μ_i, the individual-level covariate effects, and the within-area distributions of the covariates.

It is easy to extend this model to include information from a sample of individuals in each area whose outcomes and exposures are known. We simply model the risk of the individual level outcome with the logistic regression (1). Then the covariate effects α and β and the intercept μ_i are shared by the models for both the aggregate and individual-level data. Thus, we can fit a joint model which combines the information from the two sources of data. This is termed *hierarchical related regression*.

Note that it is not necessary to have individual data within all of the areas i. In practice, sample survey data will be available from varying numbers of individuals between areas.

3 Using *ecoreg*

The *ecoreg* package implements a fairly general case of the models described in Section 2. We assume you have already downloaded and installed the *ecoreg* package. To apply these methods, you should have one or both of

- An aggregate dataset with one record for each aggregate group, for example a geographical area, or a stratum within area, for example from a population census. This contains aggregate

outcomes and exposures, and optionally some indication of the within-area distribution of the exposures.

- An individual-level dataset, for example from a sample survey study. There need not be the same number of individuals per area, and there may be some areas in the aggregate dataset with no individuals. This contains full individual-level covariate and outcome data.

First we load the *ecoreg* package into the working R session.

```
> library(ecoreg)
```

The main R function in *ecoreg* is called `eco`. This is used to fit a model to one of these datasets, or a combination of the two. The R help page for *eco* fully describes each of the function's arguments.

3.1 Example

The use of `eco` is illustrated here using a simulated dataset. We simulate aggregate data consisting of 50 groups of 100 individuals. Two contextual covariates (labelled deprivation and mean income) are generated as standard normal variables. Two covariates which are binary at the individual level, and available at the aggregate level as the proportion of non-white individuals and smokers in each area, are generated from uniform distributions. The data frame `sim.df` contains the ecological covariate data.

```
> ng <- 50
> N <- rep(100, ng)
> set.seed(31412)
> ctx <- cbind(deprivation = rnorm(ng), mean.income = rnorm(ng))
> phi <- cbind(nonwhite = runif(ng), smoke = runif(ng))
> sim.df <- as.data.frame(cbind(ctx, phi))
> sim.df[1:5, ]

  deprivation  mean.income  nonwhite     smoke
1  0.81251444 -0.861417267 0.6394023 0.8004601
2 -1.02897943 -1.687661091 0.8664181 0.8016536
3 -0.04496483 -0.003495713 0.1839627 0.1202783
4  0.08924747  1.141115986 0.6579497 0.4494347
5 -1.78817352 -1.094355893 0.3637025 0.3525199
```

A disease outcome with approximate 5% baseline prevalence, and odds ratios of 1.01, 1.02, 1.5 and 2 respectively for the four covariates, is now simulated. The function `sim.eco` is provided to simulate ecological outcome data and individual sample data, in terms of known covariates, baseline risks and odds ratios.

```
> mu <- qlogis(0.05)
> alpha.c <- log(c(1.01, 1.02))
> alpha <- log(c(1.5, 2))
> sim1 <- sim.eco(N, ctx = ~deprivation + mean.income,
+     binary = ~nonwhite + smoke, data = sim.df,
```

```
+     mu = mu, alpha.c = alpha.c, alpha = alpha,
+     isam = 10)
> sim1$y[1:5]

[1] 14 17  7  8  6

> aggdata <- as.data.frame(cbind(y = sim1$y, N = N,
+     sim.df))
```

Format of aggregate dataset We have now got an aggregate dataset `aggdata` suitable for use with the `eco` function. Generally, the aggregate dataset should be a data frame with a row corresponding to an area or group. It should have at mininum an outcome variable. This is available either as a proportion of individuals with the outcome, or the number of events and the population at risk in the area. In addition, any number of aggregate covariates can be specified. Often these will be binary covariates, expressed as proportions over the area. For covariates that are continuous at the individual level, these should be specified as within-area means, and if possible, within-area variances, after transformation to an approximate normal distribution. For example, we print the first ten rows of the `aggdata` data.

```
> aggdata[1:5, ]

   y   N deprivation  mean.income  nonwhite     smoke
1 14 100  0.81251444 -0.861417267 0.6394023 0.8004601
2 17 100 -1.02897943 -1.687661091 0.8664181 0.8016536
3  7 100 -0.04496483 -0.003495713 0.1839627 0.1202783
4  8 100  0.08924747  1.141115986 0.6579497 0.4494347
5  6 100 -1.78817352 -1.094355893 0.3637025 0.3525199
```

The number of individuals with the disease, the population of the area, the deprivation index, the mean income, the proportion of non-white individuals, the proportion of smokers, and the mean and standard deviation of the pollution exposure are labelled `y`, `N`, `deprivation`, `mean.income`, `nonwhite` and `smoke`, respectively.

The return value of `sim.eco` has a component `y` containing the ecological outcome data (the number of individuals in each area with the outcome), and a component `idata` containing the individual sample data. Here we have specified `isam=10` in the call to `sim.eco`, producing an individual sample dataset with 10 individuals for each of the 50 areas.

```
> indivdata <- sim1$idata
```

Format of individual dataset We have now got an individual dataset `indivdata` suitable for use with the *ecoreg* package. The individual dataset should be a data frame with each row corresponding to an individual. Variables may include a binary outcome and any number of covariates. For example, the first 15 rows of `indivdata` are illustrated.

```
> indivdata[1:15, ]
```

```
   group y deprivation mean.income nonwhite smoke
1      1 0   0.8125144  -0.8614173        0     1
2      1 0   0.8125144  -0.8614173        0     1
3      1 0   0.8125144  -0.8614173        1     1
4      1 1   0.8125144  -0.8614173        1     1
5      1 0   0.8125144  -0.8614173        1     1
6      1 0   0.8125144  -0.8614173        0     0
7      1 1   0.8125144  -0.8614173        1     1
8      1 0   0.8125144  -0.8614173        1     1
9      1 0   0.8125144  -0.8614173        1     1
10     1 0   0.8125144  -0.8614173        1     1
11     2 0  -1.0289794  -1.6876611        1     1
12     2 1  -1.0289794  -1.6876611        1     1
13     2 0  -1.0289794  -1.6876611        1     1
14     2 0  -1.0289794  -1.6876611        1     1
15     2 1  -1.0289794  -1.6876611        1     1
```

The area indicator, the disease status of the individual, the deprivation index and the mean income of the area in which the individual lives, indicators for non-white ethnicity and whether the individual smoked, the pollution exposure and the area indicator are labelled `group`, `y`, `deprivation`, `mean.income`, `nonwhite` and `smoke` respectively. Binary indicators are 0 or 1 corresponding to no and yes respectively. The area indicator is only necessary when using models with random area effects.

3.2 Calling `eco`

Now we give examples of calling the *eco* function to fit models to the aggregate and individual datasets.

3.2.1 Aggregate data alone

Firstly, we fit the correct model to the simulated data, with two contextual covariates and two individual binary covariates, using the aggregate data alone.

```
> agg.eco <- eco(cbind(y, N) ~ deprivation + mean.income,
+     binary = ~nonwhite + smoke, data = aggdata)
```

- The first argument of `eco` is a formula, as used in most statistical modelling functions in R such as `lm` and `glm`. It specifies the *aggregate* component of the model, that is, the names of any covariates included in x_{ir} (equation 1).
- The argument `data` specifies a data frame which should contain all aggregate variables specified in the call to `eco`.
- The `binary` argument to `eco` is a formula whose right hand side should contain the names of any aggregate covariates considered as *individual-level* rather than contextual effects, here, non-white ethnicity and smoking. These should be binary at the individual level. `eco` will

fit the marginal model (2–3) by default for binary covariates. To fit the convolution normal-approximation model of Wakefield [7] specify `model=conditional` in the call to `eco`.

The `eco` function returns objects of class `ecoreg`. Printing an object of this class displays the estimated odds ratios $\exp(\alpha)$ associated with aggregate-level covariates, and odds ratios $\exp(\beta)$ associated with individual covariates (equation 1), along with their 95% confidence intervals, and $-2\times$ the maximised log-likelihood. In this example, the estimates are close to the true values used for simulating the data.

```
> agg.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, data = aggdata)

Aggregate-level odds ratios:
                    OR        l95        u95
(Intercept) 0.05398074 0.03906762 0.07458659
deprivation 0.95258788 0.86821545 1.04515955
mean.income 1.00683238 0.92165889 1.09987702

Individual-level odds ratios:
               OR      l95      u95
nonwhite 1.575121 1.097684 2.260219
smoke    2.005530 1.432162 2.808447

-2 x log-likelihood:  238.0643
```

3.2.2 Combining aggregate and individual data

Next we combine the aggregate data with the information from samples of individuals, as described in Section 2.3. Firstly, we form a reduced aggregate dataset, removing the individual outcomes and population totals which appear in the individual data. We assume that the sampled individuals did not contribute to the estimation of the aggregate covariates.

```
> aggdata.sub <- aggdata
> aggdata.sub$y <- aggdata$y - tapply(indivdata$y,
+     indivdata$group, sum)
> aggdata.sub$N <- aggdata.sub$N - 10
```

Again, we fit the correct model to the simulated data, with two contextual covariates and two individual binary covariates. The individual-level regression model is given in the `iformula` argument. The name of the individual-level dataset, in which the variables in the individual-level model should appear, is given in the `idata` argument.

```
> agg.indiv.eco <- eco(cbind(y, N) ~ deprivation +
+     mean.income, binary = ~nonwhite + smoke, iformula = y ~
+     deprivation + mean.income + nonwhite + smoke,
```

```
+     data = aggdata.sub, idata = sim1$idata)
> agg.indiv.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, iformula = y ~ deprivation + mean.income + nonwhite +
    smoke, data = aggdata.sub, idata = sim1$idata)

Aggregate-level odds ratios:
                    OR        l95        u95
(Intercept) 0.05179549 0.03832877 0.06999371
deprivation 0.95121091 0.86688939 1.04373431
mean.income 1.00921302 0.92420081 1.10204505

Individual-level odds ratios:
               OR      l95      u95
nonwhite 1.638318 1.182613 2.269622
smoke    2.050944 1.502937 2.798767

-2 x log-likelihood:  562.3413
```

In this example, combining with the individual sample data does not noticeably improve the precision of the estimates.

3.2.3 Importance of the between-area exposure contrasts

Suppose now that we simulate data with a much lower between-area exposure variance, such as a uniform(0, 0.2) distribution for proportion of non-white ethnicity and uniform(0.1, 0.3) for proportion of smokers. The aggregate data now contain less information about the individual-level effects. The amount of individual-level information in ecological data decreases as the between-area to the within-area variability of the exposure decreases. When there are low exposure contrasts between areas, inference may be improved by combining the ecological data with individual-level data [8].

When fitting the true individual model to the aggregate data alone, we obtain highly imprecise estimates for the individual-level effects.

```
> phi <- cbind(nonwhite = runif(ng, 0, 0.2), smoke = runif(ng,
+     0.1, 0.3))
> sim.df <- as.data.frame(cbind(ctx, phi))
> sim1 <- sim.eco(N, ctx = ~deprivation + mean.income,
+     binary = ~nonwhite + smoke, data = sim.df,
+     mu = mu, alpha.c = alpha.c, alpha = alpha,
+     isam = 10)
> aggdata <- as.data.frame(cbind(y = sim1$y, N = N,
+     sim.df))
> indivdata <- sim1$idata
> agg.eco <- eco(cbind(y, N) ~ deprivation + mean.income,
```

```
+     binary = ~nonwhite + smoke, data = aggdata)
> agg.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, data = aggdata)

Aggregate-level odds ratios:
                    OR        l95       u95
(Intercept) 0.08322794 0.05458192 0.1269081
deprivation 0.93587832 0.83970494 1.0430667
mean.income 1.13835917 1.02762594 1.2610246

Individual-level odds ratios:
               OR        l95      u95
nonwhite 0.6781861 0.04157216 11.06357
smoke    0.4947661 0.01855474 13.19305

-2 x log-likelihood:  209.9219
```

To be able to estimate the individual-level effects more accurately, we combine the aggregate data with the individual-level sample data.

```
> aggdata.sub <- aggdata
> aggdata.sub$y <- aggdata$y - tapply(indivdata$y,
+     indivdata$group, sum)
> aggdata.sub$N <- aggdata.sub$N - 10
> agg.indiv.eco <- eco(cbind(y, N) ~ deprivation +
+     mean.income, binary = ~nonwhite + smoke, iformula = y ~
+     deprivation + mean.income + nonwhite + smoke,
+     data = aggdata.sub, idata = indivdata)
> agg.indiv.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, iformula = y ~ deprivation + mean.income + nonwhite +
    smoke, data = aggdata.sub, idata = indivdata)

Aggregate-level odds ratios:
                    OR        l95        u95
(Intercept) 0.05860766 0.04638414 0.07405244
deprivation 0.91778108 0.82430940 1.02185188
mean.income 1.15270444 1.04056073 1.27693416

Individual-level odds ratios:
               OR       l95      u95
nonwhite 1.516919 0.7021149 3.277304
```

```
smoke    1.840635 1.0189851 3.324813

-2 x log-likelihood:  515.2852
```

This raises the question of whether the aggregate data do contribute any information in this case, and whether we can do just as well by analysing the individual data alone. But we find that precision is lower when only the individual data are included.

```
> indiv.eco <- eco(iformula = y ~ deprivation +
+     mean.income + nonwhite + smoke, idata = indivdata)
> indiv.eco

Call:
eco(iformula = y ~ deprivation + mean.income + nonwhite + smoke,
    idata = indivdata)

Aggregate-level odds ratios:
                    OR        l95       u95
(Intercept) 0.0730906 0.04822698 0.1107728

Individual-level odds ratios:
                  OR       l95      u95
deprivation 0.921286 0.6850914 1.238912
mean.income 1.015972 0.7695288 1.341338
nonwhite    1.753438 0.7300265 4.211553
smoke       2.260043 1.1906920 4.289769

-2 x log-likelihood:  294.8885
```

3.2.4 Normally-distributed covariates

We now add a continuous covariate to the simulated data, representing estimated exposure to air pollution. This will have a constant within-area standard deviation of 2, and within-area means varying around 1.24 with between-area standard deviation 0.1. A disease outcome is simulated with an odds ratio of 1.2 for one unit of pollution exposured, and the same odds ratios as before for the other covariates.

```
> phi <- cbind(nonwhite = runif(ng), smoke = runif(ng))
> sim.df <- as.data.frame(cbind(ctx, phi))
> sim.df$poll <- rnorm(ng, 1.24, 0.1)
> sim.df$poll.sd <- rep(0.2, ng)
> sim1 <- sim.eco(N, ctx = ~deprivation + mean.income,
+     binary = ~nonwhite + smoke, m = sim.df["poll"],
+     S = sim.df["poll.sd"], beta = log(2), data = sim.df,
+     mu = mu, alpha.c = alpha.c, alpha = alpha,
+     isam = 10)
> aggdata <- as.data.frame(cbind(y = sim1$y, N = N,
```

```
+     sim.df))
> aggdata[1:5, ]

   y   N deprivation  mean.income   nonwhite      smoke
1 14 100  0.81251444 -0.861417267 0.02777653 0.39783656
2 24 100 -1.02897943 -1.687661091 0.88876150 0.97207114
3 19 100 -0.04496483 -0.003495713 0.31863619 0.52441222
4 12 100  0.08924747  1.141115986 0.31363207 0.08960427
5 29 100 -1.78817352 -1.094355893 0.94361319 0.89781284
      poll poll.sd
1 1.140624     0.2
2 1.127442     0.2
3 1.302838     0.2
4 1.232604     0.2
5 1.211470     0.2
```

The area mean of the covariate is called `poll` in the aggregate dataset, and the area standard deviation is called `poll.sd`. We now model these data as before, including pollution as another individual-level covariate in the model.

```
> agg.eco <- eco(cbind(y, N) ~ deprivation + mean.income,
+     normal = ~poll, norm.var = poll.sd, binary = ~nonwhite +
+         smoke, data = aggdata)
> agg.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, normal = ~poll, data = aggdata, norm.var = poll.sd)

Aggregate-level odds ratios:
                    OR        l95       u95
(Intercept) 0.05453943 0.02132434 0.1394908
deprivation 0.97214238 0.90755177 1.0413299
mean.income 1.01163581 0.94869388 1.0787537

Individual-level odds ratios:
               OR      l95      u95
nonwhite 1.388107 1.064469 1.810144
smoke    1.841820 1.412568 2.401515
poll     2.094224 1.029639 4.259525

-2 x log-likelihood:  255.2786
```

The `normal` argument to `eco` is a formula whose right-hand side should contain variables denoting the group-level means of the normally-distributed covariates. These covariates will then be fitted as individual-level effects, using a model of the form of equation (8) by default, or (9) if the outcome is non-rare and `model = "poisson"` is specified. The `norm.var` is used to supply the corresponding group-level variances.

The true odds ratio of 2 for pollution is fairly well estimated. Now suppose the pollution data had a lower ratio of between-area to within-area standard deviation.

```
> sim.df$poll <- rnorm(ng, 1.24, 0.1)
> sim.df$poll.sd <- rep(0.2, ng)
> sim1 <- sim.eco(N, ctx = ~deprivation + mean.income,
+     binary = ~nonwhite + smoke, m = sim.df["poll"],
+     S = sim.df["poll.sd"], beta = log(2), data = sim.df,
+     mu = mu, alpha.c = alpha.c, alpha = alpha,
+     isam = 10)
> aggdata <- as.data.frame(cbind(y = sim1$y, N = N,
+     sim.df))
> agg.eco <- eco(cbind(y, N) ~ deprivation + mean.income,
+     normal = ~poll, norm.var = poll.sd, binary = ~nonwhite +
+         smoke, data = aggdata)
> agg.eco

Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, normal = ~poll, data = aggdata, norm.var = poll.sd)

Aggregate-level odds ratios:
                    OR        l95       u95
(Intercept) 0.0463745 0.01993314 0.1078904
deprivation 1.0245603 0.95681920 1.0970973
mean.income 1.0635478 0.99672811 1.1348469

Individual-level odds ratios:
               OR      l95      u95
nonwhite 1.161552 0.902295 1.495301
smoke    1.533980 1.179983 1.994178
poll     2.818187 1.462164 5.431795

-2 x log-likelihood:  267.6370
```

Now the confidence interval for the pollution odds ratio estimate is much wider, as there is less information in the aggregate data. However, the estimate is not biased. Again, to improve precision we can incorporate the individual-level data. Remember to include pollution in the individual level model `iformula`.

```
> indivdata <- sim1$idata
> indivdata[1:5, ]

  group y deprivation mean.income nonwhite smoke     poll
1     1 0   0.8125144  -0.8614173        0     0 1.027011
2     1 0   0.8125144  -0.8614173        0     1 1.330753
3     1 0   0.8125144  -0.8614173        0     0 1.210724
```

```
4     1 0   0.8125144  -0.8614173        0     1 1.207985
5     1 0   0.8125144  -0.8614173        0     1 0.920614
```

```
> aggdata.sub <- aggdata
> aggdata.sub$y <- aggdata$y - tapply(indivdata$y,
+     indivdata$group, sum)
> aggdata.sub$N <- aggdata.sub$N - 10
> agg.indiv.eco <- eco(cbind(y, N) ~ deprivation +
+     mean.income, normal = ~poll, norm.var = poll.sd,
+     binary = ~nonwhite + smoke, data = aggdata.sub,
+     iformula = y ~ deprivation + mean.income +
+         nonwhite + smoke + poll, idata = indivdata)
> agg.indiv.eco
```

```
Call:
eco(formula = cbind(y, N) ~ deprivation + mean.income, binary = ~nonwhite +
    smoke, normal = ~poll, iformula = y ~ deprivation + mean.income +
    nonwhite + smoke + poll, data = aggdata.sub, idata = indivdata,
    norm.var = poll.sd)

Aggregate-level odds ratios:
                    OR        l95       u95
(Intercept) 0.05070776 0.02510091 0.1024376
deprivation 1.03456045 0.96568396 1.1083495
mean.income 1.05323940 0.98682906 1.1241189

Individual-level odds ratios:
               OR       l95      u95
nonwhite 1.239174 0.9733386 1.577615
smoke    1.775614 1.3789404 2.286396
poll     2.286011 1.3377274 3.906512

-2 x log-likelihood:  727.0011
```

The precision of the estimate for pollution is improved.

3.2.5 Within-area distribution of binary covariates

To build the model (3) with more than one binary covariate, by default `eco` assumes that the covariates are independent within areas. Often this assumption is not appropriate, especially when considering, for example, socio-economically related factors.

To account for the joint within-area distribution of a set of binary or categorical covariates, use the `cross` argument to `eco`. This should be a matrix containing the same number of rows as the aggregate data, and number of columns equal to the distinct number of covariate categories into which an individual can belong. For full details of how to specify `cross`, refer to the help page for the `eco` function. The `cross` needs to be calculated by the user before calling `eco`. Individual data may be required to estimate `cross`, as typical census data do not give detailed cross-classification tables.

This is now illustrated with a hypothetical example. We introduce another covariate called `soclass` into the `aggdata` data representing the proportion of individuals in a lower social class. This is likely to be correlated with smoking at both aggregate and individual level. Suppose that we have an information from an individual-level survey, suggesting that an individual is twice as likely to smoke if in a lower social class. We wish to use this information to construct a matrix with 100 rows and four columns, representing estimates of the proportion of individuals who are in each of four categories:

1. neither smoke nor are in the lower social class (p_{00})
2. smoke but are not in the lower social class (p_{10})
3. do not smoke but are in the lower social class (p_{01})
4. smoke and are in the lower social class. (p_{11})

Let p_A be the proportion of smokers, and p_B be the proportion of individuals in the lower social class, in an area. We know that $p_{10}+p_{11} = p_A$, $p_{01}+p_{11} = p_B$ and, from the survey, $(p_{11}/p_B)/(p_{10}/(1-p_B)) = 2$. This leads, for example, to

$$p_{11} = 2p_Ap_B/(1+p_B)$$

This is enough information to construct all the estimated cross-classification probabilities, as illustrated by the following R code.

```
> aggdata$soclass <- plogis(qlogis(aggdata$smoke) +
+     runif(ng, -1, 1))
> pa <- aggdata$smoke
> pb <- aggdata$soclass
> p11 <- pa * pb * 2/(1 + pb)
> p10 <- pa - p11
> p01 <- pb - p11
> p00 <- 1 - (p01 + p10 + p11)
> cross <- cbind(p00, p10, p01, p11)
> cross[1:5, ]
```

```
             p00         p10        p01         p11
[1,] 0.514272326 0.241819040 0.08789111 0.156017524
[2,] 0.006071811 0.005809009 0.02185705 0.966262130
[3,] 0.337076166 0.185478119 0.13851162 0.338934100
[4,] 0.864384058 0.080193680 0.04601168 0.009410586
[5,] 0.063001388 0.057626100 0.03918577 0.840186744
```

This is the cross-classification matrix that we can supply to `eco` if we wanted to construct an ecological model including both smoking and social class, for example

```
> eco(cbind(y, N) ~ 1, binary = ~smoke + soclass,
+     cross = cross, data = aggdata)
```

To account for the joint within-area distribution of a set of continuous covariates, use the `norm.var` argument to `eco` to specify the joint covariance matrix of the covariates, assumed normally distributed, for each area. For further details of how to specify `norm.var` in this way, refer to the help page for the `eco` function.

ecoreg does not support specifying the within-area covariance between binary and continuous covariates. These are assumed to be independent in the model.

3.2.6 Stratification

Suppose that outcome data and covariate data are available by age and To account for differing disease risks in strata defined by age and sex, using fixed offsets determined from the whole population (as in equations 1 and 5), use the `strata`, `pstrata` and `istrata` arguments to `eco`.

- `pstrata` should be a vector with one element for each stratum, giving the assumed baseline outcome probabilities for the strata.
- `strata` should be a matrix with the same number of rows as the aggregate data. Rows represent areas, and columns represent the strata occupancy *proportions* for those areas (which are used as estimates of the observed occupancy *probabilities*). Alternatively, to account for within-area correlation between strata membership and binary covariate status, the cross-classification between strata and covariates can be specified in the `cross` argument. See the help page to `eco`.
- If individual data are modelled, `istrata` should be a variable containing the individual-level variable indicating the stratum an individual occupies. This should be a factor, whose levels correspond to the columns of the matrix `strata`.

3.2.7 Categorical covariates

As well as binary covariates, categorical covariates can also be fitted as individual-level predictors. The aggregate data for categorical covariates must be supplied separately from the main aggregate dataset, in the `categorical` argument to `eco`. See the help page for `eco`. In practice, there is not likely to be enough information in ecological data for successful ecological inference on categorical variables with large numbers of categories.

3.2.8 Random effects models

By default, `eco` assumes the baseline risk μ_i (equation 1) is constant μ between areas i. Optionally, `eco` can also fit μ_i as a normally-distributed random effect. If `random=TRUE` is specified in the call to `eco`, an area-level random intercept is included in the model. In this case the data should indicate which area each row of the data corresponds to. In the individual data, `igroups` should give the name of a variable containing the group identifiers of the individual-level data. In the aggregate data, by default, the groups are the row numbers of the dataset. Alternatively, `groups` specifies a group-level variable containing the group identifiers to be matched with the groups given in `igroups`.

`eco` uses adaptive Gauss-Hermite integration [16] to fit random effects models. The Gauss-Hermite integration can be controlled by the arguments `gh.points` and `iter.adapt` to `eco`.

`gh.points` gives the number of points to use for quadrature, while `iter.adapt` gives the number of iterations to use for the adaptive phase of the algorithm.

Random effects model fitting is relatively slow, and it may be useful to view the progress of the model fitting by specifying a `control` argument, such as `control=list(trace=1, REPORT=1)`. This is passed from `eco` to `optim`, the R function which performs optimisation of the likelihood. See `help(optim)` for further options to control optimisation.

4 Warnings and limitations

- It is easy to over-fit models, especially with several covariates. Often there is not enough information available in aggregate data.
- When fitting many covariates, it is essential to account for their within-area distribution.
- Continuous covariates must be normally-distributed or able to be transformed to normality.
- Only limited error-checking is performed at the moment. `eco` may fail with an incomprehensible error message if the model or data are specified wrongly or inconsistently.

5 *eco* reference guide

The R help page for `eco` gives details of all the allowed arguments and options to the `eco` function. To view this online in R, type:

```
> help(eco)
```

Similarly, all other functions in the package have help pages, which should always be consulted in case of doubt about how to call them. The web-browser based help interface may be convenient - type

```
> help.start()
```

and navigate to Packages . . . ecoreg, which brings up a list of all the functions in the package with links to their documentation, and a link to this manual in PDF format.

6 Similar software

- WinBUGS (`http://www.mrc-bsu.cam.ac.uk/bugs`) can be used to fit these models from a Bayesian perspective using Markov Chain Monte Carlo simulation. While computationally slower, this approach is amenable to extension to account for complexities such as random intercepts, random coefficients, spatial correlation and measurement error. WinBUGS files containing worked examples of a range of these models, using methods described by Jackson *et al.* [8] [9] are provided at `http://www.bias-project.org.uk/software`.
- The R package *eiPack* implements ecological inference for R $\times$ C contingency tables.

- The R package *MCMCpack*, available from CRAN, implements ecological inference for 2×2 tables using a Bayesian hierarchical model described by Wakefield [7].
- The R package *eco*, available from CRAN, implements ecological inference for 2×2 tables, using methods described by Imai and Lu [17].
- *EI* and *EzI* [18] by Kenneth Benoit and Gary King, implementing methods from King [19].

References

[1] S. Richardson, I. Stucker, and D. Hémon. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, 16(1):111–120, 1987.

[2] S. Greenland and H. Morgenstern. Ecological bias, confounding and effect modification. *International Journal of Epidemiology*, 18:269–284, 1989.

[3] S. Greenland and J. Robins. Ecological studies — biases, misconceptions and counterexamples. *American Journal Of Epidemiology*, 139:747–760, 1994.

[4] S. Richardson and C. Monfort. Ecological correlation studies. In *Spatial Epidemiology*, chapter 11, pages 205–220. Oxford University Press, Oxford, 2000.

[5] J. Wakefield and R. Salway. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A*, 164(1):119–137, 2001.

[6] R. L. Prentice and L. Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82:113–125, 1995.

[7] J. Wakefield. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A*, 167(3):385–445, 2004.

[8] C. H. Jackson, N. G. Best, and S. Richardson. Improving ecological inference using individual-level data. *Statistics in Medicine*, 25(12):2136–2159, 2006.

[9] C. H. Jackson, N. G. Best, and S. Richardson. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society, Series A*, 171(1):159–178, 2008.

[10] S. Richardson and N. Best. Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, 14:129–147, 2003.

[11] V. Lasserre, C. Guihenneuc-Jouyaux, and S. Richardson. Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine*, 19:45–59, 2000.

[12] D. Clayton and M. Hills. *Statistical Models in Epidemiology*. Oxford University Press, 1993.

[13] P.R. Rosenbaum and D.B. Rubin. Difficulties with regression analyses of age-adjusted rates. *Biometrics*, 40:437–443, 1984.

[14] N. Best, S. Cockings, J. Bennett, J. Wakefield, and P. Elliott. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society, Series A*, 164(1):155–174, 2001.

[15] R. Salway and J. Wakefield. Sources of bias in ecological studies of non-rare events. *Environmental and Ecological Statistics*, 12(3):321–347, 2005.

[16] Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81:624–629, 1994.

[17] K. Imai and Y. Lu. An incomplete data approach to the ecological inference problem. *(working paper, Princeton University)*, 2006. (URL: http://imai.princeton.edu/research/coarse.html).

[18] G. King. Ei: A program for ecological inference. *Journal of Statistical Software*, 11(7), 2004.

[19] G. King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, 1997.