Statistics in Environmental Research (BUC Workshop Series I)

Computer Labs

Gavin Shaddick [1] James V. Zidek [2]

[1]University of Bath, UK

[2]University of British Columbia, Canada

12th November 2015

Question 1

Consider the Scottish lip cancer data (scotdat.txt). Fit the Poisson log-linear regression model, that is $Y_i \sim \text{Poisson}(E_i \exp[\alpha_0 + \alpha_1 x_i])$ where Y_i, E_i and x_i are the number of disease counts, the expected numbers, and the proportion in agriculture, fishing and farming (AFF) in area i, $i = 1, ..., 56$.

Now fit an alternative model that includes latitude and longitude in the log-linear model, in order to investigate "confounding by location". Discuss your findings.

Question 2/a

Gordian, Haneuse and Wakefield (2006, Journal of Exposure Science and Environmental Epidemiology, 16, 49–5) report a study which investigated whether proximity to traffic at residential ocation is associated with being diagnosed with asthma as a young child.

Here we will look at the association between traffic count and another response variable, the number of respiratory illnesses per year, in particular let $Y = 0$ if a child had 4 or less illnesses, and $Y = 1$ if greater than 4.

Question 2/b

On the website you will find the data on the response and exposure variables, along with a number of confounders (jan31.csv) (you may wish to control for some or all of gender, age, a smoker in the house, income, school attended, parental smoker), a script to read these data into R, which also contains details of the variables (AlaskaScript.R). The exposure variable has been scaled to lie between 0 and 10.

Analyze these data using logistic regression models, and report your findings, including any drawbacks of the study/analysis.

Question 3 / a

In this question we will carry out disease mapping for Ohio lung cancer mortality data (ohio.dat) from 1988 that is on the website along with various R functions (OhioMap.R) for producing maps for counties within Ohio.

(a) Provide a map of the SMRs, with expected numbers adjusted for gender, race and age.

(b) Provide a map of the estimated standard errors of the SMRs,

$$\widehat{\text{s.e.}}(\text{SMR}_i) = \frac{\widehat{\theta}_i^{1/2}}{E_i^{1/2}}.$$

(c) Provide histograms of $\theta_i|\mathbf{y}, \widehat{\mu}, \widehat{\alpha}$, for $i = 1, 2, 3, 4$.

Question 3/b

(d) Provide a map of the posterior mean smoothed EB estimates $\widehat{\text{RR}}_i$, and compare with the SMR map.

(e) Provide a map of the posterior standard deviations of the smoothed estimates $\widehat{\text{RR}}_i$, which are given by

$$\text{sd}(\widehat{\text{RR}}_i) = \frac{\mu(\widehat{\alpha} + y_i)^{1/2}}{\widehat{\alpha} + E_i\widehat{\mu}},$$

and compare with the standard error of the SMR map.

(f) Calculate the posterior probabilities that RR_i exceeds the threshold 1.2, and map these quantities.

(g) Examine the proportionality assumption $p_{ij} = \theta_i \times q_j$ using a suitable method. For example you might look at SMR maps by gender, race, age.

Question 4/a

In this question we will investigate the properties of the empirical Bayes estimates. Specifically consider the mean squared error of the collection of estimates $\widehat{\theta}_i$, compared to the true values of θ_i, $i = 1, \ldots, n$:

$$\text{MSE}(\widehat{\theta}_i) = \frac{1}{n} \sum_{i=1}^{n} (\widehat{\theta}_i - \theta_i)^2.$$

We will investigate via simulation the size of the mean squared error for different choices of the distribution from which we simulate the random effects.

Question 4/b

For $n = 50$, and $E_i \sim \text{Unif}(0.1, 50)$ simulate $Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i)$ with $\theta_i \sim p_\theta()$ and calculate the SMRs Y_i/E_i and EB estimates $\widehat{\mu}\left(\frac{\widehat{\alpha}+Y_i}{\widehat{\alpha}+\widehat{\mu}E_i}\right)$ for the choices of p_θ:

(a) Ga(5,5)

(b) Lognormal(0,1)

(c) Unif(0.1,10)

and comment on what you find.

Question 5 / a

In this question we will simulate health data to examine the construction of relative risk estimates for point data.

Consider the locations of the larynx cancer controls in South Lancashire, and suppose that these represent the population at risk. There is an incinerator in the study region, whose location is contained in the `southlancs` dataset. .

Question 5/b

Suppose that population at risk individual i can become a case with probability $\exp(\beta_0 + \beta_1 d_i)$, where d_i is the distance between the incinerator and $\boldsymbol{s}_i$, the residential location of individual i; if the incinerator is harmful then β_1 will be negative, so that the risk increases as d_i decreases.

Simulate cases from the population at risk, experimenting with different values of β_0 and β_1; $\exp(\beta_0)$ is the risk at the incinerator, and $\exp(\beta_1)$ is the relative risk (< 1) corresponding to a decrease in distance of one unit.

Question 6

In the `geoR` library there are data `ca20` which you should explore/analyze using geostatistical techniques. For example, you may:

(a) Look at empirical semi-variograms (clouds and binned).

(b) Examine Monte Carlo intervals of no spatial dependence.

(c) Fit variogram models to the data.

(d) Carry out kriging and examine the resultant surfaces.

Question 7

Simulate and plot a discrete time $t = 0, 1, \ldots, 50$ AR(1) process for several different lag one autocorrelation function values, say $\alpha_1 = 0.1, 0.5, 0.9$.

What misleading conclusions might a casual observer make about this process?

Question 8

The next problem concerns the ozone dataset for NY State (LabOzone.txt), specifically the hourly ozone time series for Site 9. The readme file for the lab describes the data for the complete set of 9 sites.

Develop a time series model for the temporal process associated with that site along with a 24 hour ahead forecaster.

Question 9 / a

In this question we will fit several theoretical variogram to a variable of your choice in the meuse data set from gstat package. We will find the best fitted model based on the SSE criteria and by using cross validation.

Use the fit.variogram() function from gstat package. Set the option print.SSE of this function to TRUE. Read the help page for this function carefully. Concentrate on one of the metal variables in the meuse data set and fit at least four different families of variogram models to the empirical variogram computed by the variog() function.

Question 9/b

You may do the analysis on the original or make a transformation if you like.

```
library(gstat) data(meuse)
vgm1 <- variogram(log(zinc)~1, ~x+y, meuse,
print.SSE=TRUE) plot(vgm1)
meuse.vfit <- fit.variogram(vgm1, vgm(1,"Sph",300,1))
plot(vgm1,model=meuse.fit)
```

Based on the SSE criteria choose the best fitted model.

Question 9/c

Now we will use cross validation to choose between a set of models. We will use the krige.cv() function from the gstat package. Read the help page carefully. When doing cross validation choose to use the method of one-leave-out by specifying nfold=1. For example you can do like this,

```
data(meuse)
m <- vgm(.59, "Sph", 874, .04)
x <- krige.cv(log(zinc)~1, ~x+y,
model = m, data = meuse, nmax = 40, nfold=1)
```

Question 9/d

Use the following functions to calculate the mean error (ME), the mean squared error (MSE), and the mean squared deviation ratio (MSDR) diagnostics.

```
ME <- function(xv.obj){ tmp <- xv.obj$error
return(sum(tmp)/length(tmp))
}
MSE <- function(xv.obj){ tmp <- xv.obj$error
return(sum(tmp^2)/length(tmp))
}
MSDR <- function(xv.obj){ e2 <- xv.obj$error^2
s2 <- xv.obj$krige.var
msdr <- sum(e2/s2)/length(e2) return(msdr)
}
```

To get the diagnostics do the following on the cross-validation object x computed above ME(x), MSE(x), MSDR(x)

PACKAGES TO INSTALL

- `geoR`
  (Note if you are using a Mac you will need to install an up to date version of xquartz, `http://www.xquartz.org`)
- `MASS`
- `maps`

Source files: `PolyMap.R`
`OhioMap.R`

```
> install.packages("name_of_package")
```