Practical 5: Analysis of point pattern data

March 18, 2019

In this practical we are going to consider a population based case-control study on incident pleural malignant mesotheliomas in the area of Casale Monferrato, diagnosed between 1 January 1987 and 30 June 1993. Casale is a medium-size town in the North-West of Italy where a large asbestos cement (AC) plant was active from 1907 to 1985.

Results from this study were published originally by Magnani et al. 2001. Successively, part of the data were re-analysed within a Bayesian framework by Dreassi et al. 2008. The data that you are going to use are related to this re-analysis.

The study investigates occupational and domestic asbestos exposure in relationship to the largest Italian asbestos cement (AC) factory. In total, 103 cases and 271 controls, are included in the present re-analysis.

The data are stored in a `.RData` file called `data_Casale.RData`. The data.frame `data` includes the following variables:

`case` = cases (1), controls (0)

`sex` = male (1), female (0)

`abs` = occupational exposure to AC industry, with (0) non exposed and (1) exposed

`age` = age in years

`dom` = domestic exposure to asbestos material, with (0) non exposed and (1) exposed

`rel` = exposure of any relative, with (0) non exposed and (1) exposed

`x` = x coordinate and `y` = y coordinate

`dist` = distance from the AC factory, in meters

DETAILS OF THE MODEL

In the model, sex (female is the reference level), age (continuous), domestic exposure to asbestos material (dom, binary) and occupational (AC) exposure of any relatives (rel, binary) are introduced as multiplicative terms, while occupational (AC) exposure (asb, binary) is introduced as an additive term in the excess risk function.

The effect of distance is included in the additive term, too. The function f(.) describes the decay of risk with increasing distance.

The rationale of this modelling choice is that different exposures to the same agent sum up (i.e. combine additively, see Breslow, 1986). The authors considered here the two main risk factors (occupation and environmental exposure) in the additive term of the predictor. The other covariates are left in the multiplicative term, even if it can be argued that for mesotheliomas the only relevant exposure is asbestos. For environmental exposure measured by residential distance the motivation to choose an additive specification is also to ensure that the risk be unchanged at infinite distance from the source.

The distance function is obviously not increasing and tends asymptotically to zero when distance goes to infinity. The authors specified an exponential decay with threshold (see Diggle and Rowlingson, 1994) functional forms for f(.):

```
f(d_i)=alpha exp(-beta  d_i^2)
```

where alpha is the excess relative risk at source and beta the parameter of the exponential decrease function.

Before starting the practical

Organise your work:

- Create a separate subdirectory in your home directory to save your files created during this practical (e.g. "C:/SpatialAnalysis2019/Practicals/Practical5").
- Copy all the files from the blackboard to your subdirectory (created just above).
- You will use the following R packages: spatstat, dplyr, R2OpenBUGS, mcmcplots.
- If not installed yet, then install the required packages using the function `install.packages()`
- Then, load needed libraries:

```r
library(spatstat)
library(dplyr)
library(R2OpenBUGS)
library(mcmcplots)
```

Data

- Load the data in `R`

```r
load(file="data_Casale.RData")
```

- Print the first rows of the data.frame object `data`

```r
head(data)
```

```
##   case age sex asb dom rel      x       y      dist
## 1    0  53   1   0   0   0 457855 4998437  2254.926
## 2    0  52   1   0   0   0 439688 5003621 16780.630
## 3    0  58   0   0   0   0 445980 4987700 14320.130
## 4    0  83   0   0   0   1 458310 4998190  2708.400
## 5    1  52   1   0   0   0 461163 4989960 10026.210
## 6    0  83   0   0   0   0 441983 5000830 13853.240
```

- Produce the summary statistics of these data

```r
summary(data)
```

```
##       case             age             sex              asb        
##  Min.   :0.0000   Min.   :32.00   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:57.00   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :66.00   Median :1.0000   Median :0.0000  
##  Mean   :0.2754   Mean   :65.38   Mean   :0.6043   Mean   :0.1123  
##  3rd Qu.:1.0000   3rd Qu.:75.00   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :94.00   Max.   :1.0000   Max.   :1.0000  
##       dom              rel              x                y          
##  Min.   :0.0000   Min.   :0.000   Min.   :395832   Min.   :4918538  
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:447499   1st Qu.:4992439  
##  Median :0.0000   Median :0.000   Median :455580   Median :4997786  
##  Mean   :0.3984   Mean   :0.123   Mean   :452901   Mean   :4995835  
##  3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:457323   3rd Qu.:4999391  
##  Max.   :1.0000   Max.   :1.000   Max.   :514617   Max.   :5069181  
##       dist         
##  Min.   :   63.53  
##  1st Qu.: 2664.55  
##  Median : 7601.60  
##  Mean   :10802.07  
##  3rd Qu.:12989.02  
##  Max.   :87073.60
```

- Compute the proportion of exposed and unexposed in the study

```r
with(data, table(asb)) %>% prop.table()
```

```
## asb
##         0         1 
## 0.8877005 0.1122995
```

- Compute the number of cases and controls with occupational exposure in the AC factory

```r
table(data$case, data$asb)
```

```
##    
##       0   1
##   0 257  14
##   1  75  28
```

- Compute the number of cases and controls by gender

```r
table(data$case, data$sex)
```

```
##
##       0   1
##   0 105 166
##   1  43  60
```

- Compute the number of cases and controls with domestic exposure to asbestos material

```r
table(data$case, data$dom)
```

```
##
##       0   1
##   0 173  98
##   1  52  51
```

- Compute the number of cases and controls with relatives working in the AC factory

```r
table(data$case, data$rel)
```

```
##
##       0   1
##   0 249  22
##   1  79  24
```

- Compute the average age, and its standard deviation, in cases and controls

```r
tapply(data$age, data$case, mean)
```

```
##       0       1
## 65.4428 65.2233
```

```r
tapply(data$age, data$case, sd)
```

```
##        0        1
## 11.71085 11.89903
```

- Compute the average age, and its standard deviation, in males and females

```r
tapply(data$age, data$sex, mean)
```

```
##        0        1
## 67.78378 63.80973
```

```r
tapply(data$age, data$sex, sd)
```

```
##        0        1
## 10.81982 12.08356
```

- Compute the distances in km and call the new varaible as `dista`

```r
dista <- data$dist/1000 # distances in km
```

- Compute the range (min and max) of the distances

```r
min(dista)
```

```
## [1] 0.06352952
```

```r
max(dista)
```

```
## [1] 87.0736
```

- plot the data. We use `spatstat`. A point pattern is represented in `spatstat` by an object of class `ppp`, that is a planar point pattern. See `help(ppp.object)` for further details.

A `ppp` object includes:

`x` Vector of x coordinates of data points

`y` Vector of y coordinates of data points

`window` of observation, an object of class `owin`

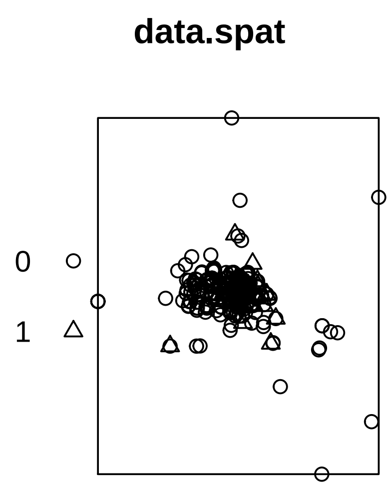
...

`marks`, (optional) vector of mark values

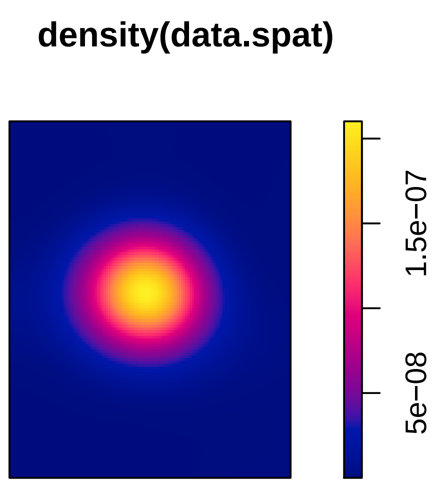
```r
# vector of x and y coordinates
x <- as.vector(data[,7])
y <- as.vector(data[,8])

# owin: is a window which defines the spatial extent of the ppp object
# here we create two vectors xrange and yrange for setting the window
xrange <- range(x, na.rm=T)
yrange <- range(y, na.rm=T)

W <- owin(xrange,yrange)

# marks (need to be a factor)
m <- factor(data$case, levels=0:1)

# create a ppp object, specifying the marks
data.spat = ppp(x, y, window=W, marks=m)
```

```
## Warning: data contain duplicated points
```

```r
# plot
plot(data.spat, which.marks = m)
```




```r
# kernel smoother of point density
plot(density(data.spat))
```



Model with BUGS

- Open and edit the model *Model.txt*. In particular, include statements in your model code to calculate the following quantities (note: revise practical 4 for this):

(a) the odds ratio associated with `sex`

(b) the odds ratio associated with `age`

(c) the odds ratio associated with the domestic exposure `dom`

(d) the odds ratio associated with the exposure of relatives `rel`

- Now run the model in `OpenBUGS` via `R`. To do so, follow the steps:

1. Prepare the list of data and call it `DataCas`

```
DataCas <- list("case"=data$case, "asb"=data$asb,"sex"=data$sex,
                "age"=data$age, "rel"=data$rel,"dom"=data$dom,
                "dista"=dista)
```

2. Prepare the list of initial values for the model

```
inits <- list(
  list(alpha=2,beta=0, alpha0=0, alphaage=0, alphasex=0, alpharel=0, alphadom=0,
alphaasb=0), # chain 1
list(alpha=1,beta=0.1, alpha0=0.1, alphaage=0.1, alphasex=0.1, alpharel=0.1, alphadom=0.1,
alphaasb=0.1) # chain 2
)
```

3. Prepare the set of parameters to be monitored

```
parameters <- c("aage","aasb","asex","arel",
                "adom","alpha","beta","alpha0","cost")
```

4. Prepare the MCMC setting. Here, we say we want to run two chains for 10,000 iterations, discarding the first 5,000 values, and thinning the remaining values by 10 (note, the Authors of this study used a very large number of iterations…)

```
ni <- 10000  # nb iterations
nt <- 10    # thinning interval
nb <- 5000  # nb iterations as burn-in
nc <- 2     # nb chains
```

5. Run the model using the function `bugs()`

```r
model.sim <- bugs(data = DataCas, parameters = parameters,
                  inits = inits,
                  model.file = "Model.txt",
                  n.chains = nc, n.iter = ni, n.burnin = nb,
                  n.thin = nt, debug = FALSE,
                  working.directory = getwd(),
                  codaPkg = FALSE, summary.only = FALSE,
                  bugs.seed = 5)
```

- Visualise the main diagnostic plots using `mcmcplot`.

```r
mcmcplot(model.sim)
```

- Compute the Gelman Rubin diagnostic

```r
gelman.diag(model.sim) # Gelman Rubin diagnostic
```

- What do you conclude about the convergenge of the model?
- Obtain the posterior summary statistics of the parameters monitored

```r
print(model.sim, digits = 2)
```

```
## Inference for Bugs model at "Model.txt",
## Current: 2 chains, each with 10000 iterations (first 5000 discarded), n.thin = 10
## Cumulative: n.sims = 10000 iterations saved
##            mean    sd   2.5%    25%    50%    75%  97.5% Rhat n.eff
## aage       1.00  0.01   0.98   0.99   1.00   1.01   1.02    1  5600
## aasb      26.29 11.58   9.38  17.69  24.47  32.89  53.88    1 10000
## asex       0.89  0.24   0.50   0.71   0.86   1.03   1.46    1 10000
## arel       1.54  0.59   0.70   1.12   1.45   1.86   2.97    1 10000
## adom       1.43  0.39   0.80   1.14   1.38   1.65   2.34    1  3100
## alpha      9.00  2.59   4.61   7.15   8.75  10.63  14.66    1   490
## beta       0.01  0.01   0.01   0.01   0.01   0.02   0.03    1  2100
## alpha0    -3.02  0.85  -4.71  -3.59  -3.01  -2.44  -1.37    1  2700
## cost       0.07  0.07   0.01   0.03   0.05   0.09   0.25    1  2700
## deviance 381.47  3.85 376.00 378.60 380.80 383.60 390.70    1  2100
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = Dbar-Dhat)
## pD = 6.84 and DIC = 388.30
## DIC is an estimate of expected predictive error (lower deviance is better).
```

- What is the risk of mesothelioma in relationship to occupational exposure in the AC factory? Report the estimated OR and the associate 95% credible interval

ANSWER: OR = 26.29 (95%CI 9.38 to 53.88)

- What is the risk of mesothelioma in relationship to relatives' occupational exposure in the AC factory? Report the estimated OR and the associate 95% credible interval

ANSWER: OR = 1.54 (5%CI 0.70 to 2.97)

- What is the risk of mesothelioma in relationship to domestic exposure from asbestos-containing items? Report the estimated OR and the associate 95% credible intervals

ANSWER: OR = 1.43 (5%CI 0.80 to 2.34)

- What is the posterior mean (and 95%CI) of the parameter evaluating the excess risk at the source?

ANSWER: `alpha` posterior mean = 9.00 (95%CI 4.61 to 14.66)