Raised Incidence Model for the Chorley Ribble Data

This is the summary of how to implement the Raised Incidence Model in R using the software package `spatstat` and the cancer data in the Chroley-Ribble Area following the book by Baddeley et al. (2016).

This Point Process Model approach was first developed by Diggle (1990) and then reformulated to a binary regression model by Diggle & Rowlingson (1994).

1 Data Description: Chorley- Ribble Data

(Page 9-10)

Spatial locations of cases of larynx and cancer of the lung and a disused industrial incinerator.

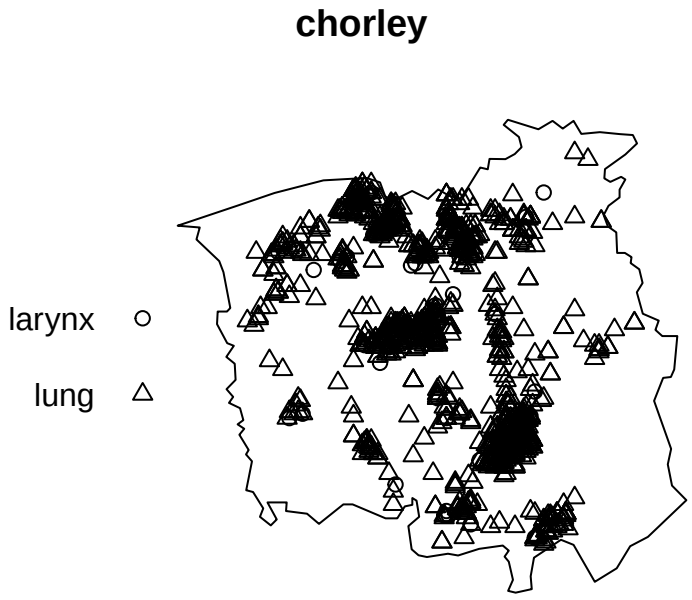
- Residential locations of new cases of cancer of the larynx (58 cases) and cancer of the lung (978 cases) in the Chorley and South Ribble Health Authority of Lancashire, England, between 1974 and 1983
- Spatial location of a disused industrial incinerator

1.1 Basic Summaries of the Chorley-Ribble Data

(Page 108-109)

```
library(spatstat)
```

```
plot(chorley)
```



```
summary(chorley)
```

```
## Marked planar point pattern:  1036 points
## Average intensity 3.287268 points per square km
##
## *Pattern contains duplicated points*
##
## Coordinates are given to 1 decimal place
## i.e. rounded to the nearest multiple of 0.1 km
##
## Multitype:
##        frequency proportion intensity
## larynx        58 0.05598456 0.1840363
## lung         978 0.94401540 3.1032320
##
## Window: polygonal boundary
## single connected closed polygon with 131 vertices
## enclosing rectangle: [343.45, 366.45] x [410.41, 431.79] km
##                      (23 x 21.38 km)
## Window area = 315.155 square km
## Unit of length: 1 km
## Fraction of frame area: 0.641
```

First, we split the whole dataset, to have the cases of larynx cancer separate from the controls of lung cancer. *(Page 320)*

```
S <- split(chorley)
larynx <- S$larynx # cases
lung <- S$lung # controls
```

2 Initital Model by Diggle (1990)

2.1 Null Model

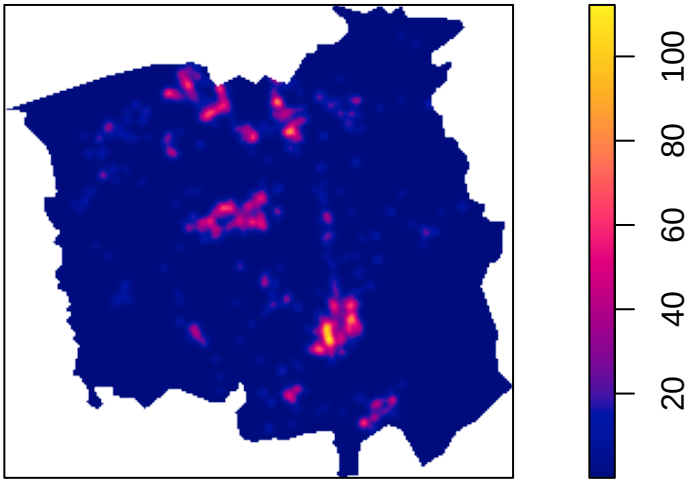
Here, we fit an inhomogeneous Poisson model with intensity that is proportional to the covariate $\lambda(u) = kZ(u)$ where Z is the covariate and κ the parameter to be estimated.

An important example of a baseline covariate $Z(u)$ is the spatially varying density of human population. The spatial point pattern of cases of a rare disease could reasonably be expected to follow a Poisson point process with intensity above, where k is the (constant) disease risk per head of population. The smoothed intensity of lung cancer cases can serve as a surrogate for the spatially varying density of the susceptible population. In this specific example, we follow Diggle in treating the cases of laryngeal cancer as the response, and taking the lung cancers as a covariate.

We apply kernel smoothing (Diggle, 1985) to the lung cancer locations to obtain an unnormalised estimate of the spatially varying population density of susceptibles.

Page 320 (Chapter 9.3.7 Models with offsets)

```
smo <- density(lung, sigma = 0.15, eps = 0.1, positive = TRUE)
plot(smo)
```



- `sigma`: adopted the smoothing bandwidth $\sigma = 0.15km$ chosen by Diggle (1990)
  - we could also employ some technique for automatic bandwidth selection, but according to the book, would give a wide range of result between 0.07 and 2 km.
- `eps`: The pixels are eps=0.1km wide
- `positive`: Logical value indicating whether to force all density values to be positive numbers. Here, it is set to `True` to ensure that negative or zero pixel values (due to numerical error) are replaced by a small positive number.

```
smo <- eval.im(pmax(smo, 1e-10))
```

- `eval.im`: performs pixel-by-pixel calculations on an image, or on several images, and returns a new image. It is used to avoid numerical instability (*p. 320*).

```
contour(smo) # contour plot
```

2.1.1 Fitting the Model, i.e. finding it's parameters

The resulting pixel image smo serves as the baseline for our model. The null model, of 'constant relative risk', postulates that the larynx cases are a Poisson process with intensity

$$\lambda_0(u) = \kappa Z(u)$$

at location u. The parameter κ adjusts for the relative abundance of the two types of data points; it is the baseline relative risk of larynx and lung cancer, multiplied by the ratio of sampling fractions used when these data were sampled from the cancer registry.

When fitting the model, we want to find the parameter κ:

```
# Finding rho
ppm(larynx ~ offset(log(smo)))
```

```
## Nonstationary Poisson process
##
## Log intensity:  ~offset(log(smo))
##
## Fitted trend coefficient:  (Intercept) = -2.938568
##
##               Estimate      S.E.   CI95.lo   CI95.hi Ztest      Zval
## (Intercept) -2.938568 0.1313064 -3.195924 -2.681212   *** -22.37947
```

Why using offset

We use the offset function to model a count variable as a rate. When you fit a model with an offset, the exponentiated regression coefficient of a predictor variable tells us how much the expected rate changes multiplicatively for a one unit increase in the predictor variable. (Source: To Offset or Not: Using Offsets in Count Models (cornell.edu))

When fitting a baseline model it is absolutely crucial to express the baseline term in the form offset(log(baseline)).

Using Quadscheme

```
Q <- quadscheme(larynx, eps = 0.1)
```

The command `quadscheme` generates a quadrature scheme from the point pattern. It is used to approximate an MLE for an inhomogeneous Poisson point process since the likelihood function involves an integral over the spatial window, wherefore the likelihood cannot be computed exactly, but must be approximated numerically. `ppm` by default uses Berman-Turner quadrature approximation.

`quadscheme(X, eps=0.1)` dummy point spacing 0.1 units.

```
chorley0fit <- ppm(Q ~ offset(log(smo)))
chorley0fit
```

```
## Nonstationary Poisson process
##
## Log intensity:  ~offset(log(smo))
##
## Fitted trend coefficient:  (Intercept) = -2.823815
##
##               Estimate      S.E.  CI95.lo   CI95.hi Ztest      Zval
## (Intercept) -2.823815 0.1313064 -3.08117 -2.566459   *** -21.50553
```

The fitted coefficient (Intercept)=-2.824 is the estimate of $\log\kappa$ in $\log\lambda_0(u) = \log\kappa + \log Z(u) = \theta + \log Z(u)$, so converting back to the form $\lambda_0(u) = \kappa Z(u)$, the fitted model is

$$\lambda_0(u) = e^{-2.939} Z(u) = 0.05292 Z(u)$$

In this example, note that the fitted parameter κ is not the estimated risk of laryngeal cancer per head of population, because the lung data are a subsample from the cancer registry, and lung cancer cases are a subset of the susceptible population. The best way to estimate the risk of laryngeal cancer assuming constant risk is to divide the total number of laryngeal cancer cases by an estimate of the total susceptible population, since the constant risk model does not involve spatial information. The model fitted above is useful mainly for comparison against alternative models where the risk of laryngeal cancer is spatially varying.

2.2 Alternative Model

$$\lambda_1(x) = \kappa\lambda_0(x)(1 + \alpha e^{-\beta x' x})$$

We start by defining the squared distance to the incinerator:

```
d2incin <- function(x, y, xincin = 354.5, yincin = 413.6) {
  (x - xincin)^2 + (y - yincin)^2
}
```

The arguments xincin,yincin are the coordinates of the incinerator

```
raisin <- function(x, y, alpha, beta) {
  1 + alpha * exp(-beta * d2incin(x, y))
}
```

```
chorleyDfit <- ippm(Q ~ offset(log(smo) + log(raisin)),
  start = list(alpha = 5, beta = 1)
)
```

Here, `ippm` is used in contrast to ppm, because it performs the iterative maximization for the profile likelihood. We need to use it because we have a general non-loglinear model.

```
chorleyDfit
```

```
## Nonstationary Poisson process
##
## Log intensity:  ~offset(log(smo) + log(raisin))
##
## Fitted trend coefficient:  (Intercept) = -2.895816
##
## Irregular parameters (covfunargs) fitted by 'ippm':
## alpha = 22.24097
## beta = 0.8892389
##               Estimate      S.E.   CI95.lo  CI95.hi Ztest      Zval
## (Intercept) -2.895816 0.1313064 -3.153172 -2.63846   *** -22.05388
```

2.3 Spatial Inference: comparing the two models

For the Chorley-Ribble data the key question is whether distance to the incinerator has an effect on the intensity of laryngeal cancer, after allowing for the fact that the intensity of laryngeal cancer cases will depend on the spatially varying density of the population at risk. A likelihood ratio test can be used to compare two models, both involving a term related to the population density, but only one model including a term related to distance from the incinerator.

The null hypothesis is the smaller model chorley0fit which does not include the covariate effect; the alternative hypothesis is the larger model chorleyDfit including the covariate effect.

```
anova(chorley0fit, chorleyDfit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: ~offset(log(smo))    Poisson
## Model 2: ~offset(log(smo) + log(raisin))      Poisson
##   Npar Df Deviance Pr(>Chi)
## 1    1
## 2    3  2   9.3105 0.009511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result indicates strong evidence against H0 in favor of the raised incidence alternative.

3 Conditional Raised Incidence Model by Diggle and Rowlingson (1994)

We can also find the code for the binary regression model by Diggle & Rowlingson (1994) for one putative source. Here, the controls are also specific point data and not an estimated intensity function as above.

```r
incin <- as.ppp(chorley.extra$incin, W = Window(chorley))
dincin <- distfun(incin)
Q <- quadscheme.logi(larynx, lung)
fit <- ppm(Q ~ dincin, method = "logi")
```

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

Here an alternative fitting methods is used. Conditional logistic regression for Poisson models can be extended to Gibbs models as logistic composite likelihood and is available in `ppm` by setting method="logi"

```r
fit
```

```
## Warning: vcov is not implemented for dummy type 'given' - using 'poisson'
## formula

## Nonstationary Poisson process
##
## Log intensity:  ~dincin
##
## Fitted trend coefficients:
## (Intercept)      dincin
## -1.58154748 -0.01216818
##
##                Estimate       S.E.     CI95.lo     CI95.hi Ztest       Zval
## (Intercept) -1.58154748 0.34136376 -2.25060816 -0.91248680   *** -4.6330268
## dincin      -0.01216818 0.03465235 -0.08008553  0.05574917       -0.3511503
```

4 References

Baddeley, A., Rubak, E., & Turner, R. (2016). Spatial Point Patterns Methodology and Applications with R. Chapman & Hall/CRC.

Diggle, P. J. (1985). A Kernel Method for Smoothing Point Process Data. *Applied Statistics*, *34*(2), 138. https://doi.org/10.2307/2347366

Diggle, P. J. (1990). A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *153*(3), 349. https://doi.org/10.2307/2982977

Diggle, P. J., & Rowlingson, B. S. (1994). A Conditional Approach to Point Process Modelling of Elevated Risk. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 433. https://doi.org/10.2307/2983529