Introduction to disease mapping - Part 1

Bayesian modelling for spatial and spatio-temporal data

MSc in Epidemiology

Week 6

General framework - I

- **Data** for a region of interest/reference area, geographical level and a specific period, e.g. England, ward level, 2009-2012
  - O_i: Observed number of cases in area i
    - Lung cancer deaths in males aged 45+
    - Congenital anomalies
  - n_i: Population at risk in area i
    - Male population aged 45+
    - Live births and stillbirths
- **Parameter of interest:** Relative risk, λ, in each area i compared with the chosen reference area

General framework - II

- Standard statistical model if rare disease and/or small areas

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

where:

- E_i is the expected number of cases in area i (fixed and know function of n_i)
- λ_i estimated by Standardised Mortality/Incidence Ratio
  - The standardized ratio is the maximum likelihood estimator for the unknown area-specific relative risk of events of interests in area i

$$\hat{\lambda}_i = \text{SMR}_i \text{ or } \text{SIR}_i = \frac{O_i}{E_i} \quad \text{and} \quad \text{Var}(\hat{\lambda}_i) = \frac{\lambda_i}{E_i} \rightarrow \hat{\text{Var}}(\hat{\lambda}) = \frac{O_i}{E_i^2}$$

So that areas with small E_i have high associated variance

Recall: $X \sim \text{Poisson}(\mu) \Leftrightarrow \text{E}(X) = \text{Var}(X) = \mu$

Expected numbers of cases - definition

- Expected number of cases if the population had the same stratum-specific mortality/incidence rates as in a reference area
- Adjustments (strata): age, gender ...

Indirect standardisation

$$E_i = \sum_k n_{ik} r_k$$

with

r_k: disease rate for stratum k in the reference population

n_{ik}: population at risk in area i, stratum k

If internal comparison: $\sum_{i=1}^{N} O_i = \sum_{i=1}^{N} E_i$

Expected numbers of cases - calculation

Lung cancer incidence in males, all ages, using the rates in England and Wales as reference, for the period 1985-2009

Strata	Reference area=EW			Ward A		
Age group	Population	Observed	Age-specific rate per 100,000 males	Population	Observed	Expected
	n_k	O_k	$r_k = \frac{O_k}{n_k}$	n_{ik}	O_{ik}	$E_{ik} = \frac{n_{ik} * r_k}{100000}$
0−4	41,400,692	15	0.04	11,438	0	0.00
5−9	41,143,722	6	0.01	9,697	0	0.00
10−14	41,469,696	9	0.02	9,026	0	0.00
15−19	43,087,823	39	0.09	8,650	0	0.01
20−24	45,441,353	79	0.17	12,409	0	0.02
25−29	46,873,725	172	0.37	16,963	0	0.06
30−34	46,927,658	518	1.10	17,303	0	0.19
35−39	46,936,367	1,465	3.12	13,847	0	0.43
40−44	45,304,711	4,136	9.13	11,843	1	1.08
45−49	41,657,557	9,835	23.61	9,457	5	2.23
50−54	38,451,416	20,929	54.43	8,561	3	4.66
55−59	35,842,426	40,427	112.79	7,613	8	8.59
60−64	32,480,032	68,230	210.07	6,968	5	14.64
65−69	28,231,499	95,794	339.32	6,290	15	21.34
70−74	23,315,240	110,371	473.39	5,098	27	24.13
75−79	17,297,264	102,038	589.91	4,049	22	23.89
80−84	10,498,214	68,273	650.33	2,616	20	17.01
85+	6,289,452	38,748	616.08	1,312	12	8.08
TOTAL	632,648,846	561,084		163,140	118	126.38

$\text{SIR}_A = \frac{118}{126.38} = 0.93 \rightarrow$ Fewer incident cases of lung cancer for males in ward A than expected in EW after adjusting for differences in age

Indirect standardization using R [1]

- In R we can perform indirect standardization using the package `SpatialEpi`.
- As an example we calculate the standardized incidence ratios (SIRs) of lung cancer in Pennsylvania in 2002 using the data available in the `SpatialEpi` package.
- The data contain the number of lung cancer cases and the population of Pennsylvania at county level, stratified by race (white and non-white), gender (female and male) and age (under 40, 40-59, 60-69 and 70+).
- We obtain the expected counts in each county, representing the total number of disease cases we would expect if the population in the county behaved the way the population of Pennsylvania behaves (i.e. we perform an internal standardization).

Indirect standardization using R [2]

- We use the command `expected()` from `SpatialEpi`, which takes three arguments:
  - `population`: vector of population counts for each stratum in each area
  - `cases`: vector of the observed number of cases in each area
  - `n.strata`: number of strata considered
- Caution: all counts are sorted by area first and then within each area the counts for all strata are listed (even if 0 count) in the same order.

Indirect standardization using R [3]

```
library(SpatialEpi)
library(tidyverse)

# read in the Pennsylvania data from SpatialEpi package
data("pennLC")
LC <- pennLC$data
head(LC)
```

```
> head(LC)
  county cases population race gender      age
1  adams     0        365    o      f    40.59
2  adams     1         68    o      f    60.69
3  adams     0         73    o      f      70+
4  adams     0       1492    o      f Under.40
5  adams     0        387    o      m    40.59
6  adams     0         69    o      m    60.69
```

```
# obtain the total number of cases by county
LC.OE <- group_by(LC, county) %>% summarize(O = sum(cases))
head(LC.OE)
```

```
> head(LC.OE)
# A tibble: 6 x 2
  county        O
  <fct>     <int>
1 adams        55
2 allegheny  1275
3 armstrong    49
4 beaver      172
5 bedford      37
6 berks       308
```

Indirect standardization using R [4]

```
# sort the data by county, race, gender and age
LC <- arrange(LC, county, race, gender, age)

# compute the expected (there are 2 races, 2 genders and 4 age groups for each county,
# so the number of strata is: 2 x 2 x 4 = 16)
expected <- expected(
  population = LC$population,
  cases = LC$cases, n.strata = 16)

# add the vector of the expected cases to the data frame of the cases
LC.OE$E <- expected[match(LC.OE$county, unique(LC$county))]

# compute the SIRs as the ratio of the observed to the expected counts
LC.OE$SIR <- LC.OE$O/LC.OE$E
head(LC.OE)
```

```
> head(LC.OE)
# A tibble: 6 x 4
  county        O      E   SIR
  <fct>     <int>  <dbl> <dbl>
1 adams        55   69.6 0.790
2 allegheny  1275 1182.  1.08
3 armstrong    49   67.6 0.725
4 beaver      172  173.  0.997
5 bedford      37   44.2 0.837
6 berks       308  301.  1.02
```