Spatial Statistics for Epidemiologists

Part 2: Modelling individual-level data

Leonhard Held and Birgit Schrödle

27.-29. October 2010

Preliminaries Modelling disease risk in relation to a point source

Brief review of generalized linear regression

The linear regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \text{ with } \epsilon_i \sim N(0, \sigma^2)$$

can be re-written as $Y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$.
Two generalizations are now possible:

1. Different distributions for the outcome variable Y_i (Binomial, Poisson, ...)
2. The expected value $\mathrm{E}(Y_i) = \mu_i$ is related through a link-funktion h with the linear predictor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$:

$$h(\mu_i) = \eta_i \text{ bzw. } \mu_i = h^{-1}(\eta_i)$$

Preliminaries Modelling disease risk in relation to a point source

Examples

1. Log-linear Poisson regression

$$\begin{aligned} Y_i &\sim Po(\lambda_i) \\ \log(\lambda_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

2. Log-linear Poisson regression with offset

$$\begin{aligned} Y_i &\sim Po(\lambda_i) \\ \log(\lambda_i) &= \log(e_i) + \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

3. Logistic regression

$$\begin{aligned} Y_i &\sim B(n_i, \pi_i) \\ \mathrm{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

Preliminaries Modelling disease risk in relation to a point source

Parameter estimation and interpretation

- The unknown parameter vector $\boldsymbol{\beta}$ is estimated through Maximum-Likelihood. $\rightarrow \hat{\boldsymbol{\beta}}_{ML}$
- Let β_i denote the i-th component of $\boldsymbol{\beta}$. Interpretation of β_i depends on the underlying model:
  - For log-linear regression, $\exp(\beta_i)$ is the relative risk
  - For logistic regression $\exp(\beta_i)$ is the odds ratio

Relative risk and odds ratio

- Relative risk (RR) is the probability that a member of an exposed group will develop a disease relative to the probability that a member of an unexposed group will develop that same disease:

$$\text{RR} = \frac{\Pr(\text{disease}|\text{exposed})}{\Pr(\text{disease}|\text{unexposed})}$$

- Odds ratio (OR) is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

$$\text{OR} = \frac{\Pr(\text{disease}|\text{exposed})/(1-\Pr(\text{disease}|\text{exposed}))}{\Pr(\text{disease}|\text{unexposed})/(1-\Pr(\text{disease}|\text{unexposed}))}$$

RR/OR $= 1$	no association between exposure and disease
RR/OR > 1	increased risk of disease among those that have been exposed
RR/OR < 1	decreased risk of disease among those that have been exposed

Wald-Test

Under $H_0 : \beta_i = 0$ the following holds:

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \overset{a}{\sim} N(0,1) \text{ and } \frac{\hat{\beta}_i^2}{[se(\hat{\beta}_i)]^2} \overset{a}{\sim} \chi_1^2$$

where

- $\hat{\beta}_i$ is the i-th component of the ML estimate $\hat{\boldsymbol{\beta}}_{ML}$
- $se(\hat{\beta}_i)$ is the corresponding standard error

Log likelihood

- The log likelihood $l(\boldsymbol{\beta})$, evaluated at the ML-estimate $\hat{\boldsymbol{\beta}}_{ML}$ is a goodness-of-fit measure: The larger $l(\hat{\boldsymbol{\beta}}_{ML})$, the better the model fit.
- A common reference model is the saturated model, the most general model possible. The saturated model typically has as many parameters as there are observations (suitably grouped). It has log likelihood $l(\hat{\boldsymbol{\beta}}_{max})$.

Deviance

- It is quite common to consider the deviance as a model choice criterion, also known as likelihood ratio statistic (LR statistic):

$$W = 2(l(\hat{\boldsymbol{\beta}}_{max}) - l(\hat{\boldsymbol{\beta}}_{ML}))$$

- If the data are a random sample from the chosen model, then W is (asymptotically) χ_p^2-distributed with p degrees of freedom, the difference in the number of parameters of the saturated and the chosen model.
- Large values of the deviance are evidence against the chosen model.

$\rightarrow$ Calculation of p-value

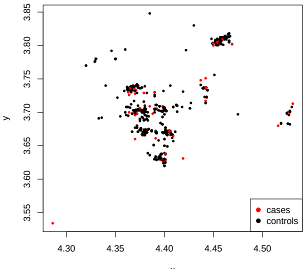
Likelihood ratio test

- The deviance is the test statistic of the LR ratio test comparing the chosen with the saturated model.
- To compare two non-saturated models, we compute the difference of the deviances.
- If the data come from the simpler model, then the deviance is asymptotically χ^2-distributed with degrees of freedom equal to the difference of the number of parameters of the two models.
- Large values of the LR test statistic speak against the simpler model.
- A comparison is only possible, if the two models are nested.

Spatial case-control studies

- Interest is in possible spatial variation of disease risk
- Adjustments for underlying population necessary
- The spatial case-control design samples controls randomly from the underlying population at risk
  - Completely randomized design
  - Stratified randomized design
  - Matched design

Example: Childhood asthma in North Derbyshire, England



Explanatory variables

```
> print(names(asthma))

 [1] "School"     "Id"         "PostCode"   "HouseType"  "OpenFire"
 [6] "GasFire"    "GasCooker"  "GasCH"      "OilCH"      "DuctCH"
[11] "PRFire"     "DoubGlaz"   "Damp"       "Mould"      "Conden"
[16] "Dust"       "Cold"       "NumRooms"   "NumPpl"     "Dog"
[21] "Cat"        "FurFeath"   "NotFurFe"   "Nsmokers"   "Cigs"
[26] "Cigars"     "Pipe"       "Nsmoked"    "Asthma"     "Eczema"
[31] "HayFever"   "NWheezes"   "CExercise"  "CSleep"     "CIllness"
[36] "CAllergies" "CStress"    "NYearsSuf"  "Medcat"     "Age"
[41] "Gender"     "YearsHere"  "Income"     "WardCode"   "x"
[46] "y"          "roaddist2"  "road"       "d2source1"  "d2source2"
[51] "d2source3"  "d2source1B" "d2source2B"
```

Potential sources of hazard:

- Main road network: `roaddist2`
- Coking works: `d2source1`
- Chemical plant: `d2source2`
- Waste treatment center: `d2source3`

Many confounding factors, e.g.

- whether the household included at least one cigarette smoker
- whether the child suffered from hay fever

Contingency table analysis

```
> threshold <- 0.01
> exposure <- (d2source1 <= threshold)
> t <- table(exposure, Asthma)
> print(t)

        Asthma
exposure   0   1
   FALSE 327  48
   TRUE  749 168

> library(vcd)
> print(summary(oddsratio(t)))

     Log Odds Ratio Std. Error z value Pr(>|z|)
[1,]        0.42398    0.17590  2.4104 0.007968 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```


Regression modelling

- It is tempting to use logistic regression to study the risk in relation to a point source
- To model the exposure effect it is common to use
  - a binary indicator whether or not the individual is living within some subjectively chosen distance from the point source (can be extended to distance categories)
  - inverse distance from the point source

  as explanatory variable
- This approach allows easily to adjust for confounders

Logistic regression analysis with binary indicator

```
> m0 <- glm(Asthma ~ exposure, family = binomial)
> print(summary(m0))

Call:
glm(formula = Asthma ~ exposure, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6362  -0.6362  -0.6362  -0.5234   2.0277

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.9188     0.1546 -12.414   <2e-16 ***
exposureTRUE   0.4240     0.1766   2.401   0.0163 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1166.4  on 1291  degrees of freedom
Residual deviance: 1160.3  on 1290  degrees of freedom
AIC: 1164.3

Number of Fisher Scoring iterations: 4
```


Adjusting for confounders

```
> m1 <- glm(Asthma ~ exposure + HayFever + (Nsmokers >= 1), family = binomial)
> print(summary(m1))

Call:
glm(formula = Asthma ~ exposure + HayFever + (Nsmokers >= 1),
    family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9822  -0.6091  -0.5522  -0.4635   2.1375

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.1770     0.1704 -12.773  < 2e-16 ***
exposureTRUE        0.3735     0.1811   2.063   0.0392 *
HayFever            1.1122     0.1824   6.098 1.08e-09 ***
Nsmokers >= 1TRUE   0.2130     0.1552   1.373   0.1699
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1166.4  on 1291  degrees of freedom
Residual deviance: 1125.3  on 1288  degrees of freedom
AIC: 1133.3

Number of Fisher Scoring iterations: 4
```

Using inverse distance

```
> inv.d2source1 <- 1/d2source1
> m2 <- glm(Asthma ~ inv.d2source1 + HayFever + (Nsmokers >= 1),
+     family = binomial)
> print(summary(m2))

Call:
glm(formula = Asthma ~ inv.d2source1 + HayFever + (Nsmokers >=
    1), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0100  -0.5852  -0.5190  -0.5162   2.0403

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.948e+00  1.167e-01 -16.702  < 2e-16 ***
inv.d2source1      4.659e-06  5.799e-06   0.803   0.4217
HayFever           1.113e+00  1.824e-01   6.100 1.06e-09 ***
Nsmokers >= 1TRUE  2.679e-01  1.533e-01   1.747   0.0806 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1166.4  on 1291  degrees of freedom
Residual deviance: 1129.1  on 1288  degrees of freedom
AIC: 1137.1

Number of Fisher Scoring iterations: 4
```


Discussion

- A logistic regression analysis allows for adjustment for confounders
- However, all possible risk models (as a function of distance u) are rather simplistic:
  - Distance categories with predefined cutpoints
  - Using inverse distance is unrealistic: risk goes to 1 for $u \to 0$
- It is better to use a non-linear regression approach to model the risk in relation to the distance from the point source

Modelling disease risk in relation to a point source

- Assume cases and controls form independent point processes with intensities $\lambda_1(\mathbf{s}) = \rho\lambda_2(\mathbf{s})\varphi(\mathbf{s})$ and $\lambda_2(\mathbf{s})$.
- Then the probability to be a case rather than a control is

$$p(\mathbf{s}) = \frac{\rho \cdot \varphi(\mathbf{s})}{1 + \rho \cdot \varphi(\mathbf{s})}$$

- It is common to model $\varphi(\mathbf{s})$ as a function of the distance u from a point source.

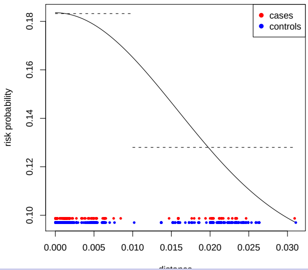
Disease risk as a function of distance

- Isotropic Gaussian model as function of distance u:

$$\varphi(u) = 1 + \alpha \exp\{-(u/\beta)^2\}$$

$\to$ Smoothly decaying risk function with $\varphi(0) \neq \infty$ and $\varphi(\infty) > 0$.
- Can be extended to allow for directional effects

Example: Childhood asthma in North Derbyshire, England



Statistical significance

- Statistical significance can be investigated using the deviance
- Under H_0, the residual deviance is χ^2-distributed with 2 degrees of freedom.
- Here we obtain the deviance 1160.4 in the fitted model and 1166.4 in the null model. The test statistic is 5.9 with associated p-value 0.051.
- Adjustments for confounding factors can be added.

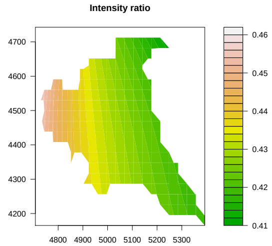
Analysis using R function `tribbleNG()`

```
> library(tribble2)
> results <- tribbleNG(ccflag = Asthma, rho = sum(Asthma == 1)/sum(Asthma ==
+     0), distances = d2source1B, alpha = 1, beta = 1)
> print(results)

Tribble model:
            value        se         Z
rho    0.09468151 0.1332499 0.7105560
alpha1 1.37491422 3.3279666 0.4131394
beta1  2.06547580 1.8627879 1.1088089

null log-likelihood:  -583.1968
     log-likelihood:  -580.2229

D:  5.947726  df:  2  p:  0.05110552
```


Spatial variation in risk

- Interest is now in non-parametric estimation of risk surface
- In the case-control setting one option is to estimate the two intensities separately via kernel smoothing and then to consider the ratio of the two intensities, which can be interpreted as a spatially-varying disease odds.
- The same smoothing parameter h should be used for cases and controls, irrespectively of the numbers of cases and controls.
- The smoothing parameter should be determined through cross-validation

Example: Childhood leukaemia in North Humberside, UK, 1974-82



Adjustments for covariates

- An alternative approach has been described based on the formulation

$$\log \frac{P(Y_i = 1)}{P(Y_i = 0)} = \mathbf{x}_i^T \boldsymbol{\beta} + l(\mathbf{s}_i)$$

  where $\mathbf{s}_i$ is the spatial location of the i-th individual and l is a spatially-varying function.
- This allows for adjustments for covariates and only one spatial surface remains to be estimated nonparametrically

Matched case-control studies in spatial epidemiology

- Matched case-control studies are often used to eliminate the effects of confounding variables on the presumption that these are not of scientific interest.
- Application of conditional logistic regression
- However, if the matching variables are not spatially neutral then they are confounded with the spatial effect, which typically is of scientific interest.
- This makes the interpretation of the resulting estimated spatial effect problematic.

Discussion

- Spatial case-control studies are attractive designs to
  - study disease risk in relation to a point source
  - estimate (residual) spatial variation

  through regression modelling.
- Matched spatial case-control studies
  - require the application of conditional logistic regression
  - are difficult to interpret if the matching variables are not spatially neutral