Statistical Methods in Cancer Epidemiology using R

Janne Pitkäniemi

Faculty of Social Sciences, University of Tampere
Finnish Cancer Registry

Lecture 2b

janne.pitkaniemi@cancer.fi

Feb,17 2020

Basic analysis of rates

- Person-time data, hazard and incidence rates,
- Comparative parameters of rates and their estimation,
- Poisson regression models and comparative parameters,
- Adjustment for confounding and evaluation of modification by Poisson regression,
- Goodness-of-fit evaluation.

Main R functions covered:

- `glm()`
- tools for extracting results from a `glm` model object

Person-time data and incidence rates

Summarized data on outcome from cohort study, in which two exposure groups, as to binary risk factor X, have been followed-up over individually variable times.

Exposure to risk factor	Number of cases	Person-time
yes	D_1	Y_1
no	D_0	Y_0
total	D_+	Y_+

Empirical **incidence rates** by exposure group:

$$I_1 = D_1/Y_1, \qquad I_0 = D_0/Y_0.$$

These provide estimates for the true $\{$**hazards**$\}$ **(or hazard rates)** λ_1 **and** λ_0 **assumed constant within exposure categories.**

Hazards and their comparison

Parameters of interest:

- **hazard ratio**

$$\rho = \frac{\lambda_1}{\lambda_0} = \frac{\text{hazard among exposed}}{\text{hazard among unexposed}}.$$

- **hazard difference**

$$\delta = \lambda_1 - \lambda_0$$

Null hypothesis $H_0 : \rho = 1 \Leftrightarrow \delta = 0 \Leftrightarrow$ exposure has no effect.

Estimation of hazard ratio

Point estimator of true hazard ratio ρ: empirical **incidence rate ratio** (IR)

$$\widehat{\rho} = \mathrm{IR} = \frac{I_1}{I_0} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}.$$

NB. The last form is particularly useful = **exposure odds ratio** (EOR).

Standard error of log(IR), 95% {error factor} & 95% CI for ρ:

$$SEL = \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}$$

$$EF = \exp\{1.96 \times \mathrm{SEL}\}$$

$$CI = [\mathrm{IR}/\mathrm{EF},\ \mathrm{IR} \times \mathrm{EF}].$$

NB. Random error depends inversely on numbers of cases.

Estimation of hazard difference

Point estimator of true hazard difference δ: empirical **incidence rate difference** (ID)

$$\widehat{\delta} = \mathsf{ID} = I_1 - I_0 = \frac{D_1}{Y_1} - \frac{D_0}{Y_0}$$

Standard error of ID, 95% error margin & 95% CI

$$\mathrm{SE} = \sqrt{\frac{I_1^2}{D_1} + \frac{I_0^2}{D_0}}$$

$$\mathrm{EM} = 1.96 \times \mathrm{SE}$$

$$\mathrm{CI} = [\mathsf{ID} - \mathrm{EM},\ \mathsf{ID} + \mathrm{EM}]$$

NB. Random error again depends inversely on no. of cases.

Example. British doctors' study (Doll & Hill 1966)

CHD mortality in males by smoking and age.\ Cases (D), person-years (Y), and mortality rates (I per 10^4 y).

	Smokers			Non-smokers		
Age(y)	D	Y	I	D	Y	I
35-44	32	52407	6	2	18790	1
45-54	104	43248	24	12	10673	11
55-64	206	28612	72	28	5710	49
65-74	186	12663	147	28	2585	108
75-84	102	5317	192	31	1462	212
Total	630	142247	44	101	39220	26

Example (cont'd).

Crude incidence rates:
$I_1 = 630/142247$ y $= 44.3$ per 10^4 y, and
$I_0 = 101/39220$ y $= 25.8$ per 10^4 y.

Crude estimate of overall hazard ratio ρ with SE, etc.

$$\widehat{\rho} = \mathrm{IR} = \frac{44.3}{25.8} = \mathbf{1.72}$$

$$\mathrm{SEL} = \sqrt{\frac{1}{630} + \frac{1}{101}} = \mathbf{0.1072}$$
$$\mathrm{EF} = \exp(1.96 \times 0.1072) = \mathbf{1.23}$$

95% CI for ρ:

$$[1.72/1.23,\ 1.72 \times 1.23] = [\mathbf{1.39},\ \mathbf{2.12}]$$

Two-tailed P < 0.001 .

Poisson regression model for rate ratio

- *Random part*: Number of cases in exposure group $j = 0, 1$

$$D_j \sim \text{Poisson}(\lambda_j Y_j),$$

  where $\mu_j = \lambda_j Y_j =$ *expected number* of cases.

- *Systematic part & link function*:
  linear predictor $\alpha + \beta X_j$ with *logarithmic* (log) link

$$\log(\lambda_j) = \alpha + \beta X_j,$$

  equivalently on the original hazard scale:

$$\lambda_j = \exp(\alpha + \beta X_j).$$

Poisson model for rate ratio (cont'd)

Interpretation,

- $X_j = \begin{cases} 1 & \text{if exposed } (j = 1), \\ 0 & \text{if unexposed } (j = 0), \end{cases}$
- $\alpha = \log(\lambda_0)$, log-baseline rate,
- $\beta = \log(\rho) = \log(\lambda_1/\lambda_0)$, logarithm of true hazard ratio,
- $e^{\beta} = \rho =$ true hazard ratio.

Special case of generalized linear models!

Example. Crude analysis of CHD mortality in R

A ready data frame contains

- four variables:
  - `age` = age group – a factor with 5 levels,
  - `smok` = smoking: 1 = `yes`, 0 = `no`,
  - `d` = number of cases,
  - `y` = person-years.
- 10 observations (one for each age-smoking combination).

Example. Analysis of CHD rates (cont'd)

	age	smok	d	y	rate
1	35-44	1	32	52407	6.1
2	35-44	0	2	18790	1.1
3	45-54	1	104	43248	24.0
4	45-54	0	12	10673	11.2
5	55-64	1	206	28612	72.0
6	55-64	0	28	5710	49.0
7	65-74	1	186	12663	146.9
8	65-74	0	28	2585	108.3
9	75-84	1	102	5317	191.8
10	75-84	0	31	1462	212.0

Fitting Poisson model for crude rate ratio

Poisson model with log-link (default) for crude rates

```
Call:
glm(formula = d/y ~ smok, family = poisson(), data = bd, weights = y)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.535   -6.031    4.612    8.162   13.644

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9618     0.0995 -59.916  < 2e-16 ***
smok          0.5422     0.1072   5.059 4.22e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.07  on 9  degrees of freedom
Residual deviance: 905.98  on 8  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 8
```

Fitting crude rate ratio model (cont'd)

Main results:

- $\widehat{\alpha} = -5.96 = \log(25.8/10^4 \text{ y}), \quad (\text{SE} = 0.10),$
- $\widehat{\beta} = 0.54 = \log(1.72), \quad (\text{SE} = 0.11)$

Function `ci.lin()` transforms results to ratio scale

```
              Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)  -5.9618 0.0995    0.0026 0.0021 0.0031
smok          0.5422 0.1072    1.7198 1.3940 2.1219
```

Compare the results with those obtained above using simple estimation & SE formulas.

Fitting crude rate ratio model (cont'd)

The Poisson model above can also be fitted as follows:

glm(d $\sim$ smok, fam =poisson(), offset=log(y))

Here `offset` refers to the logarithm of person-years `y` in formula for expected numbers of cases $\mu_j = \lambda_j \times Y_j$:

$$\log(\mu_j) = \log(\lambda_j Y_j) = \log(Y_j) + \log(\lambda_j) = \log(Y_j) + \alpha + \beta X_j,$$

$\log(Y_j)$ is an **offset** term in the linear predictor, meaning that it has a fixed value 1 for the regression coefficient.

Stratified analysis

Stratification of cohort data with person-time
– at each level k of covariate Z results are summarized:

Exposure to risk factor	Number of cases	Person-time
yes	D_{1k}	Y_{1k}
no	D_{0k}	Y_{0k}
Total	D_{+k}	Y_{+k}

Stratum-specific rates by exposure group:

$$I_{1k} = \frac{D_{1k}}{Y_{1k}}, \qquad I_{0k} = \frac{D_{0k}}{Y_{0k}}.$$

Stratum-specific comparisons

Let λ_{jk} be true rate for exposure group j ($j = 0, 1$) and stratum k ($k = 0, \ldots, K$). Let also

$$\rho_k = \frac{\lambda_{1k}}{\lambda_{0k}}, \qquad \delta_k = \lambda_{1k} - \lambda_{0k}$$

be the rate ratios and rate differences between the exposure groups in stratum k.

Two simple models assuming homogeneity:

- common rate ratio: $\rho_k = \rho$ for all k,
- common rate difference: $\delta_k = \delta$ for all k.

Only one of these can in principle hold. However, almost always neither homogeneity assumption is exactly true.

Example. British male doctors (cont'd)

CHD mortality rates (per 10^4 y) and numbers of cases (D) by age and cigarette smoking.

Mortality rate differences (ID) and ratios (IR) in age strata.

```
bd$ID<-0.0
bd$ID[bd$smok==1]<-bd$rate[bd$smok==1]-bd$rate[bd$smok==0]
bd$IR<-1.0
bd$IR[bd$smok==1]<-round(bd$rate[bd$smok==1]/bd$rate[bd$smok==0],1)
bd[order(bd$age,bd$smok),]
```

```
     age smok   d     y  rate    ID  IR
2  35-44    0   2 18790   1.1   0.0 1.0
1  35-44    1  32 52407   6.1   5.0 5.5
4  45-54    0  12 10673  11.2   0.0 1.0
3  45-54    1 104 43248  24.0  12.8 2.1
6  55-64    0  28  5710  49.0   0.0 1.0
5  55-64    1 206 28612  72.0  23.0 1.5
8  65-74    0  28  2585 108.3   0.0 1.0
7  65-74    1 186 12663 146.9  38.6 1.4
10 75-84    0  31  1462 212.0   0.0 1.0
9  75-84    1 102  5317 191.8 -20.2 0.9
```

```
rbind(bd,c("Crude",NA, sum(bd$d), sum(bd$y),sum(bd$d)/sum(bd$y),NA,NA))
```

Example (cont'd).

-Both types of comparative parameter, rate ratios ρ_k and rate differences δ_k appear heterogenous, because

- ID increases by age – at least up to 75 y,
- IR decreases by age.
- Part of this observed heterogeneity may be due to random variation.
- Yet, any single-parameter comparison by common rate ratio or rate difference

may not adequately capture the joint pattern of true rates.

$\Rightarrow$ Effect modification must be evaluated.