Statistical Methods in Cancer Epidemiology using R

Janne Pitkäniemi

Faculty of Social Sciences, University of Tampere
Finnish Cancer Registry

janne.pitkaniemi@cancer.fi

Lecture 2b

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)

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

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

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

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

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

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 $\mathrm{P} < 0.001$.

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

```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -5.9618216 0.09950368 -59.915588 0.000000e+00
smok         0.5422211 0.10718341   5.058816 4.218682e-07
```

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 ~ 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.

	age	smok	d	y	rate	ID	IR
2	35-44	0	2	18790	1.1	0	1
1	35-44	1	32	52407	6.1	5	5.5
4	45-54	0	12	10673	11.2	0	1
3	45-54	1	104	43248	24	12.8	2.1
6	55-64	0	28	5710	49	0	1
5	55-64	1	206	28612	72	23	1.5
8	65-74	0	28	2585	108.3	0	1
7	65-74	1	186	12663	146.9	38.6	1.4
10	75-84	0	31	1462	212	0	1
9	75-84	1	102	5317	191.8	-20.2	0.9
12	<NA>	0	101	39220	26	0	1
11	<NA>	1	630	142247	44	0	1

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.

Rate ratio adjustment by Poisson model

Define Poisson regression model for

- one binary exposure variable X and
- one categorical (polytomous) factor Z.
  - *Random part*: No. of cases in exposure group j ($j = 0, 1$) and covariate level k ($k = 1, \ldots, K$) is $D_{jk} \sim Poisson(\lambda_{jk} Y_{jk})$
  - *Systematic part*: $log(\lambda_{jk}) = \alpha + \beta X_j + \gamma_k$, where X_j is (0/1)
- $\alpha = \log(\lambda_{01}) =$ log-baseline rate,
- $\gamma_k = \log(\lambda_{jk}/\lambda_{j1})$,
- $\beta = \log(\rho) = log(\lambda_{1k}/\lambda_{0k})$,
- $e^{\beta} = \rho =$ true rate ratio for the effect of exposure to X.

How do we read this?

Implications of model definition

- homogeneity of true rate ratio $\rho_k = \rho$ for X across levels of Z is assumed,
- inclusion of Z leads to adjustment for Z in estimating the common effect of X,
- e^{γ_k} = rate ratio for level k of Z *vs.* level 1 is the same in both exposure groups ($j = 0, 1$)
  $\Rightarrow$ homogeneity of the effect of Z is assumed, too.
- level $k = 1$ is chosen as the *reference* level for Z (like "unexposed" is reference for X),
- before model fitting, binary *indicator* variables Z_k for levels $k = 1, \ldots, K$ of Z must be defined:

$$Z_k = \begin{cases} 1, & \text{if observation belongs to level } k, \\ 0, & \text{otherwise.} \end{cases}$$

Example. CHD in British doctors (cont'd)

Factor age has 5 levels.

Indicator variables for each age level are generated in R when defining the model, and the following model matrix is returned.

```
 m2 <- glm( d/y ~ age + smok,family=poisson(link=log),
            weights=y,data=bd)
cbind(data.frame(bd$age), model.matrix(m2))
```

	bd.age	(Intercept)	age45-54	age55-64	age65-74	age75-84	smok
1	35-44	1	0	0	0	0	1
2	35-44	1	0	0	0	0	0
3	45-54	1	1	0	0	0	1
4	45-54	1	1	0	0	0	0
5	55-64	1	0	1	0	0	1
6	55-64	1	0	1	0	0	0
7	65-74	1	0	0	1	0	1
8	65-74	1	0	0	1	0	0
9	75-84	1	0	0	0	1	1
10	75-84	1	0	0	0	1	0

Summary of the adjustment model

```
coef(summary(m2))
```

```
                Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -7.9193257  0.1917618 -41.297719 0.000000e+00
age45-54     1.4840070  0.1951034   7.606260 2.821410e-14
age55-64     2.6275051  0.1837273  14.301117 2.153261e-46
age65-74     3.3504928  0.1847992  18.130453 1.832443e-73
age75-84     3.7000965  0.1922195  19.249328 1.430052e-82
smok         0.3545356  0.1073741   3.301872 9.604175e-04
```

Fitting adjustment model (cont'd)

Results on the ratio scale

```
round(ci.lin(m2, Exp=T)[, 5:7], 4 )
```

```
             exp(Est.)    2.5%   97.5%
(Intercept)     0.0004  0.0002  0.0005
age45-54        4.4106  3.0090  6.4650
age55-64       13.8392  9.6543 19.8381
age65-74       28.5168 19.8518 40.9640
age75-84       40.4512 27.7533 58.9589
smok            1.4255  1.1550  1.7594
```

$\Rightarrow$ Age-adjusted rate ratio [95% CI] for smoking:

$$\widehat{\rho} = \mathbf{1.43}\ [\mathbf{1.16}, \mathbf{1.76}]$$

Fitted values & residuals

From the estimated coefficients we can calculate **fitted** linear predictors $\widehat{\eta}_{jk}$, hazards $\widehat{\lambda}_{jk}$ and numbers $\widehat{\mu}_{jk}$:

$$\widehat{\eta}_{jk} = \widehat{\alpha} + \widehat{\beta}x + \widehat{\gamma}_k$$

$$\widehat{\lambda}_{jk} = \exp(\widehat{\eta}_{jk}), \quad \widehat{\mu}_{jk} = \widehat{\lambda}_{jk} Y_{jk}$$

In R the two first can be extracted directly from the\ fitted model object `m2`:

```
bd$fit.eta <- m2$linear.predictor
bd$fit.rate <- fitted(m2);
bd$fit.mu <- bd$y*bd$fit.rate
```

NB. If count `d` is declared as response with `log(y)` as offset, then `fitted()` returns the fitted numbers of cases $\widehat{\mu}_{jk}$.

Fitted values & residuals (cont'd)

Deviance residual for cell jk (`resid(m2)` in R):

$$d_{jk} = \text{sign}(Y_{jk} - \widehat{\mu}_{jk}) \times \sqrt{2\left\{Y_{jk} \log\left(\frac{Y_{jk}}{\widehat{\mu}_{jk}}\right) - (Y_{jk} - \widehat{\mu}_{jk})\right\}}$$

Pearson residual (`resid(m2, type="pearson")`):

$$r_{jk} = \frac{Y_{jk} - \widehat{\mu}_{jk}}{\sqrt{\widehat{\mu}_{jk}}}.$$

Small value of either residual
$\rightarrow$ consistency of observation with model.

"Large" (in absolute value) residual
$\rightarrow$ lack of fit for that cell.

Example. Fitted values & residuals

```
bd$fit.rate<-round(10000*bd$fit.rate,1)
bd$fit.mu<-round(bd$fit.mu,1)
bd$res.D<-round(resid(m2,"dev"),2)
bd$res.P<-round(resid(m2,"pear"),2)
data.frame(bd$age,bd$smok,bd$d,bd$rate,bd$fit.rate,bd$fit.mu,bd$res.D,bd$res.P)
```

	bd.age	bd.smok	bd.d	bd.rate	bd.fit.rate	bd.fit.mu	bd.res.D	bd.res.P
1	35-44	1	32	6.1	5.2	27.2	0.90	0.93
2	35-44	0	2	1.1	3.6	6.8	-2.18	-1.85
3	45-54	1	104	24.0	22.9	98.9	0.51	0.51
4	45-54	0	12	11.2	16.0	17.1	-1.31	-1.24
5	55-64	1	206	72.0	71.7	205.3	0.05	0.05
6	55-64	0	28	49.0	50.3	28.7	-0.14	-0.14
7	65-74	1	186	146.9	147.8	187.2	-0.09	-0.09
8	65-74	0	28	108.3	103.7	26.8	0.23	0.23
9	75-84	1	102	191.8	209.7	111.5	-0.91	-0.90
10	75-84	0	31	212.0	147.1	21.5	1.92	2.05

NB! Fitted rate ratios between smokers and non-smokers:

$$\frac{5.2}{3.6} = \cdots = \frac{209.7}{147.1} = 1.43 \text{ at each age level}$$

Summary goodness-of-fit statistics

- **Deviance** (Dev)
  = sum of squares of deviance residuals,
- **Pearson's** X^2 (chi-square) = s.s. of Pearson residuals

Assume that

- the model is "true", and
- the data are not *sparse*, i.e. numbers of cases in each line of the (grouped) data matrix are sufficiently big.

Under these assumptions

- residuals d_{jk} and r_{jk} all approximately Normal(0,1),
- Both Dev and X^2 are approximately χ^2-distributed with *degrees of freedom* (df) being equal to the *residual* df = no. of obs – no. of coeff's
- Expected value of these statistics = residual df.

Goodness of fit (cont'd)

In our example

- "large" residuals in extreme age groups,
- X^2 and Dev 2.8 to 3 times higher than df.

$\rightarrow$ Some lack of fit with assumed model.

(P-values: $\mathrm{P}(X^2 > 11.2) = 0.025$, and
$\mathrm{P}(\mathrm{Dev} > 12.1) = 0.017$)

Possible causes for lack of fit

- Unrealistic random part (Poisson)?

- Misspecified systematic part? - Wrong functional form for effects? - Missing important risk factors? - Failure to take into account *effect modification* or "interaction".

In this example the last possibility is very likely:
the rate ratios are clearly heterogenous.

Yet, inclusion of 4 interaction parameters would lead to a saturated model.

Continuous covariates

Stratified analysis is adequate with

- qualitative exposure factor with few levels,
- few qualitative covariates with few levels.

Effects of continous factors can often be described by smooth functions with small number of parameters.

Stratification requires categorization of continuous variables and level specific parameters

$\Rightarrow$ loss of efficiency.

Continuous covariates (cont'd)

Regression models can accommodate functions of continuous variables

- (+) more parsimonious models,
- (+) greater flexibility in modelling,
- (-) more opportunities for misspecification, *i.e.* creation of "wrong" models.

Example (cont'd)

Treat age as continuous variable taking midpoint of each ageband (40, 50, . . ., 80) as quantitative age value.

Include linear and quadratic term of age applying

- centering at 60 years, and
- scaling, *i.e.* division by 10 years

This is achieved by

```
bd$A.L <- as.numeric(bd$age)-3
bd$A.Q <- bd$A.L*bd$A.L - 2
```

Example (cont'd)}

Model with linear and quadratic terms for age, and interaction between linear term of age and smoking:

```
 m3 <- glm( d/y ~ A.L + A.Q + smok + A.L:smok,
 family=poisson( ), weights=y,data=bd )
round(ci.lin(m3, Exp=T)[, -(3:4)], 4 )
```

	Estimate	StdErr	exp(Est.)	2.5%	97.5%
(Intercept)	-5.8368	0.1213	0.0029	0.0023	0.0037
A.L	1.1904	0.0923	3.2885	2.7441	3.9408
A.Q	-0.1977	0.0274	0.8206	0.7778	0.8659
smok	0.5183	0.1262	1.6792	1.3112	2.1506
A.L:smok	-0.3075	0.0970	0.7352	0.6079	0.8893

Interpretation of parameters?

Example (cont'd)

Observed and fitted rates & fitted rate ratios between smokers and non-smokers in each ageband

	bd.age	bd.smok	bd.d	bd.rate	bd.fit.rate	bd.fit.mu	bd.res.D	bd.res.P
1	35-44	1	32	6.1	5.6	29.6	0.44	0.44
2	35-44	0	2	1.1	1.8	3.4	-0.83	-0.77
3	45-54	1	104	24.0	24.7	106.8	-0.27	-0.27
4	45-54	0	12	11.2	10.8	11.5	0.13	0.13
5	55-64	1	206	72.0	72.8	208.2	-0.15	-0.15
6	55-64	0	28	49.0	43.3	24.7	0.64	0.65
7	65-74	1	186	146.9	144.4	182.8	0.23	0.23
8	65-74	0	28	108.3	116.9	30.2	-0.41	-0.41
9	75-84	1	102	191.8	192.9	102.6	-0.06	-0.06
10	75-84	0	31	212.0	212.5	31.1	-0.01	-0.01

Residual Dev $= 1.64$ df $= 5$; The fit is excellent!

Some closing remarks

- **Hazard differences** can also be estimated by Poisson-modelling, because `link='identity'` can be coupled with `family=poisson` in `glm()`.
- Poisson models can also be fitted on **ungrouped data**. Units of observations may even be shorter intervals of total individual follow-up times:
- Each member of the study population can contribute several **separate time-slots** as observational units.
- More about **time-splitting** in the next lectures.
- With continuous covariates it may be difficult to keep hazards positive when the *additive hazards* model with `link='identity'` is attempted to be fitted.