Epidemiologic Data Analysis using R
Part 6: Analysis of case-control studies

Janne Pitkäniemi
(Esa Läärä)

Finnish Cancer Registry, Finland, <janne.pitkaniemi@cancer.fi>
(University of Oulu, Finland, <esa.laara@oulu.fi>)

University of Tampere
Faculty of Social Sciences
Feb 26- Apr 9 2018

Contents

1. Case-control designs
2. Exposure odds ratio (EOR) and its interpretation
3. Estimation of EOR by logistic regression
4. Matched case-control studies

Main R functions to be covered

- `glm()`

Case-control design

- From given study population (base pop'n) are selected all or a random sample of
  - D **cases**, or individuals with the disease being diagnosed during certain period
  - C **controls**, or "healthy" individuals at risk.
- Exposure to risk factor X and other covariates assessed in cases and chosen controls.
- To increase efficiency and remove confounding, the sampling of controls is often *stratified* or individually *matched* for age, gender, place of residence, *etc*.

Exposure odds ratio (EOR)

With binary risk factor X the results are summarized:

Exposure	Cases	Controls	Total
yes ($X = 1$)	D_1	C_1	T_1
no ($X = 0$)	D_0	C_0	T_0
Total	D	C	T

Common effect measure:

- **exposure odds ratio**

$$\text{EOR} = \frac{D_1/D_0}{C_1/C_0} = \frac{D_1 C_0}{D_0 C_1}$$

Precision in EOR

Standard error of log(EOR), 95% error factor (EF) & 95% confidence interval (CI) for the associated parameter:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{D_1}+\frac{1}{D_0}+\frac{1}{C_1}+\frac{1}{C_0}} \\ \mathrm{EF} &= \exp\{1.96 \times \mathrm{SEL}\} \\ \mathrm{CI} &= [\mathrm{EOR}/\mathrm{EF},\ \mathrm{EOR} \times \mathrm{EF}]. \end{aligned}$$

NB. Random error depends inversely on numbers of cases and controls.

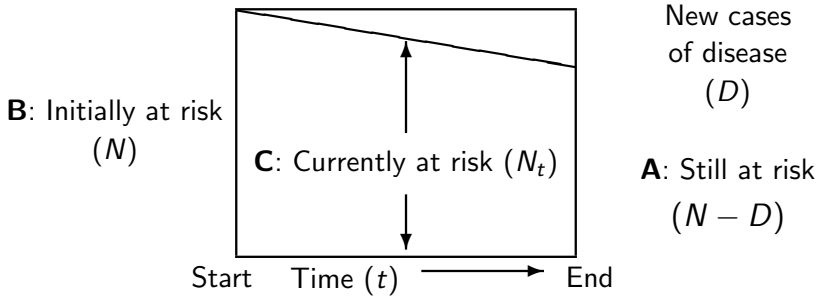
What parameter is estimated by EOR?

The answer depends on

- type of base population, from which cases emerge
  - closed population or **cohort**, or
  - open or **dynamic population**,
- time dimensionality
  – longitudinal or cross-sectional
- sampling principle of controls:
  (A) case-noncase sampling (epidemic ca-co study)
  (B) case-cohort sampling
  (C) density sampling (incl. nested case-control study)

Sampling controls from a longitudinal base

Simplified ideal situation:
Complete follow-up of a cohort of initially healthy subjects with no losses during a fixed risk period.



Possible sampling frames: **A, B** and **C**

Sampling schemes or designs for controls

A: Case-noncase sampling

- Controls chosen from those $N - D$ subjects still at risk (healthy) *at the end* of the follow-up.

B: Case-cohort sampling:

- The control group or **subcohort** is a random sample of the whole cohort (N) *at the start* of the follow-up.

B: Density sampling:

- Controls drawn *during the follow-up* from those currently at risk.
- **Nested case-control design (NCC)**
  A set of controls is sampled from the *risk set at each time t of diagnosis* of a new case,

EOR in case-noncase sampling design

- In the traditional or epidemic case-control study the controls are selected from those still healthy at the end of the risk period, during which cases are collected.
- In this design EOR estimates the **risk odds ratio**

$$\psi = \frac{\text{odds of dis. in the exp'd}}{\text{odds of dis. in the unexp'd}} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)},$$

  where π_1 and π_0 are the risks of disease in the exposed and unexposed groups, estimable from a corresponding cohort study by incidence proportions R_1 and R_0.
- **NB.** $\psi \approx \pi_1/\pi_0 = \phi$, *i.e.* close to **risk ratio**, when risks π_1 and π_0 are low = the "rare disease assumption".

Density sampling

- New incident cases occurring during given study period are identified from the base population.
- controls are randomly chosen from the population at risk at various times in the period (sometimes only once).
- For chronic disease studies this design is the most popular,
- Logically the only possibility in open populations,

Nested case-control study: *time-matched* selection:

- one or more (rarely over 5) controls chosen from the population at risk at each time t_d when a new case is diagnosed.

EOR in density sampling

- In a full cohort study the true hazard ratio $\rho = \lambda_1/\lambda_0$ is estimated by the incidence rate ratio

$$\text{IR} = \frac{I_1}{I_0} = \frac{D_1/D_0}{Y_1/Y_0}.$$

- In a case-control study with density sampling the **exposure odds** among controls C_1/C_0 estimates the exposure odds Y_1/Y_0, i.e. the distribution of person-years in the base population.
- Thus, the exposure odds ratio EOR

$$\text{EOR} = \frac{D_1/D_0}{C_1/C_0} \approx \frac{D_1/D_0}{Y_1/Y_0} = \text{IR}$$

  is a consistent estimator of the true hazard rate ratio ρ without any rare-disease assumption.

Example. Alcohol use and oesophageal cancer

(Tuyns et al 1977, see **B&D** 1980).

- 205 new cases of cancer identified in a French province during two years, and 770 randomly sampled population controls ⇒ Density sampling
- **NB.** No stratification or matching for age in design ⇒ Too many young controls in relation to few cases ⇒ inefficient!
- Exposure of interest: Daily consumption of alcohol.
- In the following table the data are summarized by dichotomized exposure and stratified by age group.
- In R the data are found: `data(esoph)`

Example: Results stratified by age

Age	Exposure $\geq$ 80 g/d	Cases	Ctrls	EOR
25-34	yes	1	9	∞
	no	0	106	
35-44	yes	4	26	5.05
	no	5	164	
45-54	yes	25	29	5.67
	no	21	138	
55-64	yes	42	27	6.36
	no	34	139	
65-74	yes	19	18	2.58
	no	36	88	
75-84	yes	5	0	∞
	no	8	31	
Total	yes	96	109	5.64
	no	104	666	(crude)

Example: (cont'd)

Modification?

- Stratum-specific EOR_ks somewhat variable.
- Random error in some of them apparently great (especially in the youngest and the oldest age groups).

Confounding?

- Is exposure associated with age in the study population?
- Look at variation in the age-specific prevalences of exposure among controls.
- Adjustment for age is generally reasonable.

Model for stratified data

Random part:

Conditional on total number of subjects

$$T_{jk} = D_{jk} + C_{jk}$$

in each level j ($j = 1, 2$) of exposure variable X and level k ($k = 1, \ldots K$) of covariate Z we assume

$$D_{jk} \sim \text{Binomial}(T_{jk};\ p_{jk}),$$

where p_{jk} is the "probability of being a case" in a group of cases & controls defined by X and Z.

Model for stratified data (cont'd)

Systematic part & logit link:

$$\text{logit}(p_{jk}) = \log\left(\frac{p_{jk}}{1 - p_{jk}}\right) = \alpha + \beta X + \gamma_k,$$

X = exposure, 1: 'exposed'; 0: 'unexposed',
$\alpha = \text{logit}(p_{11})$ = log of "pseudo baseline odds",
β = logarithm of exposure odds ratio,
$= \log(\rho)$, logarithm of true rate ratio ρ with density sampling,
γ_k = logarithm of rate ratio btw levels k and 1 of Z.

Hence $e^{\beta} = \rho$ is the common rate ratio for the exposure effect assumed constant over the levels of Z.

Example. Estimation by glm()

Input of data

```
D <- c(0,1, 5,4, 21,25, 34,42, 36,19, 8,5) # no. of cases
C <- c(106,9, 164,26, 138,29, 139,27, 88,18, 31,0) # controls
T <- D + C          # cell totals
```

Generation and naming of the levels for factors describing age group and alcohol exposure

```
agr <- gl(6,2,12)  # 6 levels for age factor
levels(agr) <- c("25-34", "35-44", "45-54",
                 "55-64", "65-74", "75-84")
alc <- gl(2,1,12)  # 2 levels for alcohol factor
levels(alc) <- c("0-79g/d", "80+g/d")
```


Example. Estimation by glm()

```
> data.frame( agrn = as.numeric(agr), agr,
              alcn = as.numeric(alc), alc, D, C, T)
   agrn   agr alcn      alc  D   C   T
1     1 25-34    1 0-79g/d  0 106 106
2     1 25-34    2  80+g/d  1   9  10
3     2 35-44    1 0-79g/d  5 164 169
4     2 35-44    2  80+g/d  4  26  30
5     3 45-54    1 0-79g/d 21 138 159
6     3 45-54    2  80+g/d 25  29  54
7     4 55-64    1 0-79g/d 34 139 173
8     4 55-64    2  80+g/d 42  27  69
9     5 65-74    1 0-79g/d 36  88 124
10    5 65-74    2  80+g/d 19  18  37
11    6 75-84    1 0-79g/d  8  31  39
12    6 75-84    2  80+g/d  5   0   5
```


Example. Estimation by glm() cont'd)

Crude estimation

```
> mod1 <- glm( D/T ~ alc, fam = binomial, w = T)
> round(ci.lin(mod1, Exp=T)[ , -(3:4)], 4)

            Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)  -1.8569 0.1054    0.1562 0.1270 0.1920
alc80+g/d     1.7299 0.1752    5.6401 4.0006 7.9515
```

Estimation adjusted for age

```
> mod2 <- update(mod1, . ~ . + agr)
```

Goodness-of-fit evaluation

```
> c( mod2$deviance, mod2$df.res )
[1] 11.04118    5
```


Example. Estimation by glm() (cont'd)

Estimation results after adjusting for age

```
> round(ci.lin(mod2, Exp=T)[ , -(3:4)], 4)

            Estimate StdErr exp(Est.)   2.5%    97.5%
(Intercept)  -5.0543 1.0094    0.0064 0.0009   0.0461
alc80+g/d     1.6699 0.1896    5.3116 3.6630   7.7022
agr35-44      1.5423 1.0659    4.6753 0.5788  37.7668
agr45-54      3.1988 1.0232   24.5022 3.2984 182.0171
agr55-64      3.7135 1.0185   40.9966 5.5688 301.8094
agr65-74      3.9669 1.0231   52.8196 7.1112 392.3239
agr75-84      3.9622 1.0650   52.5723 6.5193 423.9520
```

Matched case-control study

Matching

- For each case choose 1 or more (rarely over 4) controls with same age (*eg.* within 1 year, or in the same 5-year ageband), sex, place of living, *etc.*
- Implies stratification in design: each matched case-control set forms one stratum.
- Improves efficiency of the study & estimation of effect parameters, if matching factors are strong determinants of outcome.

Matched case-control study (cont'd)

Some principles

- Impractical to match on many other covariates than those mentioned,
- Matching on a correlate Z of risk factor X of interest, which is not causal determinant of outcome
  $\Rightarrow$ overmatching, loss of efficiency.
- *Counter-matching*: Choose controls which are different from case w.r.t. Z, close correlate of X
  $\Rightarrow$ increases efficiency.

Matched case-control study (cont'd)

- *Matched design* $\Rightarrow$ *matched analysis!*
- Ignoring matching in analysis my lead to biased results.
- Matching factors must always be accounted for in estimating the rate ratios of interest.
- With very close matching (based `e.g.` on sibship, neighbourhood) use *conditional logistic regression* modelling
  – function `clogit()` in package `survival`

Concluding remarks

- Analysis using `glm()` on individual data records from an unmatched study proceeds similarly as for grouped data.
- Matched design $\rightarrow$ matched analysis by `clogit()`.
- More complicated designs, like counter-matched and two-phase studies, require specialized methods and progamming.
- Case-cohort design: Use function `coxph()` in package `survival` but adjust standard errors *etc.* appropriately.