Epidemiologic data analysis using R

Practicals 3

Faculty of Social Sciences, University of Tampere

–

Janne Pitkäniemi

17.02.2020

Topics of practical 3

Learning objectives of this practical

- tabulation of cases, person-years and rates from individual data
- crude and adjusted rate ratio estimation by Poisson regression

Diet and heart data

Description

The diet data frame has 337 rows and 14 columns. The data concern a subsample of subjects drawn from larger cohort studies of the incidence of coronary heart disease (CHD). These subjects had all completed a 7-day weighed dietary survey while taking part in validation studies of dietary questionnaire methods. Upon the closure of the MRC Social Medicine Unit, from where these studies were directed, it was found that 46 CHD events had occurred in this group, thus allowing a serendipitous study of the relationship between diet and the incidence of CHD.

Format

This data frame contains the following columns:

```
x<-data.frame(text="
id:  subject identifier, a numeric vector.
doe:     date of entry into follow-up study, a Date variable.
dox:     date of exit from the follow-up study, a Date variable.
dob:     date of birth, a Date variable.
y:   - number of years at risk, a numeric vector.
fail:    status on exit, a numeric vector (codes 1, 3, 11, and 13 represent CHD events)
job:     occupation, a factor with levels Driver Conductor Bank worker
month:   month of dietary survey, a numeric vector
energy:  total energy intake (KCal per day/100), a numeric vector
height:  (cm), a numeric vector
weight:  (kg), a numeric vector
fat:     fat intake (g/day), a numeric vector
fibre:   dietary fibre intake (g/day), a numeric vector
energy.grp:  high daily energy intake, a factor with levels <=2750 KCal >2750 KCal
chd:     CHD event, a numeric vector (1=CHD event, 0=no event",sep=":",header=F)
```

Source

The data are described and used extensively by Clayton and Hills, Statistical Models in Epidemiology, Oxford University Press, Oxford:1993. They were rescued from destruction by David Clayton and reentered from paper printouts.

The data concern a subsample of subjects drawn from larger cohort studies of the incidence of coronary heart disease (CHD). These subjects had all completed a 7-day weighed dietary survey while taking part in validation studies of dietary questionnaire methods. Upon the closure of the MRC Social Medicine Unit, from where these studies were directed, it was found that 46 CHD events had occurred in this group, thus allowing a serendipitous study of the relationship between diet and the incidence of CHD.

1. Exploring the data.

(a) Load the diet to R data frame, see what's in there, and attach:

```
library(Epi)
data( diet )
attach(diet)
```

(b) The outcome event of interest is the first occurrence of coronary heart disease, this variable being named chd and coded 1 and 0. The person-time variable is y from the individual date of entry until the date of exit from the follow-up. NB! The numbers of cases and person-years, which are based on the data set we use in these exercises, will have somewhat different values from those given in Clayton and Hills (1993). Don't get confused of this discrepancy.

2. Computation and tabulation.

(a) Compute the observed number of cases D, total-person time Y , and the overall CHD incidence rate (per 1000 years) in this cohort. Display the values of these simultaneously rounding into 1 decimal:

```
D <- sum( chd ) ; #sum of the CHD events
Y <- sum( y ) ;  # sum of the person time at risk of CHD
rate <- 1000*D/Y # rate per 1,000 pyrs
round( c(D, Y, rate), 1)
```

(b) Function stat.table() can be used to tabulate events, person-years and rates by levels of a third variable, for example job. Try:

```
stat.table( job,
            list( sum(chd), sum(y), ratio(chd,y,1000) ),              margin=T )
```

which will produce the rates for each job category as well as the overall rate. The third argument to the ratio function changes the rates into being expressed as cases per 1000 years.

(c) Create a variable htgrp into which height is grouped in suitable intervals.

```
diet$htgrp <- cut( height,
                   breaks = c(150,170,175,195),
                   right=FALSE )
```

(d) See how height is associated with the incidence of CHD:

```
stat.table( htgrp,
    list( count(), sum(chd), sum(y), ratio(chd,y,1000) ),
    margin=T,
    data=diet )
```

(e) Find out how the rates vary between the two levels of energy.grp. Tabulate cases, personyears, and rates by energy.grp but save the table into an object with nice annotations:

```
tab.e <- stat.table( index = energy.grp,
contents = list( "Cases" = sum(chd),
                 "P-years" = sum(y),
                 "Rate/1000y" = ratio(chd, y, 1000) ),
          data = diet )
#Print the table by selective numerical precision:
print( tab.e, digits=c(sum=0, ratio=2));
```

(f) Look at the structure of the table object: str(tab.e); It is a two-dimensional table, the first index referring to the column and the 2nd index to the row of the printed table. (NB! The internal presentation of this table obeys the common way of indexing two-dimensional tables and matrices.)

(g) Now you can extract cases, person-years and rates into own vectors

```
D <- tab.e[1, ] ;
Y <- tab.e[2, ] ;
Y
rate <- tab.e[3, ] ; rate
```

Check what the individual rates and their ratio are

```
cat("rate for ", levels(diet$energy.grp)[1]," is ",rate[1],"\n");
cat("rate for ", levels(diet$energy.grp)[2]," is ",rate[2],"\n");
cat("rate ratio =",rate[1]/rate[2],"\n");
```

3. Model-based estimation of rate ratio.

(a) Use the command glm() with option family=poisson to find the crude rate ratio for the high energy group compared to the low energy group and see some results of this run:

$$log(\lambda_i) = offset_i + \alpha + \beta \times energy.grp_i$$

```
m0 <- glm( chd ~ energy.grp,
           family=poisson,
           offset=log(y),
           data = diet)
round( ci.lin( m0, Exp=T )[ , -(3:4)], 4)
```

Notice that this model is fitted on the individual records of the data frame. Now think about the interpretation of the estimated parameters; what do they mean?

(b) Change the reference level of the energy.grp factor so that high energy consumption is the reference category using function Relevel() in the Epi package. Check the levels of the factor to see that the 1st level indeed is the reference:

```
diet$eg2 <- Relevel( diet$energy.grp, ref = ">2750 KCals" )
levels(diet$eg2)
```

Refit the model with the new factor

```
m1 <- glm( chd ~ eg2 ,
           family=poisson,
           offset=log(y),
           data=diet )
round( ci.lin( m1, Exp=T )[ , -(3:4)], 4)
```

We skip the summary() but save the estimates etc. into an object

```
m1ci <- ci.lin( m1, Exp = T) # save results to r object
dim(m1ci) # 2 rows and 7 columns
m1ci
round( m1ci[ , -(3:4)] , 3 ) # remove 3rd and 4th columns
```

The point estimate should be exactly the same as obtained by direct calculation on summary rates above.

(c) Try an alternative way of fitting the Poisson model with log link and person-years as weights:

```
m1b <- glm( chd/y ~ eg2 ,
            family=poisson(link="log"),
            w = y,
            data=diet )
round( ci.lin(m1b, Exp=T)[ , -(3:4)], 3 )
```

4. Height and CHD

As height appears to be a strong predictor of CHD incidence, we shall take a look, how the rates vary by both factors.

(a) Tabulate cases, person-years, and rates by energy group and height, print and look at the structure of the table:

```
tab.eh <- stat.table(
  index = list(energy.grp, htgrp),
  contents = list( "Cases" = sum(chd),
                   "P-years" = sum(y),
                  "Rate/1000y" = ratio(chd, y, 1000) ),
data = diet )
print( tab.eh, digits=c(sum=0, ratio=2) )
dim(tab.eh)
```

(b) The first dimension in this array appears to refer to the three different quantities, and the rates are found as the 3rd item in this dimension, so let's print just them:

```
round( tab.eh[3, , ], 1)
```

(c) We may compute the rate ratios between the energy groups in all height strata:

```
IRe.h <- tab.eh[3, 1, ]/ tab.eh[3, 2, ] ; round(IRe.h, 3)
```

5. energy group and CHD controlling htgrp

Estimating the effect of energy group controlling for height (a) Fit a model that includes the main effects of eg2 and htgrp:

```
meh <- glm( chd/y ~ eg2 + htgrp, family=poisson, w = y, data=diet )
round(ci.lin(meh, Exp=T)[ , -(3:4)], 4)
```

Compare the estimate for eg2 to the crude estimate. Any change? (b) Evaluate the possible modification of the energy effect by height by updating the previous model formula to include the relevant interaction term:

```
meh2 <- update(meh, . ~ . + eg2:htgrp )
round(ci.lin(meh2, Exp=T)[ , -(3:4)], 4)
```

Perform comparison of deviances between the main effects model and the interaction model:

```
anova(meh, meh2)
```

The evidence for any interaction appears very weak. After viewing the structure of the anova object, a formal P-value is obtained for the interaction:

```
str(anova(meh, meh2))
pchisq( anova(meh, meh2)$Dev[2], anova(meh, meh2)$Df[2], lower.tail =F )
```

6. Height continuous and CHD

Height is actually a quantitative covariate, so when performing a crude categorisation we may lose essential information. As we have individual data on heights, we may treat it as a quantitative covariate in modelling.

(a) Fit a main effects model in which the categorized height is substituted by the linear term of the original quantitative height variable.

```
meh3 <- glm( chd ~ eg2 + height,
             fam=poisson,
             offset = log(y),
             data = diet)
round(ci.lin(meh3, Exp=T)[ , -(3:4)], 4)
```

(b) The coefficient & rate ratio for height appears modest when compared to those otained from the categorical model. For the purposes of more concrete interpretation of parameter estimates, it is almost always useful to perform some centering and scaling to quantitative variables. Here we choose 175 cm as the centering point and 5 cm to define the scale:

```
diet$hei.lin <- (diet$height - 175)/5
meh3b <- glm( chd ~ eg2 + hei.lin,
              fam=poisson,
              offset = log(y),
              data = diet)

round(ci.lin(meh3b, Exp=T)[ , -(3:4)], 4)
```

(c) Based on inspecting the tabulated rates across the height categories one might see a slight systematic deviation from linearity in the effect of height. We shall hence add the uadratic term upon the linear one to evaluate the "significance" of this deviation:

```
diet$hei.quad <- diet$height^2
meh4 <- update(meh3b, . ~ . + hei.quad)
round(ci.lin(meh4, Exp=T)[ , -(3:4)], 4)
anova(meh3b, meh4)
```