Epidemiologic data analysis using R

Practicals 6

Faculty of Social Sciences, University of Tampere

–

Janne Pitkaniemi

26.03.2018

# Topics of practical 6

Learning objectives of this practical

* reading ascii data files using R-studio
* analysis of case-control data

Data description

In the mid-80s a case-control study on risk factors for malignant melanoma was conducted in Denmark (Asterlind et al. The Danish case-control study of cutaneous malignant melanoma I: Importance of host factors. Int J Cancer 1988; 42: 200-206). The cases were patients with skin melanoma (excluding lentigo melanoma), newly diagnosed from 1 Oct, 1982 to 31 March, 1985, aged 20-79, from East Denmark, and they were identified from the Danish Cancer Registry. The controls (twice as many as cases) were drawn from the residents of East Denmark in April, 1984, as a random sample stratified by sex and age (within the same 5 year age group) to reflect the sex and age distribution of the cases. This is called group matching, and in such a study, it is necessary to control for age and sex in the statistical analysis. (Yes indeed: In spite of the fact that stratified sampling by sex and age removed the statistical association of these variables with melanoma from the final case-control data set, the analysis must control for variables which determine the probability of selecting subjects from the base population to the study sample.)

The population of East Denmark is a dynamic one. Sampling the controls only at one time point is a rough approximation of indidence density sampling, which ideally would spread out over the whole study period. Hence the exposure odds ratios calculable from the data are estimates of the corresponding hazard rate ratios between the exposure groups. After exclusions, refusals etc., 474 cases (92% of eligible cases) and 926 controls (82%) were interviewed. This was done face-to-face with a structured questionnaire by trained interviewers, who were not informed about the subjects case-control status. For this exercise we have selected a few host variables from the study in an ascii-file, melanoma.dat.

The variables are listed in the following table. Table 1: Variables in the melanoma dataset.

Variable Units.or.Coding Type Name  
 Case-control status 1=case, 0=control numeric cc  
 Sex 1=male, 2=female numeric sex   
 Age at interview age in years numeric age  
 Skin complexion 0=dark, 1=medium, 2=light numeric skin  
 Hair colour 0=dark brown/black, numeric hair  
 1=light brown,2=blonde, 3=red   
 eye colour 0=brown, 1=grey, green, 2=blue numeric eyes  
 Freckles 1=many, 2=some, 3=none numeric freckles  
 Naevi small no. naevi < 5mm numeric nvsmall  
 Naevi largs no. naevi >= 5mm numeric nvlarge

## 1. Reading the data

Use R-studio import data to read in ascii(text) format data.

**Note - missing value is defined by . (period)**.

library(Epi)

Warning: package 'Epi' was built under R version 3.4.4

mm <- read.table("C:/Users/janne.pitkaniemi/Projects/TRE2018/melanoma.dat", header=T, na.strings=".")  
str(mm)

'data.frame': 1400 obs. of 9 variables:  
 $ cc : int 1 1 1 0 1 0 0 0 0 1 ...  
 $ sex : int 2 1 2 2 2 2 2 1 2 2 ...  
 $ age : int 71 68 42 66 36 68 68 39 75 49 ...  
 $ skin : int 2 2 1 0 1 2 0 2 2 2 ...  
 $ hair : int 0 0 1 2 0 2 0 0 0 1 ...  
 $ eyes : int 2 2 2 1 2 2 1 2 2 2 ...  
 $ freckles: int 2 1 3 2 3 2 2 2 1 2 ...  
 $ nvsmall : int 2 3 22 0 1 0 0 3 5 6 ...  
 $ nvlarge : int 0 0 1 0 0 0 0 0 0 0 ...

head(mm, n=20)

cc sex age skin hair eyes freckles nvsmall nvlarge  
1 1 2 71 2 0 2 2 2 0  
2 1 1 68 2 0 2 1 3 0  
3 1 2 42 1 1 2 3 22 1  
4 0 2 66 0 2 1 2 0 0  
5 1 2 36 1 0 2 3 1 0  
6 0 2 68 2 2 2 2 0 0  
7 0 2 68 0 0 1 2 0 0  
8 0 1 39 2 0 2 2 3 0  
9 0 2 75 2 0 2 1 5 0  
10 1 2 49 2 1 2 2 6 0  
11 0 1 48 2 1 2 3 4 0  
12 1 2 67 0 0 2 2 1 0  
13 0 1 50 1 0 2 3 4 0  
14 1 2 38 2 0 1 3 8 0  
15 0 2 33 2 1 2 2 3 0  
16 0 2 39 1 0 1 3 0 2  
17 0 2 39 1 1 2 3 0 0  
18 1 1 50 0 1 1 1 3 1  
19 0 2 35 2 0 2 2 1 0  
20 0 2 35 2 0 1 3 5 0

## 2. House keeping

The structure of the data frame mm tells us that all the variables are numeric (integer), so first you need to do a bit of house keeping. For example the variables sex, skin, hair, eye need to be converted to factors, with labels, and freckles which is coded 4 for none down to 1 for many (not very intuitive) needs to be recoded, and relabelled.

To avoid too much typing and to leave more time to think about the analysis, just copy paste tthese the following lines to your R-script

mm$sex <- factor(mm$sex,labels=c("M","F"))  
mm$skin <- factor(mm$skin,labels=c("dark","medium","light"))  
mm$hair <- factor(mm$hair,labels=c("dark","light\_brown","blonde","red"))  
mm$eyes <- factor(mm$eyes,labels=c("brown","grey-green","blue"))  
mm$freckles <- 4 - mm$freckles  
mm$age.cat <- cut(mm$age,breaks=c(20,30,40,50,60,70,85),right=F)  
mm$freckles <- factor(mm$freckles,labels=c("none","some","many"))  
  
mm$hair2 <- Relevel(mm$hair,list("dark"=1,"other"=c(2,3,4)))  
mm$nvsma4 <- cut(mm$nvsmall,breaks=c(0,1,2,5,50),right=F)  
mm$nvlar3 <- cut(mm$nvlarge,breaks=c(0,1,2,15),right=F)

or alternatively you can use the ready made file melanoma-house.r by adding the following line to your r script source(“melanoma-house.r”)

Look again at the structure of the data frame mm and note the changes. Use the command summary(mm) to look at the univariate distributions.

str(mm) ; summary(mm)

Now let’s turn to something a bit more interesting.

## 3. Association of melanoma with one variable at a time

As a first step it is a good idea to start by looking at the effect of each of the variables, controlled for age and sex. To examine the effect of hair colour, start from simple cross-tabulation:

stat.table( list( cc, hair), list(count(), percent(hair) ), mm, T)

Now estimate the effect of hair by fitting the corresponding binomial regression model with logit link:

m.hair <- glm(cc ~ sex + age.cat + hair, family="binomial", data=mm)  
round(ci.lin(m.hair, Exp=T)[ , 5:7], 2)

Look at the effects of eyes and freckles in the same way.

## 4. Hair colour as a binary or dichotomous factor

From the above you can see that subjects at each of the 3 levels light-brown, blonde, and red, are at greater risk than subjects with dark hair, with similar odds ratios. This suggests creating a new variable hair2 which has just two levels, dark and the other three. The Relevel() function has been used for this in the house keeping script.

Use glm() to compute the odds-ratio of melanoma between persons with red, blonde or light brown hair versus those with dark hair.

## 5. The effect of freckles stratified by and adjusted for hair2

When you control the effect of an exposure for some variable you are asking a question about what would the effect be if the variable is kept constant. For example, consider the effect of freckles adjusted for hair2 in addition to age.cat and sex. We first estimate the effect of freckles separately in the two categories of hair2. This can be done by fitting a generalized model, in which the effect of freckles is stratified by hair2 (or nested in) as follows:

m.f\_h2 <- glm( cc ~ sex + age.cat + hair2/freckles,  
family = "binomial", data = mm)  
round( ci.lin( m.f\_h2, Exp=T)[ , 5:7], 2)

Based on eyeballing the hair-colour-specific effect estimates for the two categories of freckles, and their wide and substantially overlapping confidence intervals, there seems to be not much evidence against the simplifying assumption that the rate ratios associated with freckles would be homogenous in the two levels of hair colour. Thus, we will next fit a model in which the effect of freckles is adjusted for hair2 in the usual way. It is also possible to perform a deviance test for the possible modification using anova().

m.fh2 <- glm( cc ~ sex + age.cat + hair2 + freckles,  
family = "binomial", data = mm)  
round( ci.lin( m.fh2, Exp=T)[ , 5:7], 2)  
anova( m.fh2, m.f\_h2)

## 6. Naevi small and large

The distributions of nvsmall and nvlarge are very skew to the right. You can see this from the stem&leaf diagrams

with(mm, stem(nvsmall)) ; with(mm, stem(nvlarge))

Because of this it is wise to categorize them into a few classes \* small naevi into four: 0, 1, 2-4, and 5+; \* large naevi into three: 0, 1, and 2+. This has already been done in the housekeeping script. Look at the joint frequency distribution of these new variables using stat.table(). Are they strongly associated?

stat.table(index = list(nvsma4,nvlar3),  
contents = list(count(), percent(nvlar3)),  
margins = T, data=mm)

Compute the sex- and age-adjusted OR estimates associated with the number of small naevi by fitting separate logistic regression models including sex, age.cat and nvsma4 in the model formula.

m.nvs <- glm(cc ~ sex + age.cat + nvsma4, family="binomial", data=mm)  
round(ci.lin(m.nvs, Exp=T)[, 5:7], 3)

Do the same with nvlar3.

m.nvl <- glm(cc ~ sex + age.cat + nvlar3, family="binomial", data=mm)  
round(ci.lin(m.nvl, Exp=T)[ , 5:7], 3)

Now fit a glm containing age.cat, sex, nvsma4 and nvlar3 and assign the result in model object nvboth. What is the interpretation of the last two coefficients?

nvboth <- glm(cc ~ sex + age.cat + nvlar3 + nvsma4,  
family="binomial", data=mm)  
round( ci.lin(nvboth, Exp=T)[ -(1:7), 5:7], 2)

The odds ratios (with CIs) can be graphically displayed using function plotEst() in Epi. Its uses the value of ci.lin() evaluated on the fitted model object. As the intercept and the effects of age and sex are of no interest, we shall drop the corresponding rows (the 7 first ones) from the matrix produced by ci.lin(), and the plot is based just on the 1st, 5th and the 6th column of this matrix:

plotEst( exp( ci.lin( nvboth )[-(1:7), -(2:4)] ), xlog=T,vref=1 )

The xlog argument makes the OR axis logarithmic.

## 7. Nested case-control example

Recall the diet data

library(Epi)  
library(survival)  
data(diet)  
head(diet)

Recall 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.

We are interested in association between energy intake and CHD risk. For that we create a nested case-control study: Given the basic outcome variables for a cohort study: the time of entry to the cohort, the time of exit and the reason for exit (“failure” or “censoring”), this function computes risk sets and generates a matched case-control study in which each case is compared with a set of controls randomly sampled from the appropriate risk set. Other variables may be matched when selecting controls.

## Set seed for the random number generator  
set.seed(20180326)  
## Generate a nested case-control study  
dietcc <- ccwc(  
 entry = doe, # Time of entry to follow-up  
 exit = dox, # Time of exit from follow-up  
 fail = chd, # Status on exit (1 = Fail,0 =Censored)  
 origin = dob, # Origin of analysis time scale  
 controls = 2, # The number of controls to be selected for each case  
 data = diet, # data frame  
 include = energy, # List of other variables to be carried across into the case-control study  
 match = job, # List of categorical variables on which to match cases and controls  
 silent = TRUE  
 )  
  
## Show first 10 observations  
head(dietcc, 10)

The 1:2 risk set-matched dataset contains the following variables.

Set: case-control set number Map: row number of record in input dataframe Time: failure time of the case in this set Fail: failure status (1=case, 0=control) These are followed by the matching variables, and finally by thevariables in the ‘include’ list

The analysis requires conditional logistic regression. By stratifying on the matched sets, the matching factors, i.e., risk set and the job type here, are conditioned on. Conditioning on the matching factors is necessary in matched case control studies. Otherwise, selection bias will arise.

resClogit <- clogit(formula = Fail ~ scale(energy) + strata(Set), data = dietcc)  
summary(resClogit)

What is the estimate of RR and is it significant?

summary(coxph(Surv(y,chd)~scale(energy)+job,data=diet))

What is the RR using the whole cohort ?

## 8. Case-cohort analysis

Study question: Is stage of the tumor associated with the risk of relapse in Wilm’s tumor.

Data from the National Wilm’s Study.

Missing data/masurement error example. Tumor histology predicts survival, but prediction is stronger with central lab histology than with the local institution determination.

Format: A data frame with 4028 observations on the following 9 variables.

'seqno' id number  
 'instit' Histology from local institution  
 'histol' Histology from central lab  
 'stage' Disease stage  
 'study' study  
 'rel' indicator for relapse  
 'edrel' time to relapse  
 'age' age in months  
 'in.subcohort' Included in the subcohort for the example in the paper'

The in.subcohort variable indicates if the subject has been chosen as a member of the subcohort at the beginning of the study. To simulate a case-cohort study, dataset is subset to those who relapsed (cases) and those who are in the subcohort (controls).

## Load data  
data(nwtco)  
head(nwtco)

## Indicator for those relapsed OR in the subcohort (data ascertained for cases and controls only)  
selccoh <- with(nwtco, rel == 1 | in.subcohort == TRUE)  
  
## Subset to these 1154 patients. (Case-cohort dataset)  
caseCohortData <- nwtco[selccoh,]  
  
## Create factors  
caseCohortData <- within(caseCohortData, {  
  
 histol <- factor(histol, labels = c("FH","UH"))  
 stage <- factor(stage, labels = c("I","II","III","IV"))  
 age <- age / 12 # Age in years  
 })  
  
## Check  
head(caseCohortData)

Analysis with the cch() function. Cox regression is used to fit the model, thus the HR notation. The HR shown here is interpreted as the risk ratio. Fits proportional hazards regression model to case-cohort data. Returns estimates and standard errors from relative risk regression fit to data from case-cohort studies. A choice is available among the Prentice, Self-Prentice and Lin-Ying methods for unstratified data. For stratified data the choice is between Borgan I, a generalization of the Self-Prentice estimator fo unstratified case-cohort data, and Borgan II, a generalization of the Lin-Ying estimator.

Standard case-cohort analysis: simple random subcohort Fits proportional hazards regression model to case-cohort data

fit.ccP <- cch(Surv(edrel, rel) ~ stage + histol + age,  
 data = caseCohortData,  
 subcoh = ~ in.subcohort, # Vector of indicatorsfor subjects sampled as part of the sub-cohort  
 id = ~ seqno, # Vector of unique identifiers  
 cohort.size = 4028) # Vector with size of each stratum original cohort  
summary(fit.ccP)