Course work I:
Analysis of the diet study

EHA September 1 - 25, 2015

A cohort of 337 men in three occupational groups in England, aged 30 to 67 years at entry, recruited in 1950s and 1960s, followed-up until mid 1970s for incidence of CHD events. The purpose of the diet study was to investigate the incidence of CHD events in relation to age, energy intake, fibre intake and possibly body mass index.

1. Read in the external data file diet.dat into a data frame diet using read.table. The data set contains

   dob = date of birth,
   doe = date of entry into follow-up,
   dox = date of exit, end of follow-up,
   chd = indicator for status at exit: 1 = CHD event occurred, 0 = censored.

   The risk factors of interest, measured by dietary survey at entry are

   energy = total energy intake (kcal/d),
   energy.grp = energy dichotomized: 1 $=< 2750 kcal/d$, 2 $=\geq 2750 kcal/d$,
   fat = fat intake (g/d),
   fibre = dietary fibre intake (g/d).

2. Obtain summary statistics of each variable.

3. The date variables dob, doe and dox are expressed as *days since 1 January 1970*. Create variable age.entry for age at entry.

4. Using the function Lexis and splitLexis create a new data frame in which the follow-up times are split by age using agebands 30-49, 50-59 and 60-69 years, respectively. Include the risk factors and covariates in the expanded data frame. View the first 10 rows of it and compare them with the first lines of the original data frame. How many rows

there are overall in these two data frames? Use dim to detect the size of a data frame. (Note: You will need the library(Epi)).

5. Create energy.grp into a factor energy.g2 and give labels to its levels. Create a four-level factor fib.g4 of the variable fibre (expressed in units of 10 g/d) using its suitably rounded quartiles as the cutpoints.

6. Calculate individual person-time slots (y_{ik} in the lecture notes) spent in each ageband. Tabulate the numbers of CHD cases, person-years and incidence rates by factors energy.g2, fib.g4, and ageband. What do the data suggest about the association between energy intake and CHD and between fibre intake and CHD?

7. Plot the incidence rates by ageband separately for the two levels of energy.g2.

8. Summing over the other dimensions calculate the overall numbers of cases, person-years and crude incidence rates of CHD in the two energy.g2 levels as well as the crude rate ratio: high vs. low.

9. Calculate also an approximate 95% confidence interval for this rate ratio (Recall that $SE[\log(rate - -ratio)] = \sqrt{1/D1 + 1/D0}$.)

10. From the tables above form a grouped data frame with 24 rows showing the numbers of cases and person-years for each combination of ageband, fib.g4 level and energy.g2 level. Have a look at the data frame.

11. Fit a Poisson regression model for the numbers of cases as the outcome with log-link and with energy.g2 as the only regressor. What is the point estimate and the 95% confidence interval of the rate ratio? Compare the results with the crude rate ratio above. What would you think of the goodness of fit of this simple model?

References

Clayton D and Hills M. *Statistical Methods in Epidemiology*, Oxford University Press, New York (2002).

Plummer, M. and Carstensen, B. (2011). Lexis: An R class for Epidemiolgical studies with long-term follow-up. Journal of Statistical Software, 38 (5); 1:12.