Epidemiologic data analysis using R

Practicals 4

Faculty of Social Sciences, University of Tampere

–

Janne Pitkäniemi / Karri Seppä

2.3.2020

# Topics of practical 4

Learning objectives of times into agebands and expanding the data frame using functions Lexis(), splitLexis() and timeBand() in the Epi package,

* model-based adjustment for current age and evaluation of modification in rate ratio estimation,
* graphical display of age-specific rates using plot() function, and estimation results from models using plotEst() in Epi.

## 1. Diet data preliminary

We shall continue analysing the diet data. Hence, we first repeat some preliminary tasks. (a) Load the diet data frame and show 10 first rows:

library(Epi)  
data( diet )  
diet[1:10, ]

id doe dox dob y fail job month  
1 102 1976-01-17 1986-12-02 1939-03-02 10.8747433 0 Driver 1  
2 59 1973-07-16 1982-07-05 1912-07-05 8.9691992 0 Driver 7  
3 126 1970-03-17 1984-03-20 1919-12-24 14.0095825 13 Conductor 3  
4 16 1969-05-16 1969-12-31 1906-09-17 0.6269678 3 Driver 5  
5 247 1968-03-16 1979-06-25 1918-07-10 11.2744695 13 Bank worker 3  
6 272 1969-03-16 1973-12-13 1920-03-06 4.7446954 3 Bank worker 3  
7 268 1969-02-16 1986-12-02 1919-06-24 17.7905544 0 Bank worker 2  
8 206 1967-01-17 1986-12-02 1917-09-26 19.8740589 0 Bank worker 1  
9 182 1971-03-17 1976-07-27 1925-01-31 5.3634497 13 Conductor 3  
10 2 1974-12-17 1986-12-02 1924-06-03 11.9589322 0 Driver 12  
 energy height weight fat fibre energy.grp chd  
1 22.8601 181.6100 88.17984 9.168 1.4000000 <=2750 KCals 0  
2 23.8841 165.9890 58.74120 9.651 0.9350001 <=2750 KCals 0  
3 24.9537 152.4000 49.89600 11.249 1.2480000 <=2750 KCals 1  
4 22.2383 171.1960 89.40456 7.578 1.5570000 <=2750 KCals 1  
5 18.5402 177.8000 97.07040 9.147 0.9910000 <=2750 KCals 1  
6 20.3073 175.2600 61.00920 8.536 0.7650000 <=2750 KCals 1  
7 24.5261 179.0700 81.19440 11.307 1.3210000 <=2750 KCals 0  
8 23.6894 187.9600 95.25600 11.094 1.7890000 <=2750 KCals 0  
9 23.1603 173.9900 65.99880 10.140 1.4100000 <=2750 KCals 1  
10 19.8234 164.2872 70.08120 8.577 0.9490000 <=2750 KCals 0

1. 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:

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

1. Now convert the date variables corresponding to the dates of entry, exit, and birth, respectively, into fractional calendar years by function cal.yr() in Epi:

diet <- transform(diet, doe = cal.yr(doe),  
dox = cal.yr(dox), dob = cal.yr(dob) )

Check the result:

diet$doe[1:10] ; summary(diet$dob)

[1] 1976.042 1973.537 1970.205 1969.370 1968.204 1969.203 1969.127 1967.043  
 [9] 1971.205 1974.958

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1902 1916 1921 1921 1925 1941

## 2. Splitting age

Splitting into agebands and expanding the data frame

1. We first create a Lexis object from the original data frame with respect to age scale

dietL <- Lexis( entry = list(age = doe-dob),  
exit = list(age = dox-dob),  
exit.status = chd, data = diet)

NOTE: entry.status has been set to 0 for all.

round(dietL[ 1:10, 1:10],1)

age lex.dur lex.Cst lex.Xst lex.id id doe dox dob y  
1 36.9 10.9 0 0 1 102 1976.0 1986.9 1939.2 10.9  
2 61.0 9.0 0 0 2 59 1973.5 1982.5 1912.5 9.0  
3 50.2 14.0 0 1 3 126 1970.2 1984.2 1920.0 14.0  
4 62.7 0.6 0 1 4 16 1969.4 1970.0 1906.7 0.6  
5 49.7 11.3 0 1 5 247 1968.2 1979.5 1918.5 11.3  
6 49.0 4.7 0 1 6 272 1969.2 1973.9 1920.2 4.7  
7 49.7 17.8 0 0 7 268 1969.1 1986.9 1919.5 17.8  
8 49.3 19.9 0 0 8 206 1967.0 1986.9 1917.7 19.9  
9 46.1 5.4 0 1 9 182 1971.2 1976.6 1925.1 5.4  
10 50.5 12.0 0 0 10 2 1975.0 1986.9 1924.4 12.0

1. We now expand the Lexis object, such that each individual will have 1, 2, or 3 rows depending on how many agebands exist to which he is contributing follow-up time.

dietA <- splitLexis(dietL, br = c(30,50,60,70),  
time.scale = "age")

Let’s have a look at the result of this operation. Print the first few lines of the expanded data fame and compare with the original one:

round(dietA[1:10, 1:10],1)

lex.id age lex.dur lex.Cst lex.Xst id doe dox dob y  
1 1 36.9 10.9 0 0 102 1976.0 1986.9 1939.2 10.9  
2 2 61.0 9.0 0 0 59 1973.5 1982.5 1912.5 9.0  
3 3 50.2 9.8 0 0 126 1970.2 1984.2 1920.0 14.0  
4 3 60.0 4.2 0 1 126 1970.2 1984.2 1920.0 14.0  
5 4 62.7 0.6 0 1 16 1969.4 1970.0 1906.7 0.6  
6 5 49.7 0.3 0 0 247 1968.2 1979.5 1918.5 11.3  
7 5 50.0 10.0 0 0 247 1968.2 1979.5 1918.5 11.3  
8 5 60.0 1.0 0 1 247 1968.2 1979.5 1918.5 11.3  
9 6 49.0 1.0 0 0 272 1969.2 1973.9 1920.2 4.7  
10 6 50.0 3.8 0 1 272 1969.2 1973.9 1920.2 4.7

round(diet[1:10, 1:6],1)

id doe dox dob y fail  
1 102 1976.0 1986.9 1939.2 10.9 0  
2 59 1973.5 1982.5 1912.5 9.0 0  
3 126 1970.2 1984.2 1920.0 14.0 13  
4 16 1969.4 1970.0 1906.7 0.6 3  
5 247 1968.2 1979.5 1918.5 11.3 13  
6 272 1969.2 1973.9 1920.2 4.7 3  
7 268 1969.1 1986.9 1919.5 17.8 0  
8 206 1967.0 1986.9 1917.7 19.9 0  
9 182 1971.2 1976.6 1925.1 5.4 13  
10 2 1975.0 1986.9 1924.4 12.0 0

For some ids more than 1 row were created. The sizes (dimensions) of the original and the age-split data frame, respectively, can be compared by:

dim(diet) ; dim(dietA)

[1] 337 16

[1] 729 21

1. It is useful to convert the dieatA$age variable given by Lexis() to be a factor named ageband.

dietA$ageband <- timeBand(dietA, "age", "factor")

Apart from the intended agebands starting from 30 y and ending at 70 y, splitLexis() creates agebands for the tails of the age scale, too. The result can be checked:

table(dietA$ageband)

(-Inf,30] (30,50] (50,60] (60,70] (70,Inf]   
 0 196 293 240 0

The 1st and the 5th ageband do not contain any observations, so they can safely be merged with the neighbouring ones, 2nd and 4th, respectively, after which the remaining three agebands may be renamed:

dietA$ageband <- Relevel(dietA$ageband, list(1:2, 3, 4:5))  
levels(dietA$ageband) <- c("30-<50", "50-<60", "60-<70")  
table(dietA$ageband)

30-<50 50-<60 60-<70   
 196 293 240

1. Rename also the ageband-specific event indicator dietAd\_ik and the agebandspecific person-time slot dieaty\_ik, analogous to the notation used in the lecture notes:

dietA$y\_ik <- dietA$lex.dur  
dietA$d\_ik <- dietA$lex.Xst

1. A check of the overall no. of cases and p-years:

sum(dietA$y\_ik); sum(dietA$d\_ik)

[1] 4603.669

[1] 46

These should be the same as obtained in Practical 3.

## 3. Tabulate rates

Tabulation of cases, person-years and rates by ageband

1. At this stage we are able to produce a table in which the cases, person-years and rates are jointly classified by ageband and energy group:

tab.ae <- stat.table(index = list( Ageband = ageband, "Energy group" = eg2 ),  
contents = list( D = sum(d\_ik), Y = sum(y\_ik),  
I = ratio(d\_ik, y\_ik, 1000)),  
margin = T, data = dietA )  
print(tab.ae,digits = c(sum=0, ratio=2))

----------------------------------   
 ------Energy group-------   
 Ageband >2750 <=2750 Total   
 KCals KCals   
 ----------------------------------   
 30-<50 4 2 6   
 622 381 1003   
 6.43 5.25 5.98   
   
 50-<60 6 12 18   
 1128 979 2107   
 5.32 12.25 8.54   
   
 60-<70 8 14 22   
 794 699 1493   
 10.07 20.02 14.73   
   
   
 Total 18 28 46   
 2544 2059 4604   
 7.07 13.60 9.99   
 ----------------------------------

View the internal structure of this table

str(tab.ae)

'stat.table' num [1:3, 1:4, 1:3] 4 622.38 6.43 6 1127.7 ...  
 - attr(\*, "dimnames")=List of 3  
 ..$ contents : Named chr [1:3] "D" "Y" "I"  
 .. ..- attr(\*, "names")= chr [1:3] "D" "Y" "I"  
 ..$ Ageband : chr [1:4] "30-<50" "50-<60" "60-<70" "Total"  
 ..$ Energy group: chr [1:3] ">2750 KCals" "<=2750 KCals" "Total"  
 - attr(\*, "table.fun")= chr [1:3] "sum" "sum" "ratio"

It is a 3-dimensional array, in which the 1st index refers to the contents (cases, p-years, or rates), the 2nd index to ageband, and the 3rd index to energy group.

1. We can now extract the numbers of cases, person-years, and rates into separate two-way tables indexed by ageband and energy group:

cases <- tab.ae[1, , ]  
years <- tab.ae[2, , ]  
rates <- tab.ae[3, , ]

Next, we can view these tables and their common structure:

cases; round(years,1); round(rates,1) ; str(rates)

Energy group  
Ageband >2750 KCals <=2750 KCals Total  
 30-<50 4 2 6  
 50-<60 6 12 18  
 60-<70 8 14 22  
 Total 18 28 46

Energy group  
Ageband >2750 KCals <=2750 KCals Total  
 30-<50 622.4 381.0 1003.3  
 50-<60 1127.7 979.3 2107.0  
 60-<70 794.2 699.1 1493.3  
 Total 2544.2 2059.4 4603.7

Energy group  
Ageband >2750 KCals <=2750 KCals Total  
 30-<50 6.4 5.2 6.0  
 50-<60 5.3 12.3 8.5  
 60-<70 10.1 20.0 14.7  
 Total 7.1 13.6 10.0

num [1:4, 1:3] 6.43 5.32 10.07 7.07 5.25 ...  
 - attr(\*, "dimnames")=List of 2  
 ..$ Ageband : chr [1:4] "30-<50" "50-<60" "60-<70" "Total"  
 ..$ Energy group: chr [1:3] ">2750 KCals" "<=2750 KCals" "Total"

1. The ageband specific incidence rate ratios between the two energy groups are obtained by contrasting the second column with the first one in the rates table:

IRs <- rates[ , 2] / rates[ , 1] ; round(IRs, 2)

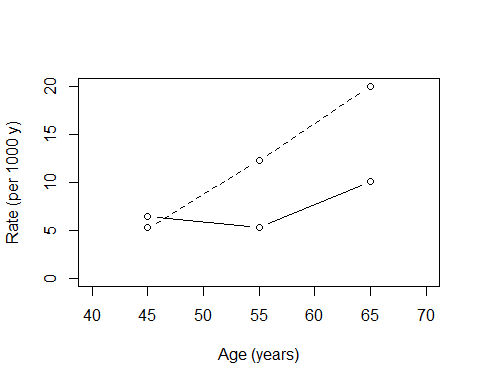
30-<50 50-<60 60-<70 Total   
 0.82 2.30 1.99 1.92

The last value on the right in the output is the crude rate ratio.

## 4. Graphical display of rates

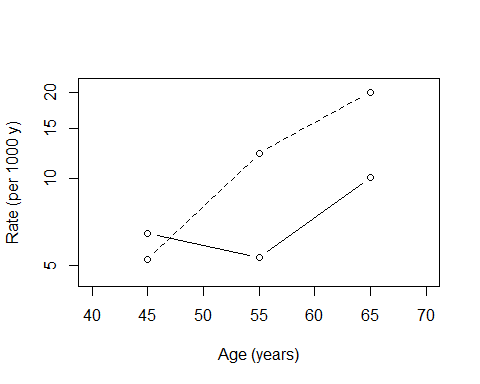
1. Plot the incidence rates in the two energy groups by ageband. From the rates table we can do this as follows (omitting the overall crude rates, i.e. value 4 in the 1st index):

agemids <- c( 45, 55, 65 ) # mid-ages of the agebands  
plot( rates[ -4, 1] ~ agemids, type="b", lty = 1, # the exposed  
xlim= c(40, 70), ylim = c( 0, 20 ) ,  
xlab = "Age (years)", ylab = "Rate (per 1000 y)" )  
lines( rates[ -4, 2] ~ agemids, type="b", lty=2 ) # the unexposed



1. Do the same in the logarithmic scale for the rates:

plot( rates[ -4, 1] ~ agemids, type="b", lty = 1,  
xlim= c(40, 70), ylim = c( 4.5, 21 ) , log = "y",  
xlab = "Age (years)", ylab = "Rate (per 1000 y)" )  
lines( rates[ -4, 2] ~ agemids, type="b", lty=2 )



These graphs are quite rough. They could be improved in many respects by means of the multitude of graphical functions and parameters available in R.

## 5. Poisson regression

Adjusting for current age by Poisson regression (a) Fit first a model for crude estimation of the effect of energy group. Are the results similar to those obtained yesterday on the original data frame?

me <- glm( d\_ik/y\_ik ~ eg2, fam = poisson, w = y\_ik, data = dietA )  
ci.lin(me, Exp=T)[ , -(3:4)]

Estimate StdErr exp(Est.) 2.5% 97.5%  
(Intercept) -4.9512148 0.2357021 0.007074809 0.004457432 0.01122909  
eg2<=2750 KCals 0.6532345 0.3021089 1.921746718 1.063018022 3.47417482

1. Fit the same model in the other way involving an offset term:

me1 <- glm( d\_ik ~ eg2, fam = poisson, offset=log(y\_ik), data = dietA )  
ci.lin(me1, Exp=T)[ , -(3:4)]

Estimate StdErr exp(Est.) 2.5% 97.5%  
(Intercept) -4.9512148 0.2356987 0.007074809 0.004457462 0.01122902  
eg2<=2750 KCals 0.6532345 0.3021023 1.921746721 1.063031585 3.47413051

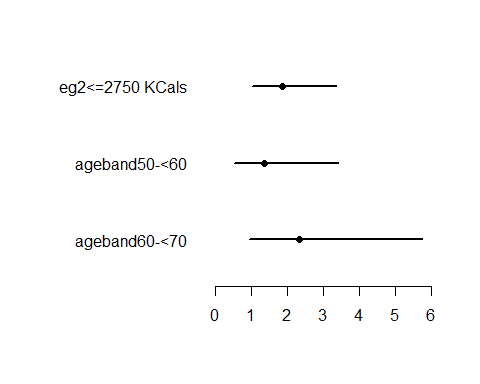
1. Update the first version above adding ageband to the model formula and save the estimates processed by function ci.lin() for further use

mea <- update( me, . ~ . + ageband)  
mea.est <- ci.lin(mea, Exp=T)  
round(mea.est[ , -(3:4)], 4)

Estimate StdErr exp(Est.) 2.5% 97.5%  
(Intercept) -5.4033 0.4390 0.0045 0.0019 0.0106  
eg2<=2750 KCals 0.6233 0.3027 1.8651 1.0306 3.3753  
ageband50-<60 0.3027 0.4721 1.3535 0.5366 3.4145  
ageband60-<70 0.8456 0.4613 2.3294 0.9431 5.7535

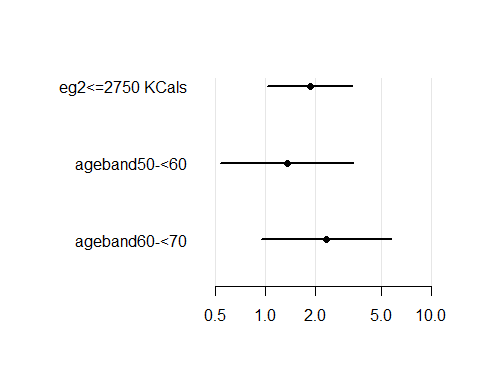
1. Display the hazard ratio estimates and CIs graphically by function plotEst() in Epi:

plotEst( mea.est[ -1, 5:7 ] )



You can have the x-axis on the logarithmic scale and provide some reference lines, too:

plotEst( mea.est[ -1, 5:7 ], xlog=T, grid=T )



Again, these figures could be improved in various ways.

## 6. Evaluating effect modification

Model mea fitted above included only the main effects of age and exposure. This model is equivalent to the assumption that the rate ratio between the low and high energy intake is the same (or homogenous) across all the agebands. A model that allows the rate ratio of interest to be different in the three agebands is obtained by including an interaction term between ageband and energy intake group in the model formula.

1. We shall now fit a model including both the main effects of eg2 and ageband and their interaction:

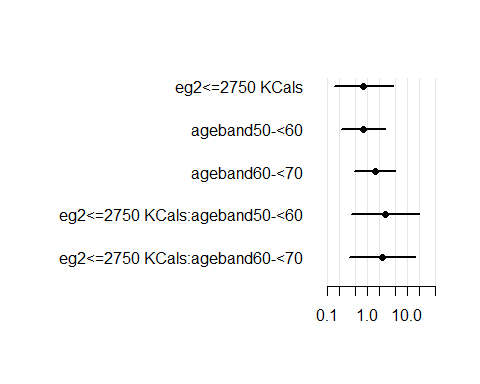
meai <- glm( d\_ik/y\_ik ~ eg2\*ageband,  
family = poisson, we = y\_ik, data = dietA )

Instead of the concise representation eg2\*ageband covering both the main effects and the interaction we could write more explicitly eg2 + ageband + eg2:agebandâ in the model formula. View the results:

est.meai <- ci.lin(meai, Exp=T)  
round(est.meai, 4)

Estimate StdErr z P exp(Est.) 2.5%  
(Intercept) -5.0473 0.5000 -10.0945 0.0000 0.0064 0.0024  
eg2<=2750 KCals -0.2023 0.8660 -0.2336 0.8153 0.8169 0.1496  
ageband50-<60 -0.1889 0.6455 -0.2927 0.7698 0.8279 0.2336  
ageband60-<70 0.4494 0.6124 0.7339 0.4630 1.5674 0.4720  
eg2<=2750 KCals:ageband50-<60 1.0365 1.0000 1.0365 0.3000 2.8193 0.3971  
eg2<=2750 KCals:ageband60-<70 0.8893 0.9728 0.9141 0.3606 2.4335 0.3615  
 97.5%  
(Intercept) 0.0171  
eg2<=2750 KCals 4.4598  
ageband50-<60 2.9336  
ageband60-<70 5.2051  
eg2<=2750 KCals:ageband50-<60 20.0138  
eg2<=2750 KCals:ageband60-<70 16.3798

plotEst( est.meai[ -1, 5:7], xlog=T, grid = T)



How would you interpret these results?

1. The confidence limits of the individual interaction parameters appear to be very wide. A global test for the interaction effect or against the H0 for homogeneity of rate ratios is provided by the corresponding likelihood ratio statistic, which contrasts the residual deviance of the model without interaction to that of the model with interaction. As there are two free interaction parameters, the appropriate reference distribution is the chi-square with 2 df.

Dint <- anova(mea, meai) ; Dint

Analysis of Deviance Table  
  
Model 1: d\_ik/y\_ik ~ eg2 + ageband  
Model 2: d\_ik/y\_ik ~ eg2 \* ageband  
 Resid. Df Resid. Dev Df Deviance  
1 725 311.44   
2 723 310.26 2 1.1734

pchisq( Dint[2,4], df = 2, lower.tail = F)

[1] 0.556164

What is your conclusion based on this test: \* H0: There is evidence for no interaction, or \* H1: There is no evidence for interaction?

## 7. Ageband as a quantitative variable

So far we have treated ageband as a categorical factor allowing a free functional form for the age effect. We have assumed, though, that the true rate within each ageband is not varying (piecewise constant hazards model). Alternatively, ageband can be treated as a continuous variable assuming a suitable smooth function for the age effect. We have three agebands, each having a width of 10 years. The mid-point of each ageband can be taken to represent the quantitative age value in that ageband.

1. We shall first convert the ageband factor into a numeric variable using the as.numeric() function, and check how succesful this operation was:

dietA$agebnum <- as.numeric(dietA$ageband)  
table(dietA$ageband, dietA$agebnum)

1 2 3  
 30-<50 196 0 0  
 50-<60 0 293 0  
 60-<70 0 0 240

1. In modelling we could use these numeric scores 1, 2, and 3 as such. A natural alternative appears be to use the mid-year values of each ageband:

dietA$agebyrs <- 35 + 10\*dietA$agebnum  
table(dietA$ageband, dietA$agebyrs)

45 55 65  
 30-<50 196 0 0  
 50-<60 0 293 0  
 60-<70 0 0 240

1. However, very often it is more advisable to use centered values for a quantitative variable. For example, instead of age values 1, 2, and 3, we could recode these into 1; 0; and 1.These values define a linear contrast between the ageband levels:

dietA$agebcontr <- dietA$agebnum - 2  
table(dietA$ageband, dietA$agebcontr)

-1 0 1  
 30-<50 196 0 0  
 50-<60 0 293 0  
 60-<70 0 0 240

Similarly, we could use centered coding for the mid-ages, too

dietA$ageby55 <- dietA$agebyrs - 55  
table(dietA$ageband, dietA$agebcontr)

-1 0 1  
 30-<50 196 0 0  
 50-<60 0 293 0  
 60-<70 0 0 240

Actually, the values 1; 0; 1 represent a centered and scaled version of the age values 45, 55, and 65; they are centered at 55 years and scaled by 10 years.

1. We shall first fit a model with quantitative age using the uncentered mid-ages in years

mea10 <- glm( d\_ik ~ agebyrs + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea10 <- ci.lin( mea10, Exp = T)  
round( est.mea10[ , 5:7], 4)

exp(Est.) 2.5% 97.5%  
(Intercept) 0.0005 0.0000 0.0064  
agebyrs 1.0469 1.0034 1.0923  
eg2<=2750 KCals 1.8574 1.0268 3.3597

Here the intercept refers to the baseline rate at age 0 years! The slope parameter for agebyrs is equal to the change in log(RR) per an increase of 1 year of age. This is not the most comprehensible parametrization. Compare the results of this with model employing the centered version of age

mea10c <- glm( d\_ik ~ ageby55 + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea10c <- ci.lin( mea10c, Exp = T)  
round( est.mea10c[ , 5:7], 4)

exp(Est.) 2.5% 97.5%  
(Intercept) 0.0065 0.0040 0.0105  
ageby55 1.0469 1.0034 1.0923  
eg2<=2750 KCals 1.8574 1.0268 3.3597

The intercept parameter has now a more meaningful interpretation (BTW: what does it mean here?). But the age effect seems to be quite modest, or is it?

1. We shall now fit a model with continuous age using the centered and scaled version of it, i.e. the contrast coding 1; 0; 1:

mea1c <- glm( d\_ik ~ agebcontr + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea1c <- ci.lin( mea1c, Exp = T)  
round( est.mea1c[ , 5:7], 4)

exp(Est.) 2.5% 97.5%  
(Intercept) 0.0065 0.0040 0.0105  
agebcontr 1.5815 1.0342 2.4183  
eg2<=2750 KCals 1.8574 1.0268 3.3597

Here the intercept refers to the baseline rate for ageband 50-59 y, as in the previous model. Now, however, the slope for agebcontr describes the change in the log(RR) per an increase of 10 years of age (instead of 1 year).

## 8. Quadratic or 2nd order polynomial function for the effect of age?

1. The model with linear effect of age assumes that in both energy groups the logarithm of the true rate changes by the same amount for the same change in age. Maybe the pattern of the age effect is not so simple. We can allow the age effect to be curved or non-linear in many ways. The simplest way is to assume a parabolic form for the age-incidence curve. This can be accomplished by fitting a quadratic term of age in addition to the linear term. It is possible to compute the quadratic term by an ordinary variable transformation from the linear term in the data frame, like

dietA$agebyrs2 <- dietA$agebyrs^2

However, this is not necessary, because we can include the quadratic term directly in the model formula using the insulate operation provided by function I():

mea20 <- glm( d\_ik ~ agebyrs + I(agebyrs^2) + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea20 <- ci.lin( mea20, Exp = T)  
round( est.mea20, 4)

Estimate StdErr z P exp(Est.) 2.5% 97.5%  
(Intercept) -3.7937 10.3829 -0.3654 0.7148 0.0225 0.0000 1.550179e+07  
agebyrs -0.0898 0.3726 -0.2410 0.8095 0.9141 0.4404 1.897200e+00  
I(agebyrs^2) 0.0012 0.0033 0.3641 0.7157 1.0012 0.9948 1.007700e+00  
eg2<=2750 KCals 0.6233 0.3026 2.0595 0.0394 1.8651 1.0306 3.375300e+00

Interpretation of parameters? Is the effect of age no more significant at all?

1. In the previous item we encountered the technical problem of collinearity, i.e. high correlation between the linear and the quadratic term of age, when expressed as years since birth. Collinearity brings about very imprecise estimates of the parameters. This can be avoided using the centered ages

mea20c <- glm( d\_ik ~ ageby55 + I(ageby55^2) + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea20c <- ci.lin( mea20c, Exp = T)  
round( est.mea20c, 4)

Estimate StdErr z P exp(Est.) 2.5% 97.5%  
(Intercept) -5.1006 0.3009 -16.9486 0.0000 0.0061 0.0034 0.0110  
ageby55 0.0423 0.0231 1.8330 0.0668 1.0432 0.9971 1.0914  
I(ageby55^2) 0.0012 0.0033 0.3641 0.7157 1.0012 0.9948 1.0077  
eg2<=2750 KCals 0.6233 0.3026 2.0595 0.0394 1.8651 1.0306 3.3753

1. Even better it may be to use the centered and scaled version, i.e. the contrast coding 1; 0; 1:

mea2c <- glm( d\_ik ~ agebcontr + I(agebcontr^2) + eg2,  
family = poisson, offset = log(y\_ik), data=dietA)  
est.mea2c <- ci.lin( mea2c, Exp = T)  
round( est.mea2c, 4)

Estimate StdErr z P exp(Est.) 2.5% 97.5%  
(Intercept) -5.1006 0.3009 -16.9486 0.0000 0.0061 0.0034 0.0110  
agebcontr 0.4228 0.2307 1.8330 0.0668 1.5262 0.9711 2.3986  
I(agebcontr^2) 0.1201 0.3297 0.3641 0.7157 1.1276 0.5908 2.1519  
eg2<=2750 KCals 0.6233 0.3026 2.0595 0.0394 1.8651 1.0306 3.3753

Either way, the analysis based on centered age is consistent with common knowledge that the incidence of CHD increases with age. Another result, implied by the Wald statistic for the quadratic term, seems to be that our data does not provide sufficient evidence against the simple model according to which the log-rate depends linearly on age. (Remember, though, that our data is also well consistent with the possibility of a curved age-logincidence pattern.)

1. We can also perform a likelihood ratio test against the simple assumption that a linear term for age is sufficient. We contrast the residual deviances: linear age model vs. quadratic age model:

anova( mea1c, mea2c )

Analysis of Deviance Table  
  
Model 1: d\_ik ~ agebcontr + eg2  
Model 2: d\_ik ~ agebcontr + I(agebcontr^2) + eg2  
 Resid. Df Resid. Dev Df Deviance  
1 726 311.57   
2 725 311.44 1 0.13219

## 9. Age-standardised incidence rates

We use popEpi package to calculate the age-standardised incidence by calendar year in the whole population of Finland. Data sibr includes the dates of diagnosis and the ages at diagnosis.

library(popEpi)  
data(sibr)  
  
coh <- transform(sibr, dg\_y = cal.yr(dg\_date) )  
coh$agegroup <- cut(coh$dg\_age, breaks=c(seq(0,85,5),Inf), right=F)  
coh$dg\_y <- floor(coh$dg\_y)

Calculate the number of cases by agegroup and calendar year

tab <- stat.table( index = list(agegroup,year=dg\_y),   
 contents = list(count() ),   
 margins = F, data = coh)

Tabulate the mid-year population in Finland in 5-year age groups

pop <- meanpop\_fi[sex==1 & year%in%1993:2012 ]  
pop$agegroup <- cut(pop$agegroup, breaks=c(seq(0,85,5),Inf), right=F)  
  
tab.pop <- stat.table( index = list(agegroup,year),   
 contents = list(meanpop=sum(meanpop) ),   
 margins = F, data = pop)

Combine the numbers of cases and person-years

#combine the numbers of cases and person-years  
tab <- merge(data.frame(as.table(tab[1,,])),   
 data.frame(as.table(tab.pop[1,,])),by=c("agegroup","year"))  
names(tab)[3:4] <- c("obs","pyrs")  
tab$year <- as.numeric(as.character(tab$year))

Choose a standard population: population of Finland in 2012

own\_weights <- tab[tab$year==2012,]  
own\_weights <- own\_weights[order(own\_weights$agegroup),]

rate() function in popEpi package can be used to calculate the rates with confidence intervals

tab1 <- rate(tab, obs="obs", pyrs="pyrs", adjust="agegroup",   
 weights="world\_1966\_18of5", print="year")  
#OR equivalently:   
#tab1 <- rate(tab, obs="obs", pyrs="pyrs", adjust="agegroup",   
# weights=stdpop18$world, print="year")  
tab2 <- rate(tab, obs="obs", pyrs="pyrs", adjust="agegroup",   
 weights=own\_weights$pyrs, print="year")

# Age-standardised incidence rate in population aged 25+

Restrict to 25+ population by setting the weights of the youngest age groups to 0.

own\_weights$pyrs25 <- own\_weights$pyrs  
own\_weights$pyrs25[1:5] <- 0  
  
#25+ population of World in 1966 as another standard  
stdpop18$world25 <- stdpop18$world  
stdpop18$world25[1:5] <- 0

Age-standardised incidence is calculated using the weights where the weights of the youngest age groups are 0.

tab1\_25 <- rate(tab, obs="obs", pyrs="pyrs", adjust="agegroup", weights=stdpop18$world25, print="year")  
tab2\_25 <- rate(tab, obs="obs", pyrs="pyrs", adjust="agegroup", weights=own\_weights$pyrs25, print="year")

Non-standardised incidence has to be calculated by excluding the youngest age groups.

tab0\_25 <- rate(tab[as.numeric(tab$agegroup)>5,], obs="obs", pyrs="pyrs", print="year")

Compare age-standardised rates using different standard populations

plot(as.numeric(tab1$year),tab1$rate\*100000, type="l",ylim=c(0, 50),  
 ylab="Incidence per 100 000 person years")  
lines(as.numeric(tab1$year),tab1$rate.adj\*100000, col="red")  
lines(as.numeric(tab2$year),tab2$rate.adj\*100000, col="green")  
  
lines(as.numeric(tab0\_25$year),tab0\_25$rate\*100000,lty=2)  
lines(as.numeric(tab1\_25$year),tab1\_25$rate.adj\*100000, col="red",lty=2)  
lines(as.numeric(tab2\_25$year),tab2\_25$rate.adj\*100000, col="green",lty=2)  
  
legend("topleft",c("Non-standardised incidence",  
 "Age-standardised incidence (World 1966)",  
 "Age-standardised incidence (Finland 2012)",  
 "Non-standardised incidence in 25+ population",  
 "Age-standardised incidence in 25+ population (World 1966)",  
 "Age-standardised incidence in 25+ population (Finland 2012)"),  
 col=c("black","green","red","black","green","red"),lty=c(1,1,1,2,2,2))

