Epidemiological Methods in Medical Research

Computer practicals

Department of Biostatistics
Institute of Public Health, University of Copenhagen
Spring 2023
http://BendixCarstensen.com/EpiPhD/F2022
Version 3.1

Compiled Wednesday 15th March, 2023, 08:34
from: C:\Bendix\teach\Epi\KU-epi\pracs/pracs.tex

Bendix Carstensen Steno Diabetes Center, Herlev, Denmark
& Department of Biostatistics,
Institute of Public Health, University of Copenhagen
b@bxc.dk
http://BendixCarstensen.com

There are two sections in this document; the exercises which contains exercises 1–11, and the solutions part which contains items 1–18, the last part being the part with parametric models that we will walk through together.

0.1 IHD exercise

The following instructions are fairly detailed. You should make sure that you know what goes on, and that consult the help-pages for the functions uses, so that you get a bit of a feeling for how the R-machinery works.

1. Load the `Epi` package and read the (modified) grouped IHD-data from the file `ihd-xtab.dta` from the data folder
   http://BendixCarstensen.com/EpiPhD/F2014/data

   ```
   > options(width=90)
   > library(Epi)
   > library(foreign)
   > ihdt <- read.table(
   +     "http://BendixCarstensen.com/EpiPhD/F2014/data/ihd-tab.txt",
   +     header=T )
   > ihdt
   ```

   Fit a Poisson model to data with exposure and age-effects:

   ```
   > mt <- glm(cases ~ factor(age) + exposure,
   +           offset = log(pyrs),
   +           family = poisson,
   +             data = ihdt)
   > round(ci.exp(mt), 3 )
   ```

   The outcome in a follow-up study is really (event, outcome), so it is more logical to have these combined in the outcome. This is done in the family `poisreg`; the same model can be fitted with a 2-column matrix (`cbind(,)`) as outcome:

   ```
   > mT <- glm(cbind(cases, pyrs) ~ factor(age) + exposure,
   +           family = poisreg,
   +             data = ihdt)
   > round(ci.exp(mT), 3 )
   ```

   Compare with the results from table 24.1 in Clayton & Hills.

2. Next, read the individual records from the file `diet.txt`; remembering to specify how missing is coded:

   ```
   > ihdi <- read.table(
   +     "http://BendixCarstensen.com/EpiPhD/F2020/diet.txt",
   +     header = TRUE,
   +     na.strings = ".",
   +     stringsAsFactors = FALSE)
   > head( ihdi )
   > str( ihdi )
   > # Turn character variables into dates and then to calendar years:XS
   ```

```
> for (i in c(2, 3, 5)) ihdi[,i] <- cal.yr(as.Date(ihdi[,i],
+                                                  format = "%m/%d/%Y"))
> str( ihdi )
> head( ihdi )
```

Now check that it looks reasonable and that you understand what the data represents.

3. Now you should set up the dataset as a `Lexis` object[1]., so that R will know when persons are at risk etc. `entry` is a named list, the names giving the names of the timescales we want to use, in this case `per` (calendar time, `period`) and age. `exit` is also a named list, with one element with the name of one of the timescales, giving the values of the exit times on this time scale. `exit.status` gives the state that persons are in at exit from the study. If `entry.status` is not specified, it is assumed that everyone starts in the *first* state, and this is noted:

```
> Lx <- Lexis(entry = list(per = doe,
+                          age = doe - dob),
+             exit = list(per = dox),
+       exit.status = factor(chd, labels = c("Well", "IHD")),
+             data = ihdi)
> summary(Lx)
```

There is a method for plotting the follow-up in boxes. Not desperately exciting but capturing the essence:

```
> boxes(Lx,
+       boxpos = TRUE,
+       scal.Y = 1000,
+      show.BE = TRUE)
```

4. The time-splitting is now done by the function `splitLexis`. To use the function we must specify which timescale to split the data on. In this case we want to split along the scale "current age", i.e. time since date of birth, here named `age`. We then specify the intervals where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70, so use the breakpoints 40, 50, 60 and 70:

```
> Ls <- splitLexis(Lx,
+             breaks = c(40, 50, 60, 70),
+          time.scale = "age")
> summary(Lx)
> summary(Ls)
> head(Ls)
```

For the fun of it you can try the default `plot` and `points` methods for a `Lexis` object. Note that grid-lines corresponding to the breaks gets inserted:

[1]Named after the German demographer, statistician and economist, Wilhelm Lexis, 1837–1914. He wrote the book "Einführung in die Theorie der Bevölkerungsstatistik, (Strassbourg, 1875)", while he was professor in Dorpat (now Tartu, Estonia), wherein he devised the so called Lexis diagram.

```
> plot  (Ls, col = gray(0.3))
> points(Ls, col = "red",
+            pch = c(NA, 16)[Ls$lex.Xst],
+            cex = 0.7)
```

On the diagram it appears that all persons are censored at age 70 and at the end of 1976, whereas some follow-up time is present before age 40.

5. The number of records are in the resulting dataset (`Ls`):

```
> nrow(Ls)
```

6. List the first 20 records:

```
> head(Ls, 20)
```

7. Now reproduce the table in Clayton & Hills:

   First use the function `timeBand` to produce a variable which is equal to the left endpoint of the intervals into which the follow-up have been split:

```
> Ls <- transform(Ls, agr = timeBand(Ls, "age", "factor"),
+                     eksp = factor(energy < 2.75,
+                                   labels = c("High", "Low")))
> str( Ls )
```

   Then make a table like the one in C& H:

```
> round(ftable(xtabs(cbind(D = (lex.Xst=="IHD"),
+                          Y = lex.dur)
+                    ~ agr + eksp,
+                    data = Ls),
+              row.vars = 1), 2)
```

   You should see that the data is not quite the same as in the book.

8. Do the grouped analysis on the slightly modified data that you can get from the data folder (which should be identical to the table you just made):

```
> ihdx <- read.table(
+     "http://BendixCarstensen.com/EpiPhD/F2020/ihd-xtab.txt", header=T )
> ihdx
> mt <- glm(cbind(cases, pyrs) ~ factor(age) + exposure,
+           family = poisreg,
+             data = ihdx)
> round(ci.exp(mt), 3 )
```

9. Estimate the effect of age and exposure from the split dataset. Remember to exclude follow-uptime before age 40 — as you saw from the table above:

```
> Ls <- subset(Ls, agr %in% levels(agr)[2:4])
> Ls$agr <- factor(Ls$agr)
> table(Ls$agr)
> head(Ls)
> mi <- glm(cbind(lex.Xst=="IHD", lex.dur) ~ factor(agr) + eksp,
+           family = poisreg,
+             data = Ls)
> round(ci.exp(mi), 3 )
> round(ci.exp(mt), 3 )
```

We see that the estimates are identical for the two ways of modeling. The point of using the individual data is that individual-level variables could be included in a model too.

10. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.

```
> mix <- update(mi, . ~ . + factor(agr):eksp)
> mtx <- update(mt, . ~ . + factor(age):exposure)
> anova(mi, mix, test="Chisq")
> anova(mt, mtx, test="Chisq")
```

11. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data. You can get the deviance from `deviance`:

```
> deviance(mt)
```

0.2 IHD exercise—solutions

The following instructions are fairly detailed. You should make sure that you know what goes on, and that consult the help-pages for the functions uses, so that you get a bit of a feeling for how the R-machinery works.

1. Load the `Epi` package and read the (modified) grouped IHD-data from the file `ihd-xtab.dta` from the data folder
   http://BendixCarstensen.com/EpiPhD/F2014/data

```
> options(width=90)
> library(Epi)
> library(foreign)
> ihdt <- read.table(
+     "http://BendixCarstensen.com/EpiPhD/F2014/data/ihd-tab.txt",
+     header=T )
> ihdt
```

```
  exposure age   pyrs cases
1        1   0  311.9     2
2        1   1  878.1    12
3        1   2  667.5    14
4        0   0  607.9     4
5        0   1 1272.1     5
6        0   2  888.9     8
```

Fit a Poisson model to data with exposure and age-effects:

```
> mt <- glm(cases ~ factor(age) + exposure,
+           offset = log(pyrs),
+           family = poisson,
+             data = ihdt)
> round(ci.exp(mt), 3 )
```

```
             exp(Est.)  2.5% 97.5%
(Intercept)      0.004 0.002 0.011
factor(age)1     1.138 0.448 2.888
factor(age)2     1.998 0.809 4.935
exposure         2.386 1.305 4.364
```

The outcome in a follow-up study is really (event, outcome), so it is more logical to have these combined in the outcome. This is done in the family `poisreg`; the same model can be fitted with a 2-column matrix (`cbind(,)`) as outcome:

```
> mT <- glm(cbind(cases, pyrs) ~ factor(age) + exposure,
+           family = poisreg,
+             data = ihdt)
> round(ci.exp(mT), 3 )
```

```
             exp(Est.)  2.5% 97.5%
(Intercept)      0.004 0.002 0.011
factor(age)1     1.138 0.448 2.888
factor(age)2     1.998 0.809 4.935
exposure         2.386 1.305 4.364
```

Compare with the results from table 24.1 in Clayton & Hills.

2. Next, read the individual records from the file `diet.txt`; remembering to specify how missing is coded:

```
> ihdi <- read.table(
+     "http://BendixCarstensen.com/EpiPhD/F2020/diet.txt",
+     header = TRUE,
+     na.strings = ".",
+     stringsAsFactors = FALSE)
> head( ihdi )
```

```
  id        doe        dox chd        dob job month  energy   height   weight    fat
1  1 08/16/1964 12/01/1976   0 01/04/1915   0     8 2.87395 175.3870 71.48737 141.71
2  2 12/16/1964 12/01/1976   0 06/03/1914   0    12 1.98234 164.2872 70.08120  85.77
3  3 11/16/1965 12/01/1976   0 02/03/1907   0    11 2.66858 169.3926 71.89560 107.67
4  4 09/16/1965 12/01/1976   0 12/25/1906   0     9 2.83669 167.0050 74.88937 132.17
5  5 09/16/1965 03/31/1976   0 04/01/1906   0     9 2.94150 174.4980 78.38208 126.35
6  6 03/16/1965 08/31/1968   0 03/23/1914   0     3 2.47351 176.5046 72.39456 103.10
  fibre
1 17.83
2  9.49
3 15.99
4 17.04
5 14.54
6 12.49
```

```
> str( ihdi )
```

```
'data.frame':        337 obs. of  12 variables:
 $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ doe   : chr  "08/16/1964" "12/16/1964" "11/16/1965" "09/16/1965" ...
 $ dox   : chr  "12/01/1976" "12/01/1976" "12/01/1976" "12/01/1976" ...
 $ chd   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dob   : chr  "01/04/1915" "06/03/1914" "02/03/1907" "12/25/1906" ...
 $ job   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ month : int  8 12 11 9 9 3 11 5 2 7 ...
 $ energy: num  2.87 1.98 2.67 2.84 2.94 ...
 $ height: num  175 164 169 167 174 ...
 $ weight: num  71.5 70.1 71.9 74.9 78.4 ...
 $ fat   : num  141.7 85.8 107.7 132.2 126.3 ...
 $ fibre : num  17.83 9.49 15.99 17.04 14.54 ...
```

```
> # Turn character variables into dates and then to calendar years:XS
> for (i in c(2, 3, 5)) ihdi[,i] <- cal.yr(as.Date(ihdi[,i],
+                                                  format = "%m/%d/%Y"))
> str( ihdi )
```

```
'data.frame':        337 obs. of  12 variables:
 $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ doe   : 'cal.yr' num  1965 1965 1966 1966 1966 ...
 $ dox   : 'cal.yr' num  1977 1977 1977 1977 1976 ...
 $ chd   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dob   : 'cal.yr' num  1915 1914 1907 1907 1906 ...
 $ job   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ month : int  8 12 11 9 9 3 11 5 2 7 ...
 $ energy: num  2.87 1.98 2.67 2.84 2.94 ...
 $ height: num  175 164 169 167 174 ...
 $ weight: num  71.5 70.1 71.9 74.9 78.4 ...
 $ fat   : num  141.7 85.8 107.7 132.2 126.3 ...
 $ fibre : num  17.83 9.49 15.99 17.04 14.54 ...
```

```
> head( ihdi )
```

```
  id      doe      dox chd      dob job month  energy   height   weight    fat fibre
1  1 1964.623 1976.916   0 1915.008   0     8 2.87395 175.3870 71.48737 141.71 17.83
2  2 1964.957 1976.916   0 1914.419   0    12 1.98234 164.2872 70.08120  85.77  9.49
3  3 1965.874 1976.916   0 1907.090   0    11 2.66858 169.3926 71.89560 107.67 15.99
4  4 1965.707 1976.916   0 1906.980   0     9 2.83669 167.0050 74.88937 132.17 17.04
5  5 1965.707 1976.245   0 1906.246   0     9 2.94150 174.4980 78.38208 126.35 14.54
6  6 1965.203 1968.664   0 1914.222   0     3 2.47351 176.5046 72.39456 103.10 12.49
```

Now check that it looks reasonable and that you understand what the data represents.

3. Now you should set up the dataset as a `Lexis` object[2]., so that R will know when persons are at risk etc. `entry` is a named list, the names giving the names of the timescales we want to use, in this case `per` (calendar time, `period`) and age. `exit` is also a named list, with one element with the name of one of the timescales, giving the values of the exit times on this time scale. `exit.status` gives the state that persons are in at exit from the study. If `entry.status` is not specified, it is assumed that everyone starts in the *first* state, and this is noted:

[2]Named after the German demographer, statistician and economist, Wilhelm Lexis, 1837–1914. He wrote the book "Einführung in die Theorie der Bevölkerungsstatistik", (Strassbourg, 1875), while he was professor in Dorpat (now Tartu, Estonia), wherein he devised the so called Lexis diagram.

```
> Lx <- Lexis(entry = list(per = doe,
+                          age = doe - dob),
+             exit = list(per = dox),
+       exit.status = factor(chd, labels = c("Well", "IHD")),
+             data = ihdi)

NOTE: entry.status has been set to "Well" for all.

> head(Lx)

 lex.id     per   age lex.dur lex.Cst lex.Xst id      doe      dox chd      dob job month
      1 1964.62 49.62   12.29    Well    Well  1 1964.623 1976.916   0 1915.008   0     8
      2 1964.96 50.54   11.96    Well    Well  2 1964.957 1976.916   0 1914.419   0    12
      3 1965.87 58.78   11.04    Well    Well  3 1965.874 1976.916   0 1907.090   0    11
      4 1965.71 58.73   11.21    Well    Well  4 1965.707 1976.916   0 1906.980   0     9
      5 1965.71 59.46   10.54    Well    Well  5 1965.707 1976.245   0 1906.246   0     9
      6 1965.20 50.98    3.46    Well    Well  6 1965.203 1968.664   0 1914.222   0     3
 energy  height weight    fat fibre
  2.874 175.387 71.487 141.71 17.83
  1.982 164.287 70.081  85.77  9.49
  2.669 169.393 71.896 107.67 15.99
  2.837 167.005 74.889 132.17 17.04
  2.942 174.498 78.382 126.35 14.54
  2.474 176.505 72.395 103.10 12.49

> summary(Lx)

Transitions:
     To
From   Well IHD  Records:  Events: Risk time:  Persons:
  Well  291  46       337       46    4603.67       337
```

Since the units of the time variables `age` and `per` is years, the risk time is also in years.

For further illustration we identify persons who had and event after quite long follow up (more than 15 years):

```
> (who <- subset(Lx, lex.Xst == "IHD" & lex.dur > 15)$lex.id[1:2])

[1] 258 280

> subset(Lx, lex.id %in% who)

 lex.id     per   age lex.dur lex.Cst lex.Xst  id      doe      dox chd      dob job
    258 1958.04 47.90   17.82    Well     IHD 258 1958.041 1975.862   1 1910.142   2
    280 1959.29 52.76   15.21    Well     IHD 280 1959.287 1974.493   1 1906.528   2
 month energy  height weight    fat fibre
     1  2.564 170.180 66.226  99.72 15.15
     4  2.814 175.895 80.287 125.35 16.92
```

There is a method for plotting the follow-up in boxes. Not desperately exciting but capturing the essence:

```
> boxes(Lx,
+       boxpos = TRUE,
+      scale.Y = 1000,
+      show.BE = TRUE)
```

4. The time-splitting is now done by the function `splitLexis`. To use the function we must specify which timescale to split the data on. In this case we want to split along the scale "current age", i.e. time since date of birth, here named `age`. We then specify the intervals where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70, so use the breakpoints 40, 50, 60 and 70:

```
> Ls <- splitLexis(Lx,
+              breaks = c(40, 50, 60, 70),
+          time.scale = "age")
> summary(Lx)

Transitions:
     To
From   Well IHD  Records:  Events: Risk time:  Persons:
  Well  291  46       337       46    4603.67       337

> summary(Ls)

Transitions:
     To
From   Well IHD  Records:  Events: Risk time:  Persons:
  Well  709  46       755       46    4603.67       337

> subset(Lx, lex.id %in% who)

 lex.id     per   age lex.dur lex.Cst lex.Xst  id      doe      dox chd      dob job
    258 1958.04 47.90   17.82    Well     IHD 258 1958.041 1975.862   1 1910.142   2
    280 1959.29 52.76   15.21    Well     IHD 280 1959.287 1974.493   1 1906.528   2
 month energy  height weight    fat fibre
     1  2.564 170.180 66.226  99.72 15.15
     4  2.814 175.895 80.287 125.35 16.92

> subset(Ls, lex.id %in% who)

 lex.id     per   age lex.dur lex.Cst lex.Xst  id      doe      dox chd      dob job
    258 1958.04 47.90    2.10    Well    Well 258 1958.041 1975.862   1 1910.142   2
    258 1960.14 50.00   10.00    Well    Well 258 1958.041 1975.862   1 1910.142   2
    258 1970.14 60.00    5.72    Well     IHD 258 1958.041 1975.862   1 1910.142   2
    280 1959.29 52.76    7.24    Well    Well 280 1959.287 1974.493   1 1906.528   2
    280 1966.53 60.00    7.96    Well     IHD 280 1959.287 1974.493   1 1906.528   2
 month energy  height weight    fat fibre
     1  2.564 170.180 66.226  99.72 15.15
     1  2.564 170.180 66.226  99.72 15.15
     1  2.564 170.180 66.226  99.72 15.15
     4  2.814 175.895 80.287 125.35 16.92
     4  2.814 175.895 80.287 125.35 16.92
```

For the fun of it you can try the default `plot` and `points` methods for a `Lexis` object. Note that grid-lines corresponding to the breaks gets inserted:

```
> plot  (Ls, col = gray(0.3))
> points(Ls, col = "red",
+            pch = c(NA, 16)[Ls$lex.Xst],
+            cex = 0.7)
```

5. What can you see from the graph?

   On the diagram it appears that all persons are censored at age 70 and at the end of 1976, whereas some follow-up time is present before age 40. Also it appears that persons have been included at different times.

   Does this matter for the conclusions on IHD occurrence?

6. The number of records are in the resulting dataset (`Ls`):

```
> nrow(Ls)
[1] 755
```

7. List the first 20 records:

```
> head(Ls, 20)
```

lex.id	per	age	lex.dur	lex.Cst	lex.Xst	id	doe	dox	chd	dob	job	month
1	1964.62	49.62	0.38	Well	Well	1	1964.623	1976.916	0	1915.008	0	8
1	1965.01	50.00	10.00	Well	Well	1	1964.623	1976.916	0	1915.008	0	8
1	1975.01	60.00	1.91	Well	Well	1	1964.623	1976.916	0	1915.008	0	8
2	1964.96	50.54	9.46	Well	Well	2	1964.957	1976.916	0	1914.419	0	12
2	1974.42	60.00	2.50	Well	Well	2	1964.957	1976.916	0	1914.419	0	12
3	1965.87	58.78	1.22	Well	Well	3	1965.874	1976.916	0	1907.090	0	11
3	1967.09	60.00	9.83	Well	Well	3	1965.874	1976.916	0	1907.090	0	11
4	1965.71	58.73	1.27	Well	Well	4	1965.707	1976.916	0	1906.980	0	9
4	1966.98	60.00	9.94	Well	Well	4	1965.707	1976.916	0	1906.980	0	9
5	1965.71	59.46	0.54	Well	Well	5	1965.707	1976.245	0	1906.246	0	9
5	1966.25	60.00	10.00	Well	Well	5	1965.707	1976.245	0	1906.246	0	9
6	1965.20	50.98	3.46	Well	Well	6	1965.203	1968.664	0	1914.222	0	3
7	1958.87	45.14	4.86	Well	Well	7	1958.873	1976.916	0	1913.734	0	11
7	1963.73	50.00	10.00	Well	Well	7	1958.873	1976.916	0	1913.734	0	11
7	1973.73	60.00	3.18	Well	Well	7	1958.873	1976.916	0	1913.734	0	11
8	1965.37	50.43	9.57	Well	Well	8	1965.370	1976.916	0	1914.942	0	5
8	1974.94	60.00	1.97	Well	Well	8	1965.370	1976.916	0	1914.942	0	5
9	1959.13	67.10	2.90	Well	Well	9	1959.125	1962.025	0	1892.029	0	2
10	1964.54	60.17	9.83	Well	Well	10	1964.538	1974.370	0	1904.371	0	7
11	1964.79	60.52	9.48	Well	Well	11	1964.790	1974.266	0	1904.267	0	10

energy	height	weight	fat	fibre
2.874	175.387	71.487	141.71	17.83
2.874	175.387	71.487	141.71	17.83
2.874	175.387	71.487	141.71	17.83
1.982	164.287	70.081	85.77	9.49
1.982	164.287	70.081	85.77	9.49
2.669	169.393	71.896	107.67	15.99
2.669	169.393	71.896	107.67	15.99
2.837	167.005	74.889	132.17	17.04
2.837	167.005	74.889	132.17	17.04
2.942	174.498	78.382	126.35	14.54
2.942	174.498	78.382	126.35	14.54
2.474	176.505	72.395	103.10	12.49
2.556	168.910	64.184	111.54	16.35
2.556	168.910	64.184	111.54	16.35
2.556	168.910	64.184	111.54	16.35
2.988	165.989	73.801	159.53	16.09
2.988	165.989	73.801	159.53	16.09

```
  2.311 165.710 49.080 115.68 15.44
  3.125 181.204 78.291 114.43 18.89
  2.161 174.193 63.912 111.16 11.66
```

8. Now reproduce the table in Clayton & Hills:

   First use the function `timeBand` to produce a variable which is equal to the left endpoint of the intervals into which the follow-up have been split:

```
> Ls <- transform(Ls, agr = timeBand(Ls, "age", "factor"),
+                     eksp = factor(energy < 2.75,
+                                   labels = c("High", "Low")))
> str( Ls )
```

```
Classes 'Lexis' and 'data.frame':        755 obs. of  20 variables:
 $ lex.id : int  1 1 1 2 2 3 3 4 4 5 ...
 $ per    : num  1965 1965 1975 1965 1974 ...
 $ age    : num  49.6 50 60 50.5 60 ...
 $ lex.dur: num  0.385 10 1.908 9.462 2.497 ...
 $ lex.Cst: Factor w/ 2 levels "Well","IHD": 1 1 1 1 1 1 1 1 1 1 ...
 $ lex.Xst: Factor w/ 2 levels "Well","IHD": 1 1 1 1 1 1 1 1 1 1 ...
 $ id     : int  1 1 1 2 2 3 3 4 4 5 ...
 $ doe    : num  1965 1965 1965 1965 1965 ...
 $ dox    : num  1977 1977 1977 1977 1977 ...
 $ chd    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dob    : num  1915 1915 1915 1914 1914 ...
 $ job    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ month  : int  8 8 8 12 12 11 11 9 9 9 ...
 $ energy : num  2.87 2.87 2.87 1.98 1.98 ...
 $ height : num  175 175 175 164 164 ...
 $ weight : num  71.5 71.5 71.5 70.1 70.1 ...
 $ fat    : num  141.7 141.7 141.7 85.8 85.8 ...
 $ fibre  : num  17.83 17.83 17.83 9.49 9.49 ...
 $ agr    : Factor w/ 5 levels "(-Inf,40]","(40,50]",..: 2 3 4 3 4 3 4 3 4 3 ...
 $ eksp   : Factor w/ 2 levels "High","Low": 1 1 1 2 2 2 2 1 1 1 ...
 - attr(*, "breaks")=List of 2
  ..$ per: NULL
  ..$ age: num [1:4] 40 50 60 70
 - attr(*, "time.scales")= chr [1:2] "per" "age"
 - attr(*, "time.since")= chr [1:2] "" ""
```

   Then make a table like the one in C& H—`xtabs` adds up the l.h.s. argument by the variables in the r.h.s. argument in the formula:

```
> tt <- xtabs(cbind(D = (lex.Xst=="IHD"),
+                   Y = lex.dur)
+             ~ agr + eksp,
+             data = Ls)
> str(tt)
```

```
 'xtabs' num [1:5, 1:2, 1:2] 0 4 6 8 0 0 2 12 14 0 ...
 - attr(*, "dimnames")=List of 3
  ..$ agr : chr [1:5] "(-Inf,40]" "(40,50]" "(50,60]" "(60,70]" ...
  ..$ eksp: chr [1:2] "High" "Low"
  ..$     : chr [1:2] "D" "Y"
 - attr(*, "call")= language xtabs(formula = cbind(D = (lex.Xst == "IHD"), Y = lex.dur) ~
```

```
> round(ftable(tt, col.vars = 2:3), 2)

          eksp    High             Low
                     D       Y       D       Y
agr
(-Inf,40]         0.00   62.25    0.00   34.08
(40,50]           4.00  560.13    2.00  346.87
(50,60]           6.00 1127.70   12.00  979.34
(60,70]           8.00  794.15   14.00  699.14
(70,Inf]          0.00    0.00    0.00    0.00

> round(ftable(tt, col.vars = 3:2), 2)

                     D              Y
          eksp    High     Low    High     Low
agr
(-Inf,40]         0.00    0.00   62.25   34.08
(40,50]           4.00    2.00  560.13  346.87
(50,60]           6.00   12.00 1127.70  979.34
(60,70]           8.00   14.00  794.15  699.14
(70,Inf]          0.00    0.00    0.00    0.00
```

You should see that the data is not quite the same as in the book.

9. Do the grouped analysis on the slightly modified data that you can get from the data folder (which should be identical to the table you just made):

```
> ihdx <- read.table(
+     "http://BendixCarstensen.com/EpiPhD/F2020/ihd-xtab.txt", header=T )
> ihdx
  exposure age    pyrs cases
1        1   0  346.87     2
2        1   1  979.34    12
3        1   2  699.14    14
4        0   0  560.13     4
5        0   1 1127.70     6
6        0   2  794.15     8

> mt <- glm(cbind(cases, pyrs) ~ factor(age) + exposure,
+           family = poisreg,
+             data = ihdx)
> round(ci.exp(mt), 3 )

             exp(Est.)  2.5% 97.5%
(Intercept)      0.005 0.002 0.012
factor(age)1     1.226 0.486 3.092
factor(age)2     2.110 0.854 5.211
exposure         1.862 1.029 3.370
```

10. Estimate the effect of age and exposure from the split dataset. Remember to exclude follow-up time before age 40 — as you saw from the table above:

```
> Ls <- subset(Ls, agr %in% levels(agr)[2:4])
> Ls$agr <- factor(Ls$agr)
> table(Ls$agr)
```

```
(40,50] (50,60] (60,70]
    196     293     240
```

```
> head(Ls)
```

```
 lex.id     per   age lex.dur lex.Cst lex.Xst id      doe      dox chd      dob job month
      1 1964.62 49.62    0.38    Well    Well  1 1964.623 1976.916   0 1915.008   0     8
      1 1965.01 50.00   10.00    Well    Well  1 1964.623 1976.916   0 1915.008   0     8
      1 1975.01 60.00    1.91    Well    Well  1 1964.623 1976.916   0 1915.008   0     8
      2 1964.96 50.54    9.46    Well    Well  2 1964.957 1976.916   0 1914.419   0    12
      2 1974.42 60.00    2.50    Well    Well  2 1964.957 1976.916   0 1914.419   0    12
      3 1965.87 58.78    1.22    Well    Well  3 1965.874 1976.916   0 1907.090   0    11
 energy  height weight    fat fibre     agr eksp
  2.874 175.387 71.487 141.71 17.83 (40,50] High
  2.874 175.387 71.487 141.71 17.83 (50,60] High
  2.874 175.387 71.487 141.71 17.83 (60,70] High
  1.982 164.287 70.081  85.77  9.49 (50,60]  Low
  1.982 164.287 70.081  85.77  9.49 (60,70]  Low
  2.669 169.393 71.896 107.67 15.99 (50,60]  Low
```

```
> mi <- glm(cbind(lex.Xst=="IHD", lex.dur) ~ factor(agr) + eksp,
+           family = poisreg,
+             data = Ls)
> round(ci.exp(mi), 3 )
```

```
                   exp(Est.)  2.5% 97.5%
(Intercept)            0.005 0.002 0.012
factor(agr)(50,60]     1.226 0.486 3.092
factor(agr)(60,70]     2.110 0.854 5.211
eksplow                1.862 1.029 3.370
```

```
> round(ci.exp(mt), 3 )
```

```
             exp(Est.)  2.5% 97.5%
(Intercept)      0.005 0.002 0.012
factor(age)1     1.226 0.486 3.092
factor(age)2     2.110 0.854 5.211
exposure         1.862 1.029 3.370
```

We see that the estimates are identical for the two ways of modeling. The point of using the individual data is that individual-level variables could be included in a model too.

11. Add an interaction between age and exposure and check that you get the same test for interaction as with the grouped data.

```
> mix <- update(mi, . ~ . + factor(agr):eksp)
> mtx <- update(mt, . ~ . + factor(age):exposure)
> anova(mi, mix, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: cbind(lex.Xst == "IHD", lex.dur) ~ factor(agr) + eksp
Model 2: cbind(lex.Xst == "IHD", lex.dur) ~ factor(agr) + eksp + factor(agr):eksp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       725     313.18
2       723     311.98  2   1.2015   0.5484
```

```
> anova(mt, mtx, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: cbind(cases, pyrs) ~ factor(age) + exposure
Model 2: cbind(cases, pyrs) ~ factor(age) + exposure + factor(age):exposure
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2     1.2015
2         0     0.0000  2   1.2015   0.5484
```

12. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data. You can get the deviance from `deviance`:

```
> deviance(mt)
[1] 1.201497
```

13. Age is not a categorical variable, it is a quantitative variable, increasing continuously, not by jumps from one category to the next. But the theory we use is based om a model for constant rates, we used it for rates constant across 10-year intervals. That is a pretty bold assumption. If we instead had intervals of, say, 4 months length then the assumption would be tenable. But then we would face not 3 but 900 age categories!

    Therefore the model would not be a model with one parameter for each age category, but a model that use the attained age in each interval as a *quantitative* variable with a possibly non-linear effect.

    So split the follow-up further in small intervals:

```
> Ls <- splitLexis(Ls,
+              breaks = seq(40, 70, 1/3),
+          time.scale = "age")
> summary(Lx)

Transitions:
     To
From   Well IHD  Records:  Events: Risk time:  Persons:
  Well  291  46       337       46    4603.67       337

> summary(Ls)

Transitions:
     To
From    Well IHD  Records:  Events: Risk time:  Persons:
  Well 13751  46     13797       46    4507.34       337
```

    So we now have many intervals for each person, say nos 6 and 305:

```
> subset(Lx, lex.id %in% c(6,305))[,1:10]

 lex.id     per   age lex.dur lex.Cst lex.Xst  id      doe      dox chd
      6 1965.20 50.98    3.46    Well    Well   6 1965.203 1968.664   0
    305 1960.04 46.55    1.49    Well     IHD 305 1960.040 1961.535   1

> subset(Ls, lex.id %in% c(6,305))[,1:10]
```

```
 lex.id     per   age lex.dur lex.Cst lex.Xst  id      doe      dox chd
      6 1965.20 50.98    0.02    Well    Well   6 1965.203 1968.664   0
      6 1965.22 51.00    0.33    Well    Well   6 1965.203 1968.664   0
      6 1965.56 51.33    0.33    Well    Well   6 1965.203 1968.664   0
      6 1965.89 51.67    0.33    Well    Well   6 1965.203 1968.664   0
      6 1966.22 52.00    0.33    Well    Well   6 1965.203 1968.664   0
      6 1966.56 52.33    0.33    Well    Well   6 1965.203 1968.664   0
      6 1966.89 52.67    0.33    Well    Well   6 1965.203 1968.664   0
      6 1967.22 53.00    0.33    Well    Well   6 1965.203 1968.664   0
      6 1967.56 53.33    0.33    Well    Well   6 1965.203 1968.664   0
      6 1967.89 53.67    0.33    Well    Well   6 1965.203 1968.664   0
      6 1968.22 54.00    0.33    Well    Well   6 1965.203 1968.664   0
      6 1968.56 54.33    0.11    Well    Well   6 1965.203 1968.664   0
    305 1960.04 46.55    0.12    Well    Well 305 1960.040 1961.535   1
    305 1960.16 46.67    0.33    Well    Well 305 1960.040 1961.535   1
    305 1960.49 47.00    0.33    Well    Well 305 1960.040 1961.535   1
    305 1960.83 47.33    0.33    Well    Well 305 1960.040 1961.535   1
    305 1961.16 47.67    0.33    Well    Well 305 1960.040 1961.535   1
    305 1961.49 48.00    0.04    Well     IHD 305 1960.040 1961.535   1
```

14. Now fit a model with the effect of age as a natural spline:

```
> (akn <- seq(50, 65, , 4))
```

```
[1] 50 55 60 65
```

```
> ms <- glm(cbind(lex.Xst=="IHD", lex.dur)
+           ~ Ns(age, knots = akn) + eksp,
+           family = poisreg,
+             data = Ls)
> summary(ms)
```

```
Call:
glm(formula = cbind(lex.Xst == "IHD", lex.dur) ~ Ns(age, knots = akn) +
    eksp, family = poisreg, data = Ls)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1168  -0.0890  -0.0795  -0.0635   4.1340

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -5.2732     0.3128 -16.860   <2e-16 ***
Ns(age, knots = akn)1  0.5950     0.5471   1.088   0.2767
Ns(age, knots = akn)2  0.8094     0.4687   1.727   0.0842 .
Ns(age, knots = akn)3  0.6243     0.4069   1.534   0.1250
ekspLow                0.6140     0.3027   2.028   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 602.61  on 13796  degrees of freedom
Residual deviance: 593.65  on 13792  degrees of freedom
AIC: 695.65

Number of Fisher Scoring iterations: 7
```

15. Now predict the rates of IHD for persons in ages 40 through 70 and with `eksp` equal to either `Low` or `High`:

```
> ndl <- data.frame(age = 40:70, eksp = "Low")
> ndh <- data.frame(age = 40:70, eksp = "High")
> prl <- ci.pred(ms, ndl)
> prh <- ci.pred(ms, ndh)
> str(prl)
```

```
 num [1:31, 1:3] 0.00704 0.00726 0.00747 0.0077 0.00793 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:31] "1" "2" "3" "4" ...
  ..$ : chr [1:3] "Estimate" "2.5%" "97.5%"
```

We now have predicted rates as function of age for each of the two exposure groups, so we use `matshade` to plot the two:

```
> matshade(ndl$age, cbind(prl, prh) * 100, plot = TRUE,
+          lwd = 3, col = c("ForestGreen","orange"),
+          log = "y", xlab = "Age (years)",
+          ylab = "IHD rate per 100 PY")
```

From the graph we can see that this is a proportional hazards model

16. Now fit an interaction:

```
> msi <- glm(cbind(lex.Xst=="IHD", lex.dur)
+            ~ Ns(age, knots = akn) * eksp,
+            family = poisreg,
+              data = Ls)
> summary(msi)
```

```
Call:
glm(formula = cbind(lex.Xst == "IHD", lex.dur) ~ Ns(age, knots = akn) *
    eksp, family = poisreg, data = Ls)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1161  -0.0929  -0.0698  -0.0647   4.1003

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -5.0572     0.3651 -13.853   <2e-16 ***
Ns(age, knots = akn)1           0.0598     0.9134   0.065    0.948
Ns(age, knots = akn)2           0.2155     0.6186   0.348    0.728
Ns(age, knots = akn)3           0.4406     0.6519   0.676    0.499
ekspLow                         0.1034     0.5561   0.186    0.852
Ns(age, knots = akn)1:ekspLow   0.8636     1.1481   0.752    0.452
Ns(age, knots = akn)2:ekspLow   1.3274     1.0420   1.274    0.203
Ns(age, knots = akn)3:ekspLow   0.2730     0.8358   0.327    0.744
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 602.61  on 13796  degrees of freedom
```

```
Residual deviance: 591.68  on 13789  degrees of freedom
AIC: 699.68

Number of Fisher Scoring iterations: 7
```

We really do not need to bother about the parametrization, we can just use the same code for the graoph of the rates in the interaction model:

```
> prl <- ci.pred(msi, ndl)
> prh <- ci.pred(msi, ndh)
> matshade(ndl$age, cbind(prl, prh) * 100, plot = TRUE,
+          led = 3, col = c("ForestGreen","orange"),
+          log = "y", xlab = "Age (years)",
+          ylab = "IHD rate per 100 PY")
```

17. We can check whether there is an interaction ("non-proportionality"):

```
> anova(ms, msi, test = "Chisq")

Analysis of Deviance Table

Model 1: cbind(lex.Xst == "IHD", lex.dur) ~ Ns(age, knots = akn) + eksp
Model 2: cbind(lex.Xst == "IHD", lex.dur) ~ Ns(age, knots = akn) * eksp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     13792     593.65
2     13789     591.68  3   1.9696   0.5787
```

18. Esoteric: We can plot the predictions from the two models together:

```
> matshade(ndl$age, cbind(ci.pred(ms , ndl),
+                         ci.pred(ms , ndh),
+                         ci.pred(msi, ndl),
+                         ci.pred(msi, ndh)) * 100,
+          plot = TRUE,
+          lwd = 3, col = c("ForestGreen","orange"),
+          lty = rep(c("solid","22"), each = 2),
+          log = "y", xlab = "Age (years)",
+          ylab = "IHD rate per 100 PY")
```