Epidemiology course for PhD students

Department of Biostatistics
Institute of Public Health
Spring 2013
http://BendixCarstensen.com/EpiF2013
Version 3

Compiled Friday 17th May, 2013, 14:18
from: C:/Bendix/undervis/Epi.CH/Epi-PhD/pracs/pracs.tex

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

Elisabeth Wreford Andersen — Department of Informatics and Mathematical Modeling
DTU Data Analysis
ewan@imm.dtu.dk

Henrik Ravn — Division of Epidemiology, Statens Serum Institut
hjn@ssi.dk

Per Kragh Andersen — Department of Biostatistics,
Institute of Public Health, University of Copenhagen
pkan@sund.ku.dk

Contents

1 Introduction **1**
1.1 Preface . . . 1
1.2 Website . . . 1
1.3 Data . . . 1

2 Exercises **3**
2.1 Vaccinations and childhood mortality in Guinea-Bissau . . . 3
2.1.1 A single risk, odds and rate . . . 3
2.1.2 Rates, risks and odds . . . 4
2.1.3 Rate ratio, risk ratio, odds ratio . . . 4
2.1.4 Confounder control: stratified analysis of odds ratio and risk ratio. . . . 5
2.1.4.1 Computing hints . . . 5
2.1.5 Survival analysis of childhood mortality in Guinea-Bissau . . . 5
2.2 Case-control study of renal cancer and trichorehtene . . . 6
2.3 IHD data from Clayton & Hills. . . . 7
2.3.1 Using continuous variables . . . 7
2.3.2 Splitting the follow-up of the IHD data . . . 8
2.3.2.1 Using SAS . . . 8
2.3.2.2 Using Stata . . . 10
2.3.2.3 Using R . . . 12
2.4 Case-control study of BCG vaccination and leprosy. . . . 15
2.5 Case-control study of malignant melanoma. . . . 15
2.5.1 Discussion of the article. . . . 16
2.5.2 Melanoma data . . . 16
2.5.3 Simple tabulation analysis . . . 16
2.5.4 Simple analysis controlling for age . . . 17
2.5.5 Introductory analyses. . . . 17
2.5.6 Trend tests and interactions. . . . 18
2.6 Testicular cancer risk and maternal parity. . . . 19
2.6.1 Discussion of the article. . . . 19
2.6.2 Practical exercises . . . 20

References **21**

3 Solutions with SAS **1**
3.1 Vaccinations and childhood mortality in Guinea-Bissau . . . 1
3.1.1 A single risk, odds and rate . . . 2

3.1.2 A single risk, odds and rate . . . 2
3.1.3 Rates, risks and odds . . . 3
3.1.4 Rate ratio, risk ratio, odds ratio . . . 4
3.1.5 Confounder control: stratified analysis of odds ratio and risk ratio. . . 6
3.1.6 Survival analysis of childhood mortality in Guinea-Bissau . . . 7
3.2 Case-control study of renal cancer and trichloroethene . . . 9
3.3 IHD data from Clayton & Hills. . . . 13
3.3.1 Using continuous variables . . . 14
3.3.2 Splitting the follow-up of the IHD-data . . . 16
3.4 Case-control study of BCG vaccination and leprosy. . . . 17
3.5 Case-control study of malignant melanoma. . . . 20
3.5.1 Discussion of the article. . . . 20
3.5.2 Melanoma data . . . 21
3.5.3 Simple tabulation analysis . . . 21
3.5.4 Introductory analyses. . . . 23
3.5.5 Trend tests and interactions. . . . 24
3.6 Testicular cancer risk and maternal parity. . . . 29
3.6.1 Practical exercises . . . 29

4 Solutions with Stata **1**
4.1 Vaccinations and childhood mortality in Guinea-Bissau . . . 1
4.1.1 A single risk, odds and rate . . . 2
4.1.2 A single risk, odds and rate . . . 2
4.1.3 Rates, risks and odds . . . 3
4.1.4 Rate ratio, risk ratio, odds ratio . . . 3
4.1.5 Confounder control: stratified analysis of odds ratio and risk ratio. . . 4
4.1.6 Survival analysis of childhood mortality in Guinea-Bissau . . . 4
4.2 Case-control study of renal cancer and trichloroethene . . . 5
4.3 IHD data from Clayton & Hills. . . . 8
4.3.1 Using continuous variables . . . 8
4.3.2 Splitting the follow-up of the IHD-data . . . 9
4.4 Case-control study of BCG vaccination and leprosy. . . . 10
4.5 Case-control study of malignant melanoma. . . . 11
4.5.1 Discussion of the article. . . . 11
4.5.2 Melanoma data . . . 12
4.5.3 Simple tabulation analysis . . . 12
4.5.4 Introductory analyses. . . . 13
4.5.5 Trend tests and interactions. . . . 13
4.6 Testicular cancer risk and maternal parity. . . . 16
4.6.1 Discussion of the article. . . . 16
4.6.2 Practical exercises . . . 18

5 Solutions with R **1**
5.1 Vaccinations and childhood mortality in Guinea-Bissau . . . 1
5.1.1 A single risk, odds and rate . . . 2
5.1.2 A single risk, odds and rate . . . 2
5.1.3 Rates, risks and odds . . . 6

5.1.4 Rate ratio, risk ratio, odds ratio . . . 11
5.1.5 Confounder control: stratified analysis of odds ratio and risk ratio . . 17
5.1.6 Survival analysis of childhood mortality in Guinea-Bissau . . . 22
5.2 Case-control study of renal cancer and trichloroethene . . . 27
5.3 IHD data from Clayton & Hills. . . . 32
5.3.1 Using continuous variables . . . 33
5.3.2 Splitting the follow-up of the IHD-data . . . 33
5.4 Case-control study of BCG vaccination and leprosy. . . . 41
5.5 Case-control study of malignant melanoma. . . . 44
5.5.1 Discussion of the article. . . . 44
5.5.2 Melanoma data . . . 45
5.5.3 Simple tabulation analysis . . . 45
5.5.4 Introductory analyses. . . . 51
5.5.5 Trend tests and interactions. . . . 54
5.6 Testicular cancer risk and maternal parity. . . . 61
5.6.1 Discussion of the article. . . . 61
5.6.2 The Lexis diagram . . . 63
5.6.3 Practical exercises . . . 64

Chapter 1

Introduction

1.1 Preface

This is the collection of exercises for the course in epidemiology for PhD-students in the spring of 2013.

The exercises are based on students using SAS as the computer program for solving the exercises, and the weight of the recap of the exercises during the course will be on SAS too.

Students are however welcome to use other software packages, provided that they bring them on their own computer, and can access the datasets to be used from the net. Some of the teachers will have some expertise in some of the other frequently used computer packages such as Stata, R and (limited) SPSS. Most of the teachers have experience using R, some have experience in Stata, whereas none of the teachers use SPSS.

1.2 Website

The course website for the statistics practicals is
http://BendixCarstensen.com/EpiF2013.

There will be links to this document, to the data, to the programs mentioned in this document and to solutions to the practicals as the course proceeds.

Whenever we refer to "from www" it means that you should go to the course website or the data website.

1.3 Data

The datasets are all found in the data folders http://192.38.117.59/~pka/epidata/ resp. http://192.38.117.59/~pka/spss-stata-data/

There is also a link to these at the course website:
http://BendixCarstensen.com/EpiF2013/data.

Description of the datasets are in the exercise texts. The detailed scheme for the course is in www.biostat.ku.dk/~nk/epiF13

Here is the scheme for the exercises to be used each of the days of the course where the are practical exercises.

Date	Subject	Exercise
5 Mar	No practicals	
12 Mar	Introduction to SAS; practicals on C&H: ch. 1, 2, 5, 13, 17)	2.1.1 2.1.2, 2.1.3
19 Mar	No practicals	
26 Mar	No teaching (Easter)	
2 Apr	Case-control studies:	
	Renal cancer and chemical exposure	2.2
	Melanoma	2.5.3
9 Apr	No practicals	
16 Apr	Confounder control	2.1.4, 2.5.4
23 Apr	No practicals	
30 Apr	Starting regression analysis	2.3, 2.4
7 May	Handling continuous explanatory variables	2.3.1
13 May	Survival analysis, comparing with other types	3.1.6
14 May	Analysis of a real Case-control study	2.5.1, 2.5.5, 2.5.6
15 May	Follow-up studies from individual records	2.3.2, 2.6.1, 2.6.2

The section numbers above are only the exercises, you will need to read the general introduction to the relevant datasets before you can do the exercises.

Chapter 2

Exercises

2.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

2.1.1 A single risk, odds and rate

Tabulate the number of children, the number of deaths and the number of person-years.

- Do the following by using the formulae from the lectures:

  1. What is the overall risk of death? Make a confidence inetrval for this proportion.
  2. What is the overall odds of death? Make a confidence interval for this odds.
  3. What is the overall *rate* of daeth (per year). Make a confidence interval for this rate.

- Do the same by using your statistical package. Do you get the same confidence intervals?

2.1.2 Rates, risks and odds

First, make a table of the number of children, the number of deaths and the number of person-years by BCG vaccination status.

- Based on this do the following calculations by hand (or a suitable program on your computer), by inserting the numbers in the formulae from the lectures:

  1. Estimate the 6-month *risk* of death for children with or without BCG vaccination (SAS users may use the program `bissau.sas`).
  2. Compute 95% confidence limits for the two risk parameters.
  3. Estimate the 6-month *odds* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the odds parameters. Compare with the risk parameters.
  4. Estimate (by hand or using SAS or another piece of software) the *rate (per day)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.
  5. Estimate the *rate (per year)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.
  6. Create a new binary variable indicating whether or not the child was DTP vaccinated at first visit and repeat the previous questions for this DTP variable.

- Calculations using a statistical model

  1. Compute the risk with 95% confidence intervals in each of the two groups. You must fit a binomial model (without intercept) with log-link and exponentiate the estimates afterwards.
  2. Compute the odds of death 95% confidence intervals in each of the two groups. You must fit a binomial model (without intercept) with logit link and exponentiate the estimates afterwards.
  3. Compute the rate of death per *year* in each of the two groups. You must fit a poisson model (without intercept) with log link and the log-person-years as offset and exponentiate the estimates afterwards.
  4. Do the same for the subdivision of data by DTP.

2.1.3 Rate ratio, risk ratio, odds ratio

Continuing from before, calculate relative effects of BCG and DTP on mortality.

1. Calculate (SAS-users may use `proc freq`) the risk ratio and odds ratio and 95% confidence interval (CI) for the effect of BCG on mortality, i.e., compare the risk/odds of dying among BCG-vaccinated vs. BCG-unvaccinated. What do you conclude?
2. Do the same for DTP (any dose, i.e. as a binary exposure). What do you conclude?
3. Test the association between BCG and DTP-any dose using a Chi-square test. In this mortality is not involved, only test wheter the occurrence of the two types of vaccination are related. What do you conclude?

4. Estimate the DTP effect (risk ratio and odds ratio) separately for each level of BCG. What happened?

5. Until now we have not accounted for the follow-up time. Repeat question 1, 2, and 4 but now by calculating the rate ratio and 95% CI for the BCG and DTP exposure.

2.1.4 Confounder control: stratified analysis of odds ratio and risk ratio.

We will continue using the data from Guinea-Bissau for practical exercises.

1. Revisit the analysis from week 2 (Q4), using just death (dead) as outcome, and estimate the DTP effect for each level of BCG.

2. Use the BCG as a potentially confounding variable and obtain the MH-estimate for the OR and RR. What are they?

3. Do the same, using age in months (`agem`) as control variable in the analysis. Is there any DTP effect?

4. Do the same, but now using both `agem` and `bcg` (that is, the cross-classification) as control variables in the analysis. Is there any DTP effect?

2.1.4.1 Computing hints

In SAS you can make an analysis controlling for confounding by including the confounder variable before the exposure and outcome variables in the table statement, and adding `cmh` as option (`cmh` = Cochran-Mantel-Haenzsel):

```
proc freq data = bissau ;
table agem * dtpany * dead / norow nocol nopct cmh ;
run ;
```

2.1.5 Survival analysis of childhood mortality in Guinea-Bissau

The `SAS` program `bissau.sas` reads the data from `www` and fits a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

1. Fit a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates and re-find estimates from today's lecture.

2. Estimate the effect of any dose of DTP, using the created variable `dtpany` adjusted only for age in months as a categorical (`class` in `SAS`) variable.

3. Now, also adjust for BCG. What happened? Can you explain?

4. Is there an interaction between DTP (`dtpany`) and BCG?

5. Make a Cox regression analysis with DTP (`dtpany`) and BCG, but now with age as time-variable, i.e. with delayed entry.

6. Repeat the Poisson and logistic regression models that you have seen during the lectures, and compare the results:

Cox RR (95%CI)	Poisson RR (95%CI)	Logistic OR (95%CI)
0.71 (0.53-0.94)	0.71 (0.53-0.94)	0.71 (0.53-0.96)

   All models should be *adjusted for age in months as a categorical variable*. In the Cox model, follow-up time was used as the time-variable. In the Poisson model, the follow-up time was used as time at risk. The logistic regression did not take the follow-up time into account.

2.2 Case-control study of renal cancer and trichorehtene

This exercise is nased on; Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichlorethe. J Cancer Res Clin Oncol, 1998, pp 374–382. [1].

The paper is available at the course homepage as
http://BendixCarstensen.com/EpiE2012/Vamvakas.1998.pdf

We will discuss the following points based on the paper:

1. What is the primary aim of the study?
2. How were the cases sampled?
3. How were the controls sampled?
4. Are they comparable; i.e. what assumptions are needed?
5. What is the (actual) study base?
6. What study base is the intended (for generalization)?
7. Is the sampling scheme incidence density sampling?
8. Can the age-effect on the occurrence of renal cancer be estimated?
9. Is age a confounder?
10. Key in the numbers in table 6 (p.380), and verify the analysis using SAS `proc freq`.
11. Is there any evidence of heterogeneity of the odds-ratio across age-classes? (*Hint:* Use the Breslow-Day-test.)
12. In particular, how does the odds-ratio estimate given by Vamvakas *et al.* compare the the Mantel-Haenszel estimate based on the same data?
13. What is the main result (in plain words)?

2.3 IHD data from Clayton & Hills.

The study is described by Clayton & Hills, Ch. 13. The tabulated data set of counts of IHD cases and person-years is avaliable from `www` in the file `ihd-tab.txt`.

The `SAS` program `ihd-reg.sas` reads the data from `www` and fits a Poisson regression model without interaction between age and exposure.

1. Fit the model from Clayton & Hills Tables 22.7-8 (p.222) and perform the tests from exercises 24.1 and 24.2 (pp.237–238). `SAS`-users may use the program `ihd-reg.sas` and notice the use of the `ESTIMATE` command to obtain a given reference group and the rate ratios with 95% confidence intervals.

2. Fit the model with interaction and re-find results from Clayton & Hills Table 24.5 (p.242) and the test for no interaction.

2.3.1 Using continuous variables

The IHD-data contains enegy consumption as a continuous variable. The dataset `diet.txt` has the following variables:

`id`	Person id
`doe`	Date of entry
`dox`	Date of exit
`chd`	CHD-status at exit: 0-no, 1-yes
`dob`	Date of birth
`job`	Not used
`month`	Notudes
`energy`	Daily energy intake in MJ
`height`	Height in cm
`weight`	Weight in cm
`fat`	Daily fat intake (g)
`fibre`	Daily fibre intake (g)

1. Read the individual diet data records from the file.

2. Create variables for the person-years, by subtracting entry date from date of exit. Also create a variable with the log-person-years.

3. Use CHD as outcome variable in a Poisson-analysis with the log-person-years as offset, using energy as a linear explanatory variable. Is there an effect on mortality?

4. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (approx. the quartiles)

5. Same question for weight and BMI (the latter you have to calculate yourself as weight/height^2).

2.3.2 Splitting the follow-up of the IHD data

The following exercise is designed to illustrate how follow-up time is subdivided in order to produce the table of events and person-years. Furthermore the aim is to show you that tabulated data and time-split data gives the same results if only age and exposure are used as variables.

We will first analyze frequency records as above (these are almost identical to Table 22.6 in C & H). Next, we shall read the individual records and construct the corresponding table of cases and person-years.

The splitting of follow-up along a timescale is quite a technical task, which is handeled somewhat differently in SAS, Stata and R, so the exercise is here given in three different versions, one for each programming language.

2.3.2.1 Using SAS

1. Import the program `ihd-lexis.sas` to the program editor. Run the first part of the program — the part reading the tabulated data and `proc genmod`. Compare with the results from table 24.1 in Clayton & Hills.

2. Next, read the individual records from the file `diet.txt`, including the `proc print` and check on the output that it looks reasonable and that you understand what the data represents.

3. Now you should import the macro `%Lexis` and use it to split into the age intervals 40–50, 50–60 and 60–70 years:

   In order to use this you must first load it from the appropriate folder folder on the net:

   ```
   options source2 ;  * List the included code in the log-window ;
   filename lexispr url "http://BendixCarstensen.com/Lexis/Lexis.sas";
   %inc lexispr ;
   ```

   Once you have specified `%inc lexispr ;` and run that line in SAS, SAS will know the macro `%lexis` and you can use it in the rest of the session.

4. The time-splitting is now done by running the SAS-macro `%Lexis`[1]. A SAS-macro is a piece of SAS-program (normally quite long) where certain small parts of the program can be changed when the program is run. The SAS-convention is that names of such programs start with a "`%`".

   To use the macro we must specify the follow-information from the input file:

   - Date of entry into the study — `doe`
   - Date of exit from the study — `dox`
   - Status at exit from the study — `chd` (1 if CHD occurred at `dox`, 0 otherwise).

[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.

Moreover, we must decide 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. To this end we must specify:

- The origin of the time-scale, i.e. where the time-scale is 0, in this case date of birth — `dob`.
- The intervals where we want the follow-up grouped, here ages 40–50, 50–60 and 60–70.
- As a purely technical thing we need to specify the conversion between the scale in which time is measured in the input dataset (in this case days) and in the specification of the grouping (in this case years) — 365.25.

In the case of `%Lexis` we must supply these 6 parameters in order to specify how to split time.

Finally we must tell the program where the the original data is, where the time-split data has to go, and what the name of the age-variable must be.

This looks like this (you do not have to write the stuff between the `/*...*/`):

```
%Lexis( data   = ihdindiv,        /* Dataset with original data       */
        out    = ihdsplit,        /* Dataset with time-split data     */
        entry  = doe,             /* Date of entry                    */
        exit   = dox,             /* Date of exit                     */
        fail   = chd,             /* Event (failure) indicator        */
        origin = dob,             /* Origin of the time-scale         */
        scale  = 365.25,          /* Conversion from input scale to breaks-scale
        breaks = 40 to 70 by 10,  /* Where to split the time scale    */
        left   = agr );           /* The name of the new age-variable */
```

Run this piece of SAS code.

(In the top of the file `http://BendixCarstensen.com/Lexis/Lexis.sas` are some more detailed explanations of how to use `%Lexis`).

5. How many records are in the resulting dataset (`ihdsplit`)

6. Take at look at the resulting data file, for example the first 20 records:

```
proc print  data = ihdsplit (obs=20) ;
run ;
```

How does this compare with the the original dataset?

7. Use `%PYtab` to tabulate ihd-cases and person-years by exposure and age-group. You must first get this from the net as you did with the `%Lexis` macro:

```
filename pytabpr url "http://BendixCarstensen.com/Lexis/PYtab.sas";
%inc pytabpr ;
```

```
%PYtab( data  = ihdsplit,
        class = exposure agr,
        fail  = chd,
        risk  = risk,
        scale = 1000 ) ;
```

Compare with the sums from the table given in the first data step in `ihd-lexis.sas`

8. Use `proc genmod` to estimate the effect of age and exposure from the split dataset. How does the estimates compare with those based on the initial tabulated dataset?

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

10. Compare the type 3 likelihood ratio statistic (`Chi-square`) for the interaction with the deviance of the model without interaction for the grouped data.

2.3.2.2 Using Stata

1. First we enter the grouped data and make a simple Poisson analysis:

```
input eksp agr pyrs cases
1 0  346.87  2
1 1  979.34 12
1 2  699.14 14
0 0  560.13  4
0 1 1127.70  6
0 2  794.15  8
end

poisson cases i.eksp i.agr , exposure(pyrs)
```

2. Then read the individual data, convert to date variables:

```
infile id str10 doe str10 dox chd str10 dob job month energy height weight fat f
*/ in 2/L using "http://BendixCarstensen.com/EpiE2012/data/diet.txt", clear

*Get the dates into date format

gen date_entry = date(doe,"MDY")
gen date_exit  = date(dox,"MDY")
gen date_birth = date(dob,"MDY")
format date_entry date_exit date_birth %td
```

3. Now tell Stata that this is survival data, that is, when persons enter, exit and whether they are dead or not at exit (`fail`), and finally which scale we are on (`origin`):

```
stset date_exit, failure(chd==1) entry(date_entry) origin(date_birth) /*
*/ scale(365.25) id(id)

display _N
```

Note that Stata generated 4 new variables `_t0`, `_t`, `_d` and `_st`, describing the survival. Read the help page for `stset` and make sure you understand what they mean. (A useful introduction to `stset` is www.pauldickman.com/survival/stset.pdf).

4. Then split the data into age groups 40–50, 50–60, 60–70 and generate a new variable called `current_age`:

```
stsplit current_age, at(40(10)70) after(date_birth)

* How many observations?
display _N
```

5. Now take a look at the data:

```
list in 1/10
browse
```

6. Tabulate IHD cases and person-years by exposure and age group. To this end we use the sytem variables `_t0` and `_t` which hold the left and the right end-points on the "analysis time scale", in this case the current age:

```
gen pyrs=_t-_t0
gen exposure = (energy < 2.75) + energy-energy

* Only count CHD cases once
gen event=_d

table current_age exposure, c(sum event sum pyrs) format(%9.2f)
```

Note that `current_age` ids 0 for all follow up before age 40 (left of first cutpoint).

7. Now use `poisson` (or `glm`) to estimate the effect of age and exposure from the split dataset. How does the estimates compare with those based on the initial tabulated dataset?

```
* drop follow-up before age 40
keep if current_age>0

poisson event i.exposure i.current_age, exposure(pyrs)
```

— the same result as with the tabulated data.

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

```
poisson event i.exposure i.current_age i.exposure#i.current_age, exposure(pyrs)
testparm i.exposure#i.current_age
est store m1

poisson event i.exposure i.current_age , exposure(pyrs)
est store m2
lrtest m1 m2
```

9. Compare the type 3 likelihood ratio statistic for the interaction with the deviance of the model without interaction for the grouped data.

```
collapse (sum) pyrs event , by(exposure current_age)

poisson event i.exposure i.current_age , exposure(pyrs)
```

2.3.2.3 Using R

The following instructions are fairly detailed. You should make sure that you know what goes on, and that consult the help-pages fro 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/EpiE2012/data"

```
> options( width=90 )
> library( Epi )
> library( foreign )
> ihdt <- read.table("http://BendixCarstensen.com/EpiE2012/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.lin( mt, E=T ), 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( "../data/diet.txt",
+                     # "http://www.biostat.ku.dk/~pka/epidata/diet.txt",
+                     header=TRUE, na.strings=".", as.is=TRUE )
> 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[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, `per`iod) 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. Id not `entry.status` is given, 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 despeartely exciting but capturing the essence:

```
> boxes( Lx, boxpos=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 names `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( Ls )
> head( Ls )
```

For the fun of it you can try the default `plot` and `points` methods for a `Lexis` object. Note that gridlines 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 )
```

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. We now 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:

[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.

```
> 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.

Now we 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/EpiE2012/data/ihd-xtab.txt", header=T )
> ihdx
> mt <- glm( cases ~ factor(age) + exposure,
+            offset = log(pyrs), family=poisson, data=ihdx )
> round( ci.lin( mt, E=T ), 3 )
```

8. 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( (lex.Xst=="IHD") ~ factor(agr) + eksp,
+            offset = log(lex.dur), family=poisson, data=Ls )
> round( ci.lin( mi, E=T ), 3 )
> round( ci.lin( mt, E=T ), 3 )
> ci.lin( mi, E=T ) / ci.lin( mt, E=T )
```

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.

9. 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" )
```

10. 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 the `summary`:

```
> summary( mt )
```

2.4 Case-control study of BCG vaccination and leprosy.

The study is decribed by Clayton & Hills, p.156. In short, 260 cases of leprosy among individuals aged less than 35 years were ascertained in a study area in Malawi. Subjects were grouped into 7 age intervals and according to absence or presence of a scar after BCG vaccination. Three sets of controls were studied:

1. a population survey of 80,622 persons
2. a random sample of 1000 persons
3. a 4 to 1 age-matched sample

The file `bcgalldata.txt` includes data from this study: for each of the 14 `age` by `scar` combinations, a text variable `status` indicates the type of person in question (`case, conall, con1000, conmatch`) and the numerical variable `n` the number of such persons.

The `SAS` program `bcg-reg.sas` reads these data and fits a logistic regression model with no interaction between `age` and `scar` using all cases and all controls.

1. Fit the model from Clayton & Hills Table 23.5 (p.232). `SAS`-users may use the program `bcg-reg.sas`; what are the reference groups?
2. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference groups (in `SAS`: use '`ESTIMATE`' statements).
3. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference.
4. Test the hypothesis of no interaction between `age` and `scar`.
5. Analyse the data set with only 1000 controls (i.e., use the controls `con1000`: Table 23.6, p.233) and compare the precision of the estimate for `scar` with that based on the entire sample.
6. Analyse the matched data set (i.e., use the controls `conmatch`: Table 23.6, p.233) and compare with the results from Table 23.7.
7. Try (erroneously) to drop `age` from the analysis of the matched data and study the consequences for the estimate of `scar`.

2.5 Case-control study of malignant melanoma.

Anne Østerlind conducted in the middle of the 80's a case-control study of risk factors for malignant melanoma in Denmark.

The review paper "Malignant melanoma in Denmark" from *Acta Oncologica*, 1990 ,[3] is from Anne Østerlind's thesis and gives an overview of the results from the study which included 1400 interviewed persons, 474 cases and 926 controls, cf. table 5 in the article.

In the article incidence changes between 1943 and 1982 are also discussed; that part of the paper will not be touched upon in this exercise.

2.5.1 Discussion of the article.

1. Explain the design, the data base and data collection, particularly how the matching was conducted.

2. How were interviews planned to minimize bias?

3. Explain the drop-out, particularly the analyses in Tables 5-7. What are the consequenses of these results for the subsequent analyses?

4. How are the analyses carried out? Are all variables included in one step or are the analyses conducted in smaller steps? How are the matching variables accounted for? Comments?

5. Explain the analyses presented in Table 9. How many logistic regression models are fitted here?

6. What is the conclusion from the analyses in the table?

7. What is the purpose of Table 11?

8. Which modifiable factors seem to affect the melanoma risk?

2.5.2 Melanoma data

We have access to a subset of the variables from the study. These are found in the file `melanom.txt`. The variables are described in the table below. Based on these data, results from AØ's Tables 9 and 10 can (almost) be reconstructed. Revised versions of those two tables are also found below.

The `SAS` program `melanom.sas` reads the data from `www` and fits a simple logistic regression model including only the variable `skin`.

2.5.3 Simple tabulation analysis

1. Make the two by two table showing the association between case-control status and whether or not the person experienced *any* sunburns before the age of 15. `SAS`-users may use the program `melanom.sas` to read in the data from `www`. Estimate the odds ratio with associated 95% confidence limits and test for no association between the risk factor and case-control status.

2. Conduct similar analyses for the factors `sex, hair, eyes, freckles, acuterea, chronrea`. Compare with Table 9 in the article.

3. The case control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables. Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

4. Same question for age.

Table 2.1: *Variables in the melanoma data set. Some variables have missing values for some of the persons, these are coded ".". In the file there is one line for each person in the study. Data are found in the file* `melanom.txt`.

`casecon`	case-control status: 1:case, 0:control
`sex`	1:man, 2:woman
`ageint`	age at interview in years
`agroup`	grouped age: 10:10–19, 20:20–29, ...
`skin`	skin colour: 0:dark, 1:medium, 2:light
`hair`	hair colour: 0:dark brown/black, 1:light brown, 2:blond, 3:red
`eyes`	eye colour: 0:brown, 1:grey/green, 2:blue
`freckles`	freckles: 1:many, 2:some, 3:none
`acuterea`	acute reaction to sunlight: 1:blisters, 2:painful sunburn, 3:mild sunburn, 4:no sunburn
`chronrea`	chronic reaction to sunlight 1:deep tan, 2:moderate tan, 3:mild tan, 4:no tan
`nvsmall`	number of naevi < 5mm
`nvlarge`	number of naevi ≥ 5mm
`nvtot`	total number of naevi
`burn15`	number of sunburns before age 15

2.5.4 Simple analysis controlling for age

1. The case-control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables.

   Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

2. Study how this association is affected by adjustment for age.

3. Study how this association is affected by adjustment for *both* age and sex.

2.5.5 Introductory analyses.

1. Estimate (log-)odds ratios for the variable `skin` (see top left in AØ's Table 9). `SAS`-users may use the program `melanom.sas`.

2. Estimate also odds ratios (in `SAS`: use `ESTIMATE` statements).

3. Conduct the other analyses in AØ's Table 9 (*left* part) where the factors `hair, eyes, freckles, acuterea, chronrea` are studied one at a time.

4. Conduct the analysis corresponding to Table 9 (*right* part) where several variables are included simultaneously (see the table footnote).

5. Reconstruct the results from AØ's Table 10 concerning number of raised naevi.

Table 2.2: *Corrected Table 9. from the paper*

Factor	Category	OR (crude)	OR (adjusted)
Skin colour	Dark	(1.0)	(1.0)
	Medium	1.4 (1.0-1.9)	**1.3 (1.0-1.8)**
	Light	1.7 (1.2-2.3)	1.3 (0.9-**1.9**)
	trend test	$p < 0.01$	$p =$**0.15**
Hair colour	Dark-brown/black	(1.0)	(1.0)
	Light-brown	1.5 (1.2-1.9)	1.5 (1.2-1.9)
	Blond/fair	1.7 (1.0-2.9)	**1.6 (0.9-2.8)**
	Red	**1.7** (1.1-2.7)	1.3 (0.8-2.0)
	trend test	$p < 0.001$	$p = 0.04$
Eye colour	Brown	(1.0)	(1.0)
	Grey/green	0.9 (0.6-1.2)	0.7 (0.5-1.1)
	Blue	1.1 (0.8-1.5)	0.9 (0.6-1.3)
	trend test	$p =$**0.32**	$p =$**0.98**
Freckles	None	(1.0)	(1.0)
	Some	1.5 (1.2-1.9)	1.5 (1.2-2.0)
	Many	3.0 (2.2-4.1)	**3.0** (2.1-4.1)
	trend test	$p < 0.001$	$p < 0.001$
Acute reaction to sunlight	No sunburn	(1.0)	(1.0)
	Mild sunburn	1.3 (1.0-1.6)	1.1 (0.8-1.4)
	Painful sunburn	1.6 (**1.0**-2.6)	1.3 (0.8-2.1)
	Blisters	2.2 (0.9-5.0)	1.6 (0.7-3.9)
	trend test	$p =$**0.005**	$p =$**0.15**
Chronic reaction to sunlight	Deep tan	(1.0)	(1.0)
	Moderate tan	1.4 (1.1-1.8)	1.2 (0.9-1.6)
	Mild tan	1.8 (1.3-2.6)	1.4 (1.0-2.1)
	No tan	2.0 (1.0-3.7)	1.2 (0.6-2.5)
	trend test	$p < 0.001$	$p =$**0.10**

NB: new variables must be defined from the original variables `nvtot, nvsmall, nvlarge`.

2.5.6 Trend tests and interactions.

6. In the analyses so far all variables have been considered as categorical ('`class`' in `SAS`) variables while all tests in Tables 9 and 10 are trend tests. Conduct the analyses which give the P-values in Table 9 (right part) for the variables `skin` and `freckles`.

7. May `freckles` be scored linearly (1, 2, 3), when this variable is studied separately? (Conduct a test for linearity/departures from trend).

8. In AØ's Table 11 `freckles` and the total number of naevi (suitably grouped) are

Table 2.3: *Corrected Table 10.*

Factor	Category	OR (crude)	OR (adjusted)
Number of raised naevi on arms, total	None	(1.0)	(1.0)
	1	1.5 (1.1-2.1)	**1.5 (1.1-2.0)**
	2-4	2.3 (1.6-3.1)	2.2 (1.6-3.1)
	5+	5.4 (3.5-8.1)	**4.9 (3.2-7.5)**
	trend test	$p < 0.001$	
Number of raised naevi on arms, < 5 mm (diameter)	None	(1.0)	(1.0)
	1	1.6 (1.1-2.2)	1.6 (1.1-**2.2**)
	2-4	2.5 (1.8-3.4)	**2.4 (1.7-3.4)**
	5+	5.0 (3.3-7.7)	**4.7 (3.0-7.4)**
	trend test	$p < 0.001$	
Number of raised naevi on arms, $\geq$ 5mm (diameter)	None	(1.0)	(1.0)
	1	1.8 (1.2-2.8)	**1.6 (1.1-2.5)**
	2+	3.6 (1.8-7.2)	**2.7 (1.3-5.5)**
	trend test	$p < 0.001$	

studied. Conduct this analysis. Is there any interaction between these two variables?

9. Study, in a similar vein, interactions between `acuterea` and `skin` and between the grouped version of `nnvtot` from question 5. and `agroup`.

10. All of AØ's analyses are conducted without accounting for the match variable age (`agroup`) (in spite of warnings given by Clayton & Hills!). Repeat some of the previous analyses adjusting for `agroup`. Are there any substantial differences? Explain!

2.6 Testicular cancer risk and maternal parity.

This exercise deals with the article "Testicular cancer risk and maternal parity: a population-based cohort study", by T. Westergaard, P.K. Andersen, J.B. Pedersen, M. Frisch, J.H. Olsen, M. Melbye. *Br. J. Cancer*, **77**,pp. 1180-1185 (1998). [4].

2.6.1 Discussion of the article.

1. What is the authors' argument for the existence of an effect of maternal parity on the risk of testicular cancer in the son?

2. Describe the design of the study:

   - a. which "sons" are included in the study?
   - b. when are they followed?
   - c. how are cases defined and ascertained?

3. Concentrating on all testicular cancers, what do you consider to be the main result reported in Table 1?

4. Explain in words the interpretation of the value RR=0.80 for parity 2+.

5. Compare this value with the corresponding crude RR (and 95 % CI) obtained without any adjustment. Explain the differences between the two results.

6. Draw a Lexis diagram to illustrate the combinations of age and calendar period which contribute person-years to the study. An empty diagram is available as http://BendixCarstensen.com/EpiE2012/blank-Lexis.pdf

7. Explain the meaning of the estimates for "Interval from ..." in the lower part of Table 1.

8. What type of analysis is reported in Table 2?

9. Discuss how, alternatively, a case-control design could have been conducted to address the same question as the cohort study reported in the article.

2.6.2 Practical exercises

The file `testis.txt`, available at `www` contains for each (non-empty) combination of the factors `SON_AGE, SON_KOH, MOTH_AGE, PARITY` the number of person-years at risk `PYRS`, the numbers of non-seminomas and seminomas, respectively `NONSEMI SEMI`, and the total number of testis cancer cases `CASES`. The first line of the file contains the variable names.

The `SAS` program `testis.sas` reads the data from `www`.

10. Compute the crude rate ratio for testis cancer for parity 2+ versus parity 1. Compare with 5. above. `SAS`-users may use the SAS program `testis.sas` (and `PROC GENMOD`).

11. Reconstruct the estimates for "parity of mother at birth of son" from the top of Table 1 in the article both for all testis cancers and for non-seminomas.

12. Reconstruct the estimates from Table 2 in the article concerning mother's age (for all testis cancers). Is there an interaction between parity and mother's age?

13. Same question for birth cohort of the son.

References

[1] Vamvakas et al. Renal cell cancer correlated with occupational exposure to trichlorethene. *J Cancer Res Clin Oncol*, pages 374–382, 1998.

[2] I. Kristensen, P. Aaby, and H. Jensen. Routine vaccinations and child survival: follow up study in Guinea-Bissau, West Africa. *BMJ*, 321(7274):1435–1438, Dec 2000.

[3] A. Østerlind. Malignant melanoma in Denmark. Occurrence and risk factors. *Acta Oncol*, 29(7):833–854, 1990.

[4] T. Westergaard, P. K. Andersen, J. B. Pedersen, M. Frisch, J. H. Olsen, and M. Melbye. Testicular cancer risk and maternal parity: a population-based cohort study. *Br. J. Cancer*, 77(7):1180–1185, Apr 1998.

Chapter 3

Solutions with SAS

The SAS-programs are available on the course web site in the folder
http://BendixCarstensen.com/EpiF2013/sas. There is also a link to this on the website.

SAS

SAS is the default programming language in this course. It is a big package with many capabilities, but a bit clumsy for simple calculations. It is expensive, but PhD-students can have a free copy through KU, but only for the duration of your PhD.

The output from SAS comes in two different files, which makes it difficult to make a safe documentation of results, and always makes the documentation hard to follow because two different files have to be read in parallel. The solutions here consists only of the program code, not of listinngs of the `.log` (log window) and `.lst` (ouptput window) as this would be too extensive.

3.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

The following SAS-programs does all the calculations required in the exercises.

3.1.1 A single risk, odds and rate

3.1.2 A single risk, odds and rate

Tabulate the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naïve 95% confidence inerval for this is:

     $$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

     but a better one is the formula:

     $$\frac{p}{p + (1-p) \overset{\times}{\div} \text{erf}}, \quad \text{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\text{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

  2. The overall odds of death is simply:

     $$\frac{222}{5274 - 222} = 0.0439$$

     and the s.e. on the log-scale is is used to compute the 95% c.i.:

     $$\text{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

     so we get:

     $$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0502)$$

  3. The overall *rate* of death (per year) is

     $$222/2409.8 = 0.0921$$

     and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

     $$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

- Using your statistical package, you get (almost) the same confidence intervals, the programs are:

The SAS-program is in http://BendixCarstensen.com/EpiF2013/sasas `bissau-sol0.sas`.

```
data bissau;
  *filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  lpy = log( fuptime/36525 ) ;
run;

title "All children in study" ;
proc means data=bissau  nway ;
    var dead fuptime ;
    output out=bcgsum
           sum= ;
run;

proc print data=bcgsum ;
run ;

title "Use log-link to produce log-probability" ;
proc genmod  data = bcgsum ;
  model dead/_freq_ = / dist=bin link=log ;
  estimate "prob" intercept 1 / Exp ;
run ;
proc genmod  data = bcgsum ;
  model dead = / dist=bin link=log ;
  estimate "prob" intercept 1 / Exp ;
run ;

title "Use logit-link to produce log-odds" ;
proc genmod  data = bcgsum ;
  model dead/_freq_ = / dist=bin link=logit ;
  estimate "odds" intercept 1 / Exp ;
run ;
proc genmod  data = bissau ;
  model dead = / dist=bin link=logit ;
  estimate "odds" intercept 1 / Exp ;
run ;

data bcgsum ;
  set bcgsum ;
  * We want rates per 100 person-years ;
  lpy = log( fuptime/36525 ) ;
run ;

title "Poisson model to derive rate" ;
proc genmod  data = bcgsum ;
  model dead = / dist=poisson link=log offset=lpy ;
  estimate "rate" intercept 1 / Exp ;
run ;
proc genmod  data = bissau ;
  model dead = / dist=poisson link=log offset=lpy ;
  estimate "rate" intercept 1 / Exp ;
run ;
```

3.1.3 Rates, risks and odds

Program is in http://BendixCarstensen.com/EpiF2013/sasas `bissau-sol1.sas`.

```
data bissau;
  *filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dtpany=1; if dtp>0 then dtpany=2;
run;

title "Bissau data" ;
proc print data = bissau (obs=25) ;
run ;

title "Analysis by BCG groups" ;
proc means data=bissau  nway ;
    class bcg;
    var dead fuptime ;
    output out=bcgsum
           sum= ;
run;

proc print data=bcgsum ;
run ;

data bcgres;
```

```
  set bcgsum ;
/* Q1: We take the results from PROC MEANS (BCG=1) */
d=dead;
n=_freq_;
y=fuptime;

/* Q2: naive CI for pi */
pi = d/n;
sdpi = sqrt(pi*(1-pi)/n);
pilow = pi - 1.96*sdpi;
piup  = pi + 1.96*sdpi;

/* Q3: odds with CI + improved CI for pi*/
omega = pi/(1-pi);
sdlogomega = sqrt(1/d+1/(n-d));
errorfact  = exp(1.96*sdlogomega);
omegalow   = omega / errorfact;
omegaup    = omega * errorfact;
pilow_2    = omegalow/(1+omegalow);
piup_2     = omegaup /(1+omegaup );

/* Q4: rate per day - note that we added fuptime in PROC MEANS above */
lambda=d/y;
errorfact_rate = exp(1.96*sqrt(1/D));
lambda_low     = lambda / errorfact_rate;
lambda_up      = lambda * errorfact_rate;

/* Q5: rate per year is rate per day times 365.25 */
lambda_year     = lambda    *365.25;
lambda_year_low = lambda_low*365.25;
lambda_year_up  = lambda_up *365.25;
run;

proc print data=bcgres; run;

/* Q6: we repeat everything using the DTPANY variable created in the first DATA step */
title "Analaysis by DTP groups" ;
proc means data=bissau  nway ;
    class dtpany;
    var dead fuptime ;
    output out=dtpsum
           sum= ;
run;

proc print data=dtpsum ;
run ;

data dtpres;
  set dtpsum ;
/* Q1: We take the results from PROC MEANS (BCG=1) */
d=dead;
n=_freq_;
y=fuptime;

/* Q2: naive CI for pi */
pi = d/n;
sdpi = sqrt(pi*(1-pi)/n);
pilow = pi - 1.96*sdpi;
piup  = pi + 1.96*sdpi;

/* Q3: odds with CI + improved CI for pi*/
omega = pi/(1-pi);
sdlogomega = sqrt(1/d+1/(n-d));
errorfact  = exp(1.96*sdlogomega);
omegalow   = omega / errorfact;
omegaup    = omega * errorfact;
pilow_2    = omegalow/(1+omegalow);
piup_2     = omegaup /(1+omegaup );

/* Q4: rate per day - note that we added fuptime in PROC MEANS above */
lambda=d/y;
errorfact_rate = exp(1.96*sqrt(1/D));
lambda_low     = lambda / errorfact_rate;
lambda_up      = lambda * errorfact_rate;

/* Q5: rate per year is rate per day times 365.25 */
lambda_year     = lambda    *365.25;
lambda_year_low = lambda_low*365.25;
lambda_year_up  = lambda_up *365.25;
run;

proc print data=dtpres; run;
```

3.1.4 Rate ratio, risk ratio, odds ratio

Program is in http://BendixCarstensen.com/EpiF2013/sasas `bissau-sol2.sas`.

```
data bissau;
  filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dead2=dead;
  if dead2=0 then dead2=2;
  if dtp=0 then dtpany=2;
  if dtp>0 then dtpany=1;
run;

/*******************************************************************************/
/* Q1 Effect of BCG on mortality */
title "BCG EFFECT";
proc freq data=bissau;
  table bcg*dead2 / nocol nopercent relrisk;
run;

/*******************************************************************************/
/* Q2 Effect of DTP on mortality */
title "DTP EFFECT";
proc freq data=bissau;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/*******************************************************************************/
/* Q3 Association between BCG and DTP */
title "ASSOCIATION BETWEEN DTP AND BCG";
proc freq data=bissau;
  table dtpany*bcg / chisq;
run;

/*******************************************************************************/
/* Q4 Effect of DTP on mortality for each value (level) of BCG */
title "DTP EFFECT among BCG vaccinated";
proc freq data=bissau;
  where bcg=1;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

title "DTP EFFECT among BCG UN-vaccinated";
proc freq data=bissau;
  where bcg=2;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/* MUCH EASIER - YOU CAN ADD BCG IN THE THIS WAY: */
title "DTP EFFECT";
proc freq data=bissau;
  table bcg*dtpany*dead2 / nocol nopercent relrisk;
run;

/*******************************************************************************/
/* Q5 RATES */
/* We need no. of deaths and the sum of fullow-up time for each value of BCG: */
title "RATE RATIOS: BCG EFFECT";
proc means data=bissau sum;
  class bcg;
  var dead fuptime;
run;
/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  rate1 = (125/554929);
  rate2 = ( 97/325258);
  rr = rate1/rate2;
  sd = sqrt((1/125) + (1/97));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
run;
proc print data=rr;
  var rr lower upper error_factor ;
run;

title "RATE RATIOS: DTP EFFECT";
proc means data=bissau sum;
  class dtpany;
  var dead fuptime;
run;

/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  rate1 = (94/364012);
  rate2 = (128/516175);
  rr = rate1/rate2;
  sd = sqrt((1/94) + (1/128));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
```

```
run;
proc print data=rr;
  var rr lower upper error_factor ;
run;

title "RATE RATIOS: DTP EFFECT BY BCG STATUS";
proc means data=bissau sum;
  class bcg dtpany;
  var dead fuptime;
run;

/* Now we calculate by hand the rate ratio with 95%-CI: */
data rr;
  bcg="YES";
  rate1 = (92/358571);
  rate2 = (33/196358);
  rr = rate1/rate2;
  sd = sqrt((1/92) + (1/33));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
  output;
  BCG="NO";
  rate1 = (2/5441);
  rate2 = (95/319817);
  rr = rate1/rate2;
  sd = sqrt((1/2) + (1/95));
  error_factor = exp(1.96*sd);
  lower = rr/error_factor;
  upper = rr*error_factor;
  output;
run;

proc print data=rr;
  var bcg rr lower upper error_factor ;
run;
```

3.1.5 Confounder control: stratified analysis of odds ratio and risk ratio.

Program is in http://BendixCarstensen.com/EpiF2013/sasas `bissau-sol3.sas`.

```
/*****************************************************************************/
/* Read the data and transform */

data bissau;
* filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt" ;
  filename bisfile "../data/bissau.txt" ;
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  dead2=dead;
  if dead2=0 then dead2=2;
  if dtp=0 then dtpany=2;
  if dtp>0 then dtpany=1;
run;

/*****************************************************************************/
/* Q1 Effect of DTP on mortality by bcg */
title "DTP EFFECT for each level of bcg";
proc freq data=bissau;
  where bcg eq 1 ;
  table dtpany*dead2 / nocol nopercent relrisk;
run;
proc freq data=bissau;
  where bcg eq 2 ;
  table dtpany*dead2 / nocol nopercent relrisk;
run;

/*****************************************************************************/
/* Q2 Effect of DTP on mortality controlled for BCG */
title "DTP EFFECT controlled for BCG";
proc freq data=bissau;
  table bcg*dtpany*dead2 / nocol nopercent relrisk cmh ;
run;

/* Compare with bcg-UNadjusted effect of DTP */
title "DTP EFFECT NOT controlled for BCG";
proc freq data=bissau;
  table dtpany*dead2 / nocol nopercent relrisk ;
run;

/*****************************************************************************/
/* Q3 Effect of DTP on mortality controlled for age in months */
```

```
title "DTP EFFECT controlled for AGE";
proc freq data=bissau;
  table agem*dtpany*dead2 / nocol nopercent relrisk cmh ;
run;

/*****************************************************************************/
/* Q4 Effect of DTP on mortality controlled for age in months AND bcg */
title "DTP EFFECT controlled for AGE and BCG";
proc freq data=bissau;
  table bcg*agem*dtpany*dead2 / nocol nopercent relrisk cmh ;
run;

/*****************************************************************************/
/* Qx Effect of DTP on mortality controlled for age in months AND bcg
   using logistic regression */
title "DTP EFFECT controlled for AGE and BCG - logistic regression";
proc genmod data=bissau;
  class bcg agem dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=logit ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;

/* Using agem as continuoius (linear) */
title "DTP EFFECT controlled for AGE (continuous) and BCG - logistic regression";
proc genmod data=bissau;
  class bcg dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=logit ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;

/* Computing the RR effect by using the log-link */
title "DTP EFFECT controlled for AGE and BCG - logistic regression";
proc genmod data=bissau;
  class bcg agem dtpany ;
  model dead2 = bcg agem dtpany
        / dist=binomial  link=log ;
  estimate "OR by dtnm / agecont" dtpany 1 -1 / exp ;
run;
```

3.1.6 Survival analysis of childhood mortality in Guinea-Bissau

The SAS program `bissau.sas` in http://BendixCarstensen.com/EpiF2013/sas reads the data from the web and defines / recodes relevant variables and fits a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

1. Fit a simple Cox regression model with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates. In the program we have just recoded `bcg` to a 0/1 variable.

2. Estimate the effect of any dose of DTP, using the created variable `dtpany` adjusted only for age in months as a categorical (`class` in SAS) variable.

3. Now, also adjust for BCG. We then find a positive risk associated with DTP.

4. Is there an interaction between DTP (`dtpany`) and BCG?

   No, there is no interaction. From the table made by `proc tabulate` we see that there are virtually no children with a DTP vaccination who is without BCG vaccination. And we see that the mortality is very low among the second smallest group, those with BCG and no DTP. So the strange results hinges on the fact that almost 4000 of the 5000 in the study eiter are vaccinated by both or by none of the two.

   We see that there is an excess risk associated with DTP of 1.3 both for BCG=0 and for BCG=1, but neither are significant.

5. Make a Cox regression analysis with DTP (`dtpany`) and BCG, but now with age as time-variable, i.e. with delayed entry.

   We see that we get pretty much the same results as when using the age at entry as control variable.

6. Repeat the Poisson and logistic regression models that you have seen during the lectures.

   We see that the three types of analysis gives witually the same rsults for the protective effect of BCG-vaccination namely a mortality RR of 0.71, just significant.

   With rather short follow-up time, and hence little scope for censoring and variation of rates over time, the three different models are virtually the same.

   The Poisson model where we use the log of the persons-years as offset is a model where we assume that the mortality is constant throughout follow-up (as opposed to the Cox-model where it is allowed to vary witout any restrictions). The Logistic regression model is a further sinmplification of the Poisson model where we ignore censoring, so essentially assume that everyone is followd or the same time.

Program is in http://BendixCarstensen.com/EpiF2013/sas as `bissau-solcox.sas`.

```
options ps=200 nocenter ;

data bissau;
* filename bisfile url "http://www.biostat.ku.dk/~pka/epidata/bissau.txt";
  filename bisfile "../data/bissau.txt";
  infile bisfile firstobs=2;
  input id fuptime dead bcg dtp age agem;
  * DTP - indicator ;
  dtpany = dtp>0 ;
  * BCG-indicator ;
  bcg = 2 - bcg ;
  outage = age + fuptime ;
  lfup = log(fuptime) ;
run;

proc print data=bissau (obs=10) ;
run ;

proc tabulate  data=bissau  noseps formchar="           ";
  class agem ;
  var fuptime dead ;
  table agem, n*f=5. dead*f=5. fuptime*f=comma10. ;
run ;

title "Q1: Simple analysis of bcg effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = bcg agem / rl;
run;

title "Q2: Simple analysis of dtp effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = dtpany agem / rl ;
run;

title "Q3: Analysis of dtp and bcg effect" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = dtpany bcg agem / rl type3 ;
run;

title "Q4: Analysis of dtp and bcg effect with interaction" ;
proc phreg  data = bissau;
  class agem ;
  model fuptime * dead(0) = dtpany bcg dtpany*bcg agem / rl type3 ;
run;
title2 "Explaining the missing interaction and showing possible confounding" ;
proc tabulate data = bissau  noseps  missing  formchar=" ----   ---" ;
  class bcg dtpany dead ;
  table ( bcg all ) * dtpany,
        ( n * f=6. pctn<dead all> * f=6.1 ) * (dead all)
        / rts=15 ;
```

```
run ;

title "Q4a: Analysis of dtp effect separately for bcg Y/N" ;
proc sort data= bissau ; by bcg ; run ;
proc phreg  data = bissau;
  by bcg ;
  class agem ;
  model fuptime * dead(0) = dtpany agem / rl type3 ;
run;

title "Q5: Analysis of dtp and bcg effect using current age as timescale" ;
proc phreg  data = bissau;
  class agem ;
  model (age,outage) * dead(0) = dtpany bcg / rl type3 ;
run;

title "Q6: Analysis of bcg effect using age at entry - Cox-model" ;
proc phreg  data = bissau ;
  class agem ;
  model fuptime * dead(0) = bcg agem / rl ;
run;
title "Q6: Analysis of bcg effect using age at entry - Poisson-model" ;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson  offset = lfup ;
  estimate "RR bcg" bcg 1 / Exp ;
run;

title "Q6: Analysis of bcg effect using age at entry - Poisson-model with the same FU for all" ;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson ;
run;

title "Q6: Analysis of bcg effect using age at entry - logistic regression-model" ;
proc genmod  data = bissau  descending ;
  class agem ;
  model dead = bcg agem / dist = binomial ;
  estimate "RR bcg" bcg 1 / Exp ;
run;

title "QX: Logistic and Poission regression - the real cheat" ;
data xx ;
  set bissau ;
  if dead eq 0 then output ;
  if dead eq 1 then do ;
     fuptime = fuptime - 1 ; dead = 0 ; output ;
 fuptime =           1 ; dead = 1 ; output ;
 end ;
run ;

proc print data=xx (obs=10) ;
run ;

proc genmod  data = xx ;
  class agem ;
  model dead/fuptime = bcg agem / dist = binomial ;
run;
proc genmod  data = xx ;
  class agem ;
  model dead/fuptime = bcg agem / dist = binomial  link = cll;
run;
proc genmod  data = bissau ;
  class agem ;
  model dead = bcg agem / dist = poisson  offset=lfup ;
run;
```

3.2 Case-control study of renal cancer and trichloroethene

The exercise is based on Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichloroethane. J Cancer Res Clin Oncol, 1998, pp 374–382. [1].

The following points were addressed:

1. What is the primary aim of the study?

   The primary aim as stated is to assess the effect of thricholorethene (C_2HCl_3) and tetrachloroethene (C_2Cl_4) on the occurrence of kidney cancer. This is based on a

described possible biological mechanism.

2. How were the cases sampled?

   Cases in the study were patients who underwent surgery between 1987-12-01 and 1992-05-01, a period of 3.5 years. Of the 78 patients 62 responded (or their next of kin, since 4 of these 62 were dead)

3. How were the controls sampled?

   Controls were sampled from the accident wards of three *other* hospitals in the same area, form the period 1993-01-01 to 1993-12-31.

4. Are they comparable; i.e. what assumptions are needed?

   In order to deem cases and controls comparable we must assume:

   (a) In principle we must assume that the controls would have been enrolled as cases *if they had had a diagnosis of kidney cancer*.

   In practice it will suffice that the controls comes from a comparable populations, that is even though they might have been subjected to surgery at a different hospital in case of kidney cancer, if this were a similar type of hospital it would have been sufficient.

   (b) Exposure distribution in the population has not changed between the case-selection period (1987–1992) and the control sampling period (1993).

   It is of course difficult to assess this without further knowledge about the industrial development in Nordrhein-Westphalia in the period 1987–1993.

   (c) The attendance of the accident ward is unrelated to the exposure of interest.

   At face value it seems so, bar of course accidents related to trichloroethene exposure itself, but we can safely assume that they are very rare compared to other. However, the attendance at the accident ward is hardly unrelated to age, and neither is exposure to to trichloroethene, so the attendance *is* related to the exposure, albeit not directly.

   (d) Recall of exposure is the same among cases and controls.

   It is not stated anywhere whether the purpose of the study was revealed to cases and controls, but in the likely event that it were, cases may be more prone to recall exposure to the thricholorethene in order to get an explanation of their (severe) disease.

   As we see there are a few potential biases, some tend to *increase* the risk estimate; the problems with age is difficult to assess.

5. What is the (actual) study base?

   There is no definite answer to this, but if we define the study base to be the state of Nordrhein-Westphalia, the validity of the study hinges on assumptions that the hospitals where cases and controls are sampled are representative of the population — with respect to the exposure of interest.

   Of course, precisely the same argument applies if we define the study base to some subset of Nordrhein-Westphalia, which includes the uptake-areas of the 4 hospitals,

and incidentally we might define the study base as the uptake area of these 4 hospitals.

6. What study base is the intended (for generalization)?

   The intended study base is presumably all industrialized counties comparable to Germany.

7. Is the sampling scheme incidence density sampling?

   The sampling scheme is certainly not incidence density sampling, that would require that cases were selected among those at risk at the time of the case-occurrence, ad that is not the case.

8. Can the age-effect on the occurrence of renal cancer be estimated?

   The age-effect cannot be estimated; that would require that the controls were a representative sample (w.r.t. age) of the study base. Sampling persons from the accident wards makes this a far-fetched assumption.

9. Is age a confounder?

   Most likely so; it is definitely associated with both the exposure and the outcome.

10. Key in the numbers in table 6 (p.380), and verify the analysis using SAS `proc freq`.

    When we enter the data from table 6 there are a few things to be aware of:

    - Unlike previously, we are entering *grouped* or `tabulated` data from a table, not individual records. This means that every line in the input dataset represents as (usually large) number of individuals.
    - This in turn means that we shall need one line in the data set per entry in the table. In addition to the number, the line must then contain variables that tells us whether it is cases/controls, exposed or not, and what age.
    - Note that the entries in table 6 in the paper are not the number of cases and controls, but the number of cases and *total* number in the strata. So we need to do a bit of subtraction to get the numbers right.
    - Since tabular data are often quite short, it is most convenient to include the data in the program code itself. This is done *inside* a data step, and the data are preceded by a line with the word `datalines;` indicating that the next lines are data. The convention in SAS is that data is terminated by a line that only contains a semicolon, `;`.

    Thus the program looks like this:

```
data a ;
  input age tri ck n ;
  datalines ;
   30   1   1   2
   40   1   1   2
   50   1   1  10
   60   1   1   1
   70   1   1   4
   30   0   1   0
   40   0   1   1
   50   0   1  12
```

```
   60   0   1  17
   70   0   1   9
   30   1   0   1
   40   1   0   4
   50   1   0   2
   60   1   0   0
   70   1   0   0
   30   0   0  21
   40   0   0  11
   50   0   0  25
   60   0   0  14
   70   0   0   6
  ;
run ;
```

The entire SAS-program can be found in the folder http://bendixcarstensen.com/EpiE2012/sas/, and the program is listed at the end of this section.

Once we have the data, we use `proc freq`, first to produce a crude estimate and then an age-adjusted analysis, and we see that the estimate reported as the Mantel-Haenszel estimate in the paper in fact is the logit-estimate, based on adding 0.5 to all cells in age-strata where one entry is 0.

11. Is there any evidence of heterogeneity of the odds-ratio across age-classes? (*Hint:* Use the Breslow-Day-test.)

    There is no evidence of age-heterogeneity, but the age-classes are quite crude (10-years).

12. In particular, how does the odds-ratio estimate given by Vamvakas *et al.* compare to the Mantel-Haenszel estimate based on the same data?

    The estimate given in the paper is somewhat smaller than the MH-estimate, but the substantial message is pretty much the same: definitely an effect, but unclear how large.

13. What is the main result (in plain words)?

    Based on this study alone, there seems to be an excess risk of kidney cancer associated with trichloroethene exposure, but its cannot really be determined whether the OR is 2.5 or 25. This is of course due to the rather small sample size.

```
options  formchar=" ----------"
         nocenter ;

data a ;
  input age tri ck n ;
  datalines ;
   30   1   1   2
   40   1   1   2
   50   1   1  10
   60   1   1   1
   70   1   1   4
   30   0   1   0
   40   0   1   1
   50   0   1  12
   60   0   1  17
   70   0   1   9
   30   1   0   1
   40   1   0   4
   50   1   0   2
   60   1   0   0
   70   1   0   0
   30   0   0  21
   40   0   0  11
   50   0   0  25
```

```
   60   0   0  14
   70   0   0   6
  ;
run ;

proc print  data = a ;
run ;

proc freq  data = a ;
  table tri * ck / norow nocol nopct relrisk ;
  weight n ;
run ;

proc freq  data = a ;
  table age * tri * ck / norow nocol nopct cmh ;
  weight n ;
run ;

proc genmod  data = a  descending ;
  class age tri ;
  model ck = tri / dist = bin ;
  freq n ;
  estimate "OR" tri -1 1 / exp ;
run ;

proc genmod  data = a  descending ;
  class age tri ;
  model ck = age tri / dist = bin ;
  freq n ;
  estimate "OR" tri -1 1 / exp ;
run ;
```

3.3 IHD data from Clayton & Hills.

The study is described by Clayton & Hills, Ch. 13. The tabulated data set of counts of IHD cases and person-years is avaliable from `www` in the file `ihd-tab.txt`.

1. Fit the model from Clayton & Hills Tables 22.7-8 (p.222) and perform the tests from exercises 24.1 and 24.2 (pp.237–238). `SAS`-users may use the program `ihd-reg.sas` and notice the use of the `ESTIMATE` command to obtain a given reference group and the rate ratios with 95% confidence intervals.

2. Fit the model with interaction and re-find results from Clayton & Hills Table 24.5 (p.242) and the test for no interaction.

These two tasks are completed in the SAS-program listed below:

```
/*
  First read in the data as from the table in the book
*/
data ihd;
input eksp alder pyrs cases;
lpyrs=log(pyrs);
cards;
0 2  311.9  2
0 1  878.1 12
0 0  667.5 14
1 2  607.9  4
1 1 1272.1  5
1 0  888.9  8
;
run;

/*
  Q1:
  We fit the model with main effects of age and exposure, and we see
  we obtain the same estimates as in Clayton \& Hill's book
*/
proc genmod data=ihd;
 class eksp alder;
 model cases=eksp alder / dist=poisson offset=lpyrs type3;
estimate "exp. vs. non-exp." eksp 1 -1 / exp;
estimate "50-59 vs. 40-49"   alder 0 1 -1 / exp;
estimate "60-69 vs. 40-49"   alder 1 0 -1 / exp;
run;
```

```
/*
  Q2:
  In order to fit the model wit interaction we simply add the term
  eksp*alder and find the test for interaction towards the end of the
  output under type 3 analysis. Clearly there is no interaction; the
  p-value is 0.433.
*/
proc genmod data=ihd;
 class eksp alder;
 model cases = eksp alder eksp*alder / dist=poisson offset=lpyrs type3;
run;
```

3.3.1 Using continuous variables

The IHD-data contains energy consumption as a continuous variable, `energy`, among other variables.

The questions were:

1. Read the individual diet data records from the file.
2. Create variables for the person-years, by subtracting entry date from date of exit. Also create a variable with the log-person-years.
3. Use CHD as outcome variable in a Poisson-analysis with the log-person-years as offset, using energy as a linear explanatory variable. Is there an effect on mortality?
4. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (approx. the quartiles)
5. Same question for weight and BMI (the latter you have to calculate yourself as weight/height^2).

— and they are addressed in the following SAS-program:

```
options nocenter ;

/* Q1 & Q2
   First we read the individual records from the diet data, and create
   person-years and log-PY as well as spline variables for
    */
data ihdindiv ;
* filename dietfile url "http://BendixCarstensen.com/Epi-PhD/diet.txt";
  filename dietfile "../data/diet.txt";
    infile dietfile firstobs=2;
  input id doe dox chd dob job month energy height weight fat fibre ;
  * This is to read the date variables correctly in and print them reabably ;
  informat doe dox dob mmddyy10.;
    format doe dox dob ddmmyy10.;
  * We must compute the exposure variable ;
  exposure = ( energy < 2.75 ) ;
  * Here are the spline variables as shown in the lectures ;
      e200 = max( 0, energy-2.0 ) ;
      e250 = max( 0, energy-2.5 ) ;
      e300 = max( 0, energy-3.0 ) ;
  * To make the follow-up comparable with the grouped data only use FU from age 40 to 70 ;
       dox = min( dox, dob+70*365.25 ) ;
       doe = max( doe, dob+40*365.25 ) ;
  * Person-YEARS(!) and the log of them ;
      pyrs = ( dox - doe ) / 36525 ;
       lpy = log( pyrs ) ;
run ;

proc print  data = ihdindiv (obs=20) ;
run ;

/* Q3
   Exposure (energy intake) -- first as grouped variable, then as
   continuous with a linear effect on log-rates
   */
proc genmod  data = ihdindiv ;
  class exposure ;
  model chd = exposure
```

```
        / dist=poisson  link=log  offset=lpy ;
  estimate "low vs. high" exposure 1 -1 / Exp ;
  estimate "Rate unexp" intercept 1 exposure 1 0 / Exp ;
  estimate "Rate   exp" intercept 1 exposure 0 1 / Exp ;
run ;

proc genmod  data = ihdindiv ;
  model chd = energy
        / dist=poisson  link=log  offset=lpy ;
  estimate "Effect of 0.1 MJ" energy 0.1 / Exp ;
run ;

/* Q4
   Analysis with linear splines in exposure
   */
proc genmod  data = ihdindiv ;
  model chd = energy e250
        / dist=poisson  link=log  offset=lpy ;
  estimate "Effect above 2.5MJ" energy 1 e250 1 / Exp ;
run ;
proc genmod  data = ihdindiv ;
  model chd = energy e200 e250 e300
        / dist=poisson  link=log  offset=lpy ;
run ;

/* Q5
   Compute BMI (in kg/m2) and quartiles fof weigt & BMI to inform the
   placement of the knots. Note we shoudl base the knot palcement on
   the distribution of the variable among those with events, because
   this is where most of the information is.
   */
data ihdindiv ;
  set ihdindiv ;
  if height > 100 ;
  bmi = weight / ((height/100)**2) ;
run ;

proc tabulate  data = ihdindiv  missing noseps ;
  var height weight bmi ;
  class chd ;
  table n nmiss min q1 median q3 max,
        ( all chd ) * ( height*f=8.0 weight*f=8.1 bmi*f=6.1 ) / rts=10 ;
run ;

proc print data = ihdindiv ;
  where height < 100 ;
run ;

/* Breaks approx at quartiles */
data ihdindiv ;
  set ihdindiv ;
  bmi20 = max( 0, bmi-20 ) ;
  bmi25 = max( 0, bmi-25 ) ;
  bmi30 = max( 0, bmi-30 ) ;
  wt65  = max( 0, weight-65 ) ;
  wt75  = max( 0, weight-75 ) ;
  wt80  = max( 0, weight-80 ) ;
run ;

/* Models with linear BMI, 1 and 3 knots */
proc genmod  data = ihdindiv ;
  model chd = bmi / dist=poisson  link=log  offset=lpy ;
run ;
proc genmod  data = ihdindiv ;
  model chd = bmi bmi25 / dist=poisson  link=log  offset=lpy ;
run ;
proc genmod  data = ihdindiv ;
  model chd = bmi bmi20 bmi25 bmi30 / dist=poisson  link=log  offset=lpy ;
run ;

/* Models with linear weight, 1 and 3 knots */
proc genmod  data = ihdindiv ;
  model chd = weight / dist=poisson  link=log  offset=lpy ;
run ;
proc genmod  data = ihdindiv ;
  model chd = weight wt75 / dist=poisson  link=log  offset=lpy ;
run ;
proc genmod  data = ihdindiv ;
  model chd = weight wt65 wt75 wt80 / dist=poisson  link=log  offset=lpy ;
run ;
```

3.3.2 Splitting the follow-up of the IHD-data

The exercise gives quite a thorough introduction to the exercise, sor there is no need to repeat that here. The analyses can be done using the SAS-program `IHD-Lexis-sol` which is listed here:

```
* The grouped data (slightly altered from the book by C&H) ;
data ihdfreq;
  input eksp agr pyrs cases;
  lpyrs = log( pyrs );
datalines;
1 0  346.87  2
1 1  979.34 12
1 2  699.14 14
0 0  560.13  4
0 1 1127.70  6
0 2  794.15  8
;
run;

* Make a Poisson regression of the rates ;
proc genmod  data = ihdfreq ;
  class eksp agr ;
  model cases = eksp agr
                / dist = poisson  offset = lpyrs  type3 ;
  estimate "low vs. high" eksp -1 1 / Exp ;
  estimate "Rate E1 A0" intercept 1 eksp 1 0 agr 1 0 0 / Exp ;
run;

/* Q2
   Read the individual records from the diet data
   */
data ihdindiv ;
* filename dietfile url "http://BendixCarstensen.com/Epi-PhD/diet.txt";
  filename dietfile url "http://www.biostat.ku.dk/~pka/epidata/diet.txt";
    infile dietfile firstobs=2 ;
  input id doe dox chd dob job month energy height weight fat fibre ;
  * This is to read the date variables correctly in and print them reabably ;
  informat doe dox dob mmddyy10.;
    format doe dox dob ddmmyy10.;
  * We must compute the exposure variable ;
  exposure = ( energy < 2.75 ) ;
  drop job month energy height weight fat fibre ;
run ;

* Take a look at the individual records ;
proc print  data = ihdindiv (obs=20) ;
run ;

/* Q3
   Include the two macros for time-splitting and tabulation ;
   */
options source2 ;
filename Lexispr url "http://BendixCarstensen.com/Lexis/Lexis.sas";
%inc Lexispr ;
filename PYtabpr url "http://BendixCarstensen.com/Lexis/PYtab.sas";
%inc PYtabpr ;

/* Q4
   Timesplitting using the Lexis macro
   */
%Lexis( data   = ihdindiv,        /* Dataset with original data       */
        out    = ihdsplit,        /* Dataset with time-split data     */
        entry  = doe,             /* Date of entry                    */
        exit   = dox,             /* Date of exit                     */
        fail   = chd,             /* Event (failure) indicator        */
        origin = dob,             /* Origin of the time-scale         */
        scale  = 365.25,          /* Convert from input scale to braks-scale */
        breaks = 40 to 70 by 10,  /* Where to split the time scale    */
        left   = agr );           /* The name of the new age-variable */

/* Q6
   We can use proc print to inspect the resulting records
   */
proc print  data = ihdsplit (obs=20) ;
run ;

/* Q7
   We can use %PYtab to tabulate person-years and detahs:
   */
%PYtab( data  = ihdsplit,
        class = exposure agr,
        fail  = chd,
        risk  = risk,
        scale = 1000 ) ;
```

```
/* Q7a
   But we can also use proc tabulate to replicate the table from C&H
   in a readable form
   */
proc tabulate  data=ihdsplit  missing  noseps formchar=" ----   ---" ;
  class agr exposure ;
  var risk chd ;
  table agr="Age" all="Total",
        ( exposure="Exposure" all="Total" ) *
        sum=" " * ( chd="D"*f=6. risk="Y"*f=8.1 )
        / rts = 15 ;
run ;

/* Q8
   Now fit the model to the time-split data - results are as for the
   grouped data */
proc genmod  data = ihdsplit ;
  class agr exposure ;
  model chd = agr exposure / dist=poisson offset=lrisk ;
run ;

/* Q9
   Add an interaction between age and exposure and check that you
   get the same test for interaction as with the grouped data.
   */
proc genmod  data = ihdsplit ;
  class agr exposure ;
  model chd = agr exposure agr*exposure / dist=poisson offset=lrisk type3 ;
run ;

/* Q10
   We fit the main effects model to the grouped data (6 obs-dataset)
   and observe that the deviance is the same as the type-3 test for
   interaction based on the individual data
   */
proc genmod  data = ihdfreq ;
  class eksp agr ;
  model cases = eksp agr
              / dist = poisson  offset = lpyrs  type3 ;
run;
```

3.4 Case-control study of BCG vaccination and leprosy.

The study is decribed by Clayton & Hills, p.156. In short, 260 cases of leprosy among individuals aged less than 35 years were ascertained in a study area in Malawi. Subjects were grouped into 7 age intervals and according to absence or presence of a scar after BCG vaccination. Three sets of controls were studied:

1. a population survey of 80,622 persons
2. a random sample of 1000 persons
3. a 4 to 1 age-matched sample

1. Fit the model from Clayton & Hills Table 23.5 (p.232). SAS-users may use the program `bcg-reg.sas`; what are the reference groups?
2. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference groups (in SAS: use '`ESTIMATE`' statements).
3. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference.
4. Test the hypothesis of no interaction between `age` and `scar`.

5. Analyse the data set with only 1000 controls (i.e., use the controls `con1000`: Table 23.6, p.233) and compare the precision of the estimate for `scar` with that based on the entire sample.

6. Analyse the matched data set (i.e., use the controls `conmatch`: Table 23.6, p.233) and compare with the results from Table 23.7.

7. Try (erroneously) to drop `age` from the analysis of the matched data and study the consequences for the estimate of `scar`.

The following SAS program reads the data and produces the solutions to these tasks:

```
options nocenter ps=300 ;

data bcgdata;
filename bcgfile url 'http://www.biostat.ku.dk/~pka/epidata/bcgalldata.txt';
filename bcgfile url '../data/bcgalldata.txt';
  infile bcgfile firstobs=2;
  input age scar status $ n;
run;

/*
  Data overview
*/
proc print data=bcgdata (obs=50) ;
run ;

/*
  To do a particular analysis we must select cases and a particular
  set of controls with the where-statement, so the first analysis is
  with the total set of controls.

  Q1, Q2:
  We use the estimate statements to give the ORs of leprosy between
  age-classes, using both the first and last age-class as reference.
  Note that the s.e. of the log-odds-ratio is very large when we use
  the first age-class with only 2 cases as reference.
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age scar;
  model status = age scar / dist=bin type3 ;
  weight n;
  estimate "BCG scar   " scar -1 0 / exp ;
  estimate "Age 2 vs. 1" age -1 1 0 0 0 0 0 / exp ;
  estimate "Age 3 vs. 1" age -1 0 1 0 0 0 0 / exp ;
  estimate "Age 4 vs. 1" age -1 0 0 1 0 0 0 / exp ;
  estimate "Age 5 vs. 1" age -1 0 0 0 1 0 0 / exp ;
  estimate "Age 6 vs. 1" age -1 0 0 0 0 1 0 / exp ;
  estimate "Age 7 vs. 1" age -1 0 0 0 0 0 1 / exp ;
run;

/*
  Q2x:
  log-Odds-ratios with a specific reference can also be obtained using
  the param and ref options in the class statement
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0") ;
  model status = age scar / dist=bin type3 ;
  weight n;
  /* There is one parameter less, so we must omit the "-1"
     from the estimate statement */
  estimate "Age 2 vs. 1" age 1 0 0 0 0 0 / exp ;
run;

/*
  Q3:
  We now use the age-class 20-24 (age class 5) as reference
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age scar;
  model status = age scar / dist=bin type3 ;
  weight n;
  estimate "BCG scar   " scar -1 0 / exp ;
  estimate "Age 1 vs. 5" age 1 0 0 0 -1 0 0 / exp ;
  estimate "Age 2 vs. 5" age 0 1 0 0 -1 0 0 / exp ;
  estimate "Age 3 vs. 5" age 0 0 1 0 -1 0 0 / exp ;
```

```
  estimate "Age 4 vs. 5" age 0 0 0 1 -1 0 0 / exp ;
  estimate "Age 6 vs. 5" age 0 0 0 0 -1 1 0 / exp ;
  estimate "Age 7 vs. 5" age 0 0 0 0 -1 0 1 / exp ;
run;

/*
  Q4:
  We can test for interaction - the hypothesis that OR is the same in
  all classes. So we expand the model to one with separate OR for each
  age-class. The test for interaction is found undet Type 3 test. But
  is only the test for interaction if both age and scar are included
  as separate effects too.
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0") ;
  model status = age scar age*scar / dist=bin type3 ;
  weight n;
run;

/*
  If we want to use the trick of getting the effect directly,
  leaving out the scar from the model statment you must use the
  standard parametrization otherwise you fit a goofy model. Also note
  that you can omit trailing 0s from the estimate statement. Finally,
  the order of the estimates in the inetraction term depends on the
  order of the variables in the CLASS statement.
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age scar ;
  model status = age age*scar / dist=bin type3 ;
  weight n ;
  estimate "OR1" age*scar -1 1 / exp ;
  estimate "OR2" age*scar 0 0 -1 1 / exp ;
  estimate "OR3" age*scar 0 0 0 0 -1 1 / exp ;
  estimate "OR4" age*scar 0 0 0 0 0 0 -1 1 / exp ;
  estimate "OR5" age*scar 0 0 0 0 0 0 0 0 -1 1 / exp ;
  estimate "OR6" age*scar 0 0 0 0 0 0 0 0 0 0 -1 1 / exp ;
  estimate "OR7" age*scar 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 / exp ;
run;

/*
  Q5:
  Re-do analysis for reference - then the 1000 randomly selected
  controls what happens is that the estimate barely changes, but that
  the s.e. of the log-OR increases some 15%
*/
proc genmod data=bcgdata;
  where status='case' or status='conall' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0");
  model status = age scar / dist=bin type3 ;
  weight n ;
run;

proc genmod data=bcgdata;
  where status='case' or status='con1000' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0") ;
  model status = age scar / dist=bin type3 ;
  weight n ;
run;

/*
  Q6:
  Analysis using the matched data. This gives a slightly different
  estimate, but the interesting thing is that the s.e. shrinks a bit
  because of the better use of the controls
*/
proc genmod data=bcgdata;
  where status='case' or status='conmatch' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0") ;
  model status = age scar / dist=bin type3 ;
  weight n ;
run;

/*
  Q7:
  If we erroneously omit the matching variable (age) from the analysis
  we get a more precise estimate, but the estimate is quite strongly
  biased towards 0
*/
proc genmod data=bcgdata;
  where status='case' or status='conmatch' ;
  class age  (param=ref ref="1")
        scar (param=ref ref="0") ;
```

```
  model status = scar / dist=bin type3 ;
  weight n ;
run;
```

3.5 Case-control study of malignant melanoma.

Anne Østerlind conducted in the middle of the 80's a case-control study of risk factors for malignant melanoma in Denmark.

The review paper "Malignant melanoma in Denmark" from *Acta Oncologica*, 1990[3] is from Anne Østerlind's thesis and gives an overview of the results from the study which included 1400 interviewed persons, 474 cases and 926 controls, cf. table 5 in the article.

In the article incidence changes between 1943 and 1982 are also discussed; that part of the paper will not be touched upon in this exercise.

3.5.1 Discussion of the article.

1. Explain the design, the data base and data collection, particularly how the matching was conducted.

   The study is a group-matched (stratified) case-control study. Based on knowledge of the age- and sex-distribution of melanoma cases in Denmark a sample of the population with the same age- and sex-distribution was requested from the CPR.

   The data base (study base) is persons living in the eastern part of Denmark.

   Data were collected by personal interview.

2. How were interviews planned to minimize bias?

   Interviewers were blinded to the case/control status of the persons they interviewed for the study.

3. Explain the drop-out, particularly the analyses in Tables 5-7. What are the consequences of these results for the subsequent analyses?

   The drop-outs are differential between age-classes, older seem to be less likely to respond. It would have been more informative to have had the response probabilities for each subcategory, instead of the relative distribution separately for responders and non-responders.

4. How are the analyses carried out? Are all variables included in one step or are the analyses conducted in smaller steps? How are the matching variables accounted for? Comments?

   Analyses are conducted first with one variable at a time, giving marginal results, and then jointly to give mutually controlled results.

   Sex, but not age, is included in the controlled analyses, so strictly speaking there is a potential bias. Particularly if confounding is expected, that is if the distribution of the risk factors are different across age-classes. This is however not particularly likely.

5. Explain the analyses presented in Table 9. How many logistic regression models are fitted here?

Table 9 contains 6 analyses of one variable and 1 analysis with all variables, that is a total of 7 analyses.

6. What is the conclusion from the analyses in the table?

   Table 9 seems to indicate that freckles and to some extent skin and hair colour are the major (non-modifiable) risk factors for melanoma of the skin.

7. What is the purpose of Table 11?

   Table 11 shows the joint effect of freckles and naevi on the risk of melanoma. Clearly the risk increases by both, but there is no assessment of whether an interaction is present or not. Incidentally, as you will see, there is none.

8. Which modifiable factors seem to affect the melanoma risk?

   Number of sunburns before age 15. So one would in general expect that sunlight exposure increases the occurrence of melanoma. However detailed assessment of exposure history for each piece of skin for each study participant is of course impossible. So this seems to be pretty much as good as it gets.

3.5.2 Melanoma data

The SAS-program listed below reads data and does the analyses requested in the questions. Title statemenst and comments explains what question the code relates to.

3.5.3 Simple tabulation analysis

1. Make the two by two table showing the association between case-control status and whether or not the person experienced *any* sunburns before the age of 15. SAS-users may use the program `melanom.sas` to read in the data from `www`. Estimate the odds ratio with associated 95% confidence limits and test for no association between the risk factor and case-control status.

2. Conduct similar analyses for the factors `sex, hair, eyes, freckles, acuterea, chronrea`. Compare with Table 9 in the article.

3. The case control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables. Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

4. Same question for age.

These analyses can be accomplished using the following SAS program:

```
data mel;
* filename melfile url "http://www.biostat.ku.dk/~pka/epidata/melanom.txt";
* infile melfile firstobs=2 ;
infile "../data/melanom.txt" firstobs=2 ;
input casecon sex ageint agroup skin hair eyes freckles acuterea
      chronrea nvsmall nvlarge nvtot burn15;
run;

/* Q1 ----------------------------------------------------------------------
   Create binary variable for sunburns in new DATA step */
```

```
data mel;
  set mel;
  if burn15 = 0 then burnbin = 0 ;
  if burn15 > 0 then burnbin = 1 ;
  if burn15 = . then burnbin = . ;
/* or a little shorter and less easy to read (adding and subtracting
  burnbin makes sure that missing values are carried over to the new
  variable) */
  brbin = ( burn15 > 0 ) + burnbin - burnbin ;
run;

proc freq data=mel;
tables casecon * burnbin / chisq relrisk;
run;

/* Q2 ---------------------------------------------------------------------
   NB the analysis of sex is unjustified due to matching on sex */

proc freq data=mel;
tables casecon * sex / chisq relrisk;
run;

/* Q2 ---------------------------------------------------------------------
   The analysis of hair is done in 3 steps looking at 3 2 by 3 tables with
   dark haired people as reference throughout */

proc freq data = mel;
where hair=0 or hair=1;
    tables casecon*hair/chisq relrisk;
run;

proc freq data = mel;
where hair=0 or hair=2;
    tables casecon*hair/chisq relrisk;
run;

proc freq data = mel;
where hair=0 or hair=3;
    tables casecon*hair/chisq relrisk;
run;

/* Q2 ---------------------------------------------------------------------
   Similar approach for the other risk factors */

proc freq data=mel;
where eyes=0 or eyes=1;
    tables casecon*eyes/chisq relrisk;
run;

proc freq data=mel;
where eyes=0 or eyes=2;
    tables casecon*eyes/chisq relrisk;
run;

proc freq data=mel;
where skin=0 or skin=1;
    tables casecon*skin/chisq relrisk;
run;

proc freq data=mel;
where skin=0 or skin=2;
    tables casecon*skin/chisq relrisk;
run;

/* The factors freckles and acuterea need recoding first */

data mel;
  set mel;
  newfreck = 3 - freckles ;
  newacute = 4 - acuterea ;
  chronrea = chronrea - 1 ;
run;

/* freckels, one level at a time */
proc freq data=mel;
where newfreck=0 or newfreck=1;
    tables casecon*newfreck/chisq relrisk;
run;

proc freq data=mel;
where newfreck=0 or newfreck=2;
    tables casecon*newfreck/chisq relrisk;
run;

/* acute reaction, one level at a time */
proc freq data=mel;
where newacute=0 or newacute=1;
    tables casecon*newacute/chisq relrisk;
run;
```

```
proc freq data=mel;
where newacute=0 or newacute=2;
    tables casecon*newacute/chisq relrisk;
run;

proc freq data=mel;
where newacute=0 or newacute=3;
    tables casecon*newacute/chisq relrisk;
run;

/* chronic reaction, one level at a time */
proc freq data=mel;
where chronrea=0 or chronrea=1;
    tables casecon*chronrea/chisq relrisk;
run;

proc freq data=mel;
where chronrea=0 or chronrea=2;
    tables casecon*chronrea/chisq relrisk;
run;

proc freq data=mel;
where chronrea=0 or chronrea=3;
    tables casecon*chronrea/chisq relrisk;
run;

/* chronic reaction, one level at a time - controlling for sex */
proc freq data=mel;
where chronrea=0 or chronrea=1;
    tables sex * casecon * chronrea / chisq relrisk cmh ;
run;

proc freq data=mel;
where chronrea=0 or chronrea=2;
    tables sex * casecon * chronrea / chisq relrisk cmh ;
run;

proc freq data=mel;
where chronrea=0 or chronrea=3;
    tables sex * casecon * chronrea / chisq relrisk cmh ;
run;

/* Using proc genmod for chronic reaction */
proc genmod  data = mel  descending ;
  class chronrea ;
  model casecon = chronrea
        / dist=bin  type3 ;
  estimate "kronisk 1 vs 4" chronrea 1 0 0 -1 / exp ;
  estimate "kronisk 2 vs 4" chronrea 0 1 0 -1 / exp ;
  estimate "kronisk 3 vs 4" chronrea 0 0 1 -1 / exp ;
run ;

/* Using proc genmod for chronic reaction controlling for sex */
proc genmod  data = mel ;
  class chronrea sex ;
  model casecon = sex chronrea
        / dist=bin  type3 ;
  estimate
run ;

/* Using proc genmod for chronic reaction controlling for age and sex */
proc genmod  data = mel  descending ;
  class brbin sex agroup ;
  model casecon = brbin
        / dist=bin  type3 ;
  estimate "solsk < 15" brbin -1 1 / exp ;
run ;

proc genmod  data = mel  descending ;
  class brbin sex agroup ;
  model casecon = brbin sex agroup
        / dist=bin  type3 ;
  estimate "solsk < 15" brbin -1 1 / exp ;
run ;
```

3.5.4 Introductory analyses.

1. Estimate (log-)odds ratios for the variable `skin` (see top left in AØ's Table 9). `SAS`-users may use the program `melanom.sas`.

2. Estimate also odds ratios (in `SAS`: use `ESTIMATE` statements).

3. Conduct the other analyses in AØ's Table 9 (*left* part) where the factors `hair, eyes, freckles, acuterea, chronrea` are studied one at a time.

4. Conduct the analysis corresponding to Table 9 (*right* part) where several variables are included simultaneously (see the table footnote).

5. Reconstruct the results from AØ's Table 10 concerning number of raised naevi.

   NB: new variables must be defined from the original variables `nvtot, nvsmall, nvlarge`.

3.5.5 Trend tests and interactions.

6. In the analyses so far all variables have been considered as categorical ('`class`' in `SAS`) variables while all tests in Tables 9 and 10 are trend tests. Conduct the analyses which give the P-values in Table 9 (right part) for the variables `skin` and `freckles`.

7. May `freckles` be scored linearly (1, 2, 3), when this variable is studied separately? (Conduct a test for linearity/departures from trend).

8. In AØ's Table 11 `freckles` and the total number of naevi (suitably grouped) are studied. Conduct this analysis. Is there any interaction between these two variables?

9. Study, in a similar vein, interactions between `acuterea` and `skin` and between the grouped version of `nnvtot` from question 5. and `agroup`.

10. All of AØ's analyses are conducted without accounting for the match variable age (`agroup`) (in spite of warnings given by Clayton & Hills!). Repeat some of the previous analyses adjusting for `agroup`. Are there any substantial differences? Explain!

These analyses can be accomplished using the following SAS program:

```
options ps=300 ;
data mel;
* filename melfile url "http://www.biostat.ku.dk/~pka/epidata/melanom.txt";
* infile melfile firstobs=2 ;
infile "../data/melanom.txt" firstobs=2 ;
input casecon sex ageint agroup skin hair eyes freckles acuterea
      chronrea nvsmall nvlarge nvtot burn15;
run;

/* Q1 Create binary variable for sunburns in new DATA step */

data mel;
  set mel;
  if burn15 = 0 then burnbin = 0 ;
  if burn15 > 0 then burnbin = 1 ;
  if burn15 = . then burnbin = . ;
/* or a little shorter and less easy to read (adding and subtracting
  burnbin makes sure that missing values are carried over to the new
  variable) */
  brbin = ( burn15 > 0 ) + burnbin - burnbin ;
run;
proc print data = mel(obs=10) ; run;

/* Introductory analyses ********************************************/
/* Q1 & 2
   Effect of skincolor on the odds of MM.
   Remember the descending option to proc genmod to get the response modelling right
   */
proc genmod  data = mel  descending;
  class skin ;
  model casecon = skin / dist=binomial ;
  estimate "OR light  vs. dark  " skin -1  0 1 / exp ;
  estimate "OR medium vs. dark  " skin -1  1 0 / exp ;
```

```
  estimate "OR ligth vs. medium " skin  0 -1 1 / exp ;
run ;

proc genmod  data = mel  descending;
  model casecon = skin / dist=binomial ;
run ;

/* Q3
   Same analysis for hair eyes freckles akurea and chronrea
   */
proc genmod  data = mel  descending;
  class hair ;
  model casecon = hair / dist=binomial ;
  estimate "OR light vs. black  " hair -1 1 0 0 / exp ;
  estimate "OR blond vs. black  " hair -1 0 1 0 / exp ;
  estimate "OR red   vs. black  " hair -1 0 0 1 / exp ;
run ;
proc genmod  data = mel  descending;
  class eyes ;
  model casecon = eyes / dist=binomial ;
  estimate "OR grey vs. brown" eyes -1 1 0 / exp ;
  estimate "OR blue vs. brown" eyes -1 0 1 / exp ;
run ;
proc genmod  data = mel  descending;
  class freckles ;
  model casecon = freckles / dist=binomial ;
  estimate "OR some vs. none" freckles 0 1 -1 / exp ;
  estimate "OR many vs. none" freckles 1 0 -1/ exp ;
run ;
proc genmod  data = mel  descending;
  class acuterea ;
  model casecon = acuterea / dist=binomial ;
  estimate "OR blisters vs. none" acuterea  1 0 0 -1 / exp ;
  estimate "OR painful  vs. none" acuterea  0 1 0 -1 / exp ;
  estimate "OR mild     vs. none" acuterea  0 0 1 -1 / exp ;
run ;
proc genmod  data = mel  descending;
  class chronrea ;
  model casecon = chronrea / dist=binomial ;
  estimate "OR moderate vs. deep" chronrea -1 1 0 0 / exp ;
  estimate "OR mild     vs. deep" chronrea -1 0 1 0 / exp ;
  estimate "OR no tan   vs. deep" chronrea -1 0 0 1 / exp ;
run ;

/* Q4
   Here repeat all the previous analyses where all varaibles are included
   simultaneously and adjusted for sex
   */
proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea ;
  model casecon = sex skin hair eyes freckles acuterea chronrea / dist=binomial ;
  estimate "OR light  vs. dark  " skin -1 0 1 / exp ;
  estimate "OR medium vs. dark  " skin -1 1 0 / exp ;
  estimate "OR light vs. black  " hair -1 1 0 0 / exp ;
  estimate "OR blond vs. black  " hair -1 0 1 0 / exp ;
  estimate "OR red   vs. black  " hair -1 0 0 1 / exp ;
  estimate "OR grey vs. brown" eyes -1 1 0 / exp ;
  estimate "OR blue vs. brown" eyes -1 0 1 / exp ;
  estimate "OR some vs. none" freckles 0 1 -1 / exp ;
  estimate "OR many vs. none" freckles 1 0 -1/ exp ;
  estimate "OR blisters vs. none" acuterea  1 0 0 -1 / exp ;
  estimate "OR painful  vs. none" acuterea  0 1 0 -1 / exp ;
  estimate "OR mild     vs. none" acuterea  0 0 1 -1 / exp ;
  estimate "OR moderate vs. deep" chronrea -1 1 0 0 / exp ;
  estimate "OR mild     vs. deep" chronrea -1 0 1 0 / exp ;
  estimate "OR no tan   vs. deep" chronrea -1 0 0 1 / exp ;
run ;

proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea ;
  model casecon = sex skin hair eyes freckles acuterea chronrea / dist=binomial type3 ;
run ;


/* Q5
   Construct the total numver of nevi and group then 1 / 2-4 /5+
   */
data mel ;
  set mel ;
  gnvsml = ( nvsmall>0 ) + ( nvsmall>1 ) + ( nvsmall>4 ) + nvsmall-nvsmall ;
  gnvlrg = ( nvlarge>0 ) + ( nvlarge>1 ) + ( nvlarge>4 ) + nvlarge-nvlarge ;
  gnvtot = ( nvtot  >0 ) + ( nvtot  >1 ) + ( nvtot  >4 ) + nvtot  -nvtot   ;
run ;

/* UNadjusted analyses */
proc genmod  data = mel  descending;
  class gnvsml ;
  model casecon = gnvsml / dist=binomial ;
  estimate "OR gnvsml   1 vs. none" gnvsml -1 1 0 0 / exp ;
```

```
  estimate "OR gnvsml 2-4 vs. none" gnvsml -1 0 1 0 / exp ;
  estimate "OR gnvsml  5+ vs. none" gnvsml -1 0 0 1 / exp ;
  estimate "OR gnvsml   1 vs. none" gnvsml -1 1 0 0 / exp ;
  estimate "OR gnvsml 2-4 vs.    1" gnvsml 0 -1 1 0 / exp ;
  estimate "OR gnvsml  5+ vs.  2-4" gnvsml 0 0 -1 1 / exp ;
run ;

proc genmod  data = mel  descending;
  class gnvlrg ;
  model casecon = gnvlrg / dist=binomial ;
  estimate "OR gnvlrg   1 vs. none" gnvlrg -1 1 0 0 / exp ;
  estimate "OR gnvlrg 2-4 vs. none" gnvlrg -1 0 1 0 / exp ;
  estimate "OR gnvlrg  5+ vs. none" gnvlrg -1 0 0 1 / exp ;
run ;

proc genmod  data = mel  descending;
  class gnvtot ;
  model casecon = gnvtot / dist=binomial ;
  estimate "OR gnvtot   1 vs. none" gnvtot -1 1 0 0 / exp ;
  estimate "OR gnvtot 2-4 vs. none" gnvtot -1 0 1 0 / exp ;
  estimate "OR gnvtot  5+ vs. none" gnvtot -1 0 0 1 / exp ;
run ;

/* Adjusted analyses */
proc genmod  data = mel  descending;
  class gnvsml sex freckles hair skin ;
  model casecon = gnvsml sex freckles hair skin / dist=binomial type3 ;
  estimate "OR gnvsml   1 vs. none" gnvsml -1 1 0 0 / exp ;
  estimate "OR gnvsml 2-4 vs. none" gnvsml -1 0 1 0 / exp ;
  estimate "OR gnvsml  5+ vs. none" gnvsml -1 0 0 1 / exp ;
run ;

proc genmod  data = mel  descending;
  class gnvlrg sex freckles hair skin ;
  model casecon = gnvlrg sex freckles hair skin / dist=binomial ;
  estimate "OR gnvlrg   1 vs. none" gnvlrg -1 1 0 0 / exp ;
  estimate "OR gnvlrg 2-4 vs. none" gnvlrg -1 0 1 0 / exp ;
  estimate "OR gnvlrg  5+ vs. none" gnvlrg -1 0 0 1 / exp ;
run ;

proc genmod  data = mel  descending;
  class gnvtot sex freckles hair skin ;
  model casecon = gnvtot sex freckles hair skin / dist=binomial ;
  estimate "OR gnvtot   1 vs. none" gnvtot -1 1 0 0 / exp ;
  estimate "OR gnvtot 2-4 vs. none" gnvtot -1 0 1 0 / exp ;
  estimate "OR gnvtot  5+ vs. none" gnvtot -1 0 0 1 / exp ;
run ;

/* Trend tests and interactions ***************************************/

/* Q6
   Trend tests for skin and freckles
   */
data mel ;
  set mel ;
  vskin = skin ;
  vfreck = freckles ;
run ;

proc genmod  data = mel  descending;
  class skin ;
  model casecon = vskin / dist=binomial type3 ;
  run ;
proc genmod  data = mel  descending;
  class skin ;
  model casecon = vskin skin / dist=binomial type3 ;
  run ;
proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea ;
  model casecon = sex vskin hair eyes freckles acuterea chronrea / dist=binomial type3 ;
  run ;
proc genmod  data = mel  descending;
  class freckles ;
  model casecon = vfrevk / dist=binomial type3 ;
  run ;
proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea ;
  model casecon = sex skin hair eyes acuterea chronrea vfreck / dist=binomial type3 ;
  run ;

/* Q7
   Test for linearity - test if the categorical variable can be removed
   */
proc genmod  data = mel  descending;
  class freckles ;
  model casecon = vfreck freckles / dist=binomial type3 ;
  run ;

proc genmod  data = mel  descending;
```

```
  class sex skin hair eyes freckles acuterea chronrea ;
  model casecon = sex skin hair eyes acuterea chronrea  vfreck freckles / dist=binomial type3 ;
  run ;

/* Q8
   Interaction between freckles and total naevi.
   First the table
   */
proc tabulate  data=mel  missing noseps formchar=" ----   ---" ;
  class casecon freckles gnvtot ;
  table freckles all,
        ( gnvtot all ) * casecon="Ca/Co" * f= 5.
        / rts = 15 ;
run ;
/* Without missing */
proc tabulate  data=mel noseps formchar=" ----   ---" ;
  class casecon freckles gnvtot ;
  table freckles all,
        ( gnvtot all ) * casecon="Ca/Co" * f= 5.
        / rts = 15 ;
run ;

/* Estimates of OR for each combination of freckles and nvtot,
   relative to the lowest level of both.
   */
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = freckles * gnvtot / dist=binomial ;
  estimate "1,2" freckles * gnvtot 0 0 0 0  0 0 0 0 -1 1 0 0 / exp ;
  estimate "1,3" freckles * gnvtot 0 0 0 0  0 0 0 0 -1 0 1 0 / exp ;
  estimate "1,4" freckles * gnvtot 0 0 0 0  0 0 0 0 -1 0 0 1 / exp ;
  estimate "2,1" freckles * gnvtot 0 0 0 0  1 0 0 0 -1 0 0 0 / exp ;
  estimate "2,2" freckles * gnvtot 0 0 0 0  0 1 0 0 -1 0 0 0 / exp ;
  estimate "2,3" freckles * gnvtot 0 0 0 0  0 0 1 0 -1 0 0 0 / exp ;
  estimate "2,4" freckles * gnvtot 0 0 0 0  0 0 0 1 -1 0 0 0 / exp ;
  estimate "3,1" freckles * gnvtot 1 0 0 0  0 0 0 0 -1 0 0 0 / exp ;
  estimate "3,2" freckles * gnvtot 0 1 0 0  0 0 0 0 -1 0 0 0 / exp ;
  estimate "3,3" freckles * gnvtot 0 0 1 0  0 0 0 0 -1 0 0 0 / exp ;
  estimate "3,4" freckles * gnvtot 0 0 0 1  0 0 0 0 -1 0 0 0 / exp ;
  run ;
/* Same model, but differently parametrized to make the type3
   likelihood ratio test for interaction meaningful */
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = freckles gnvtot freckles*gnvtot / dist=binomial type3 ;
  run ;
/* Likelihood ratio test for quantitatlive interaction by defining a
   quantitative interaction variable, a product of nvtot */
data mel ;
  set mel;
  fnv = (4-freckles)*(gnvtot+1) ;
  vfreck = freckles ;
  vnvtot = gnvtot ;
run ;

title1 "Parametric 1 d.f. interaction between freckles and naevi" ;
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = freckles gnvtot fnv / dist=binomial type3 ;
  run ;

title1 "Marginal test for linearity by freckles" ;
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = vfreck freckles gnvtot / dist=binomial type3 ;
  run ;
title1 "Marginal test for trend by freckles" ;
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = vfreck gnvtot / dist=binomial type3 ;
  run ;

title1 "Marginal test for linearity by naevi" ;
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = freckles vnvtot gnvtot / dist=binomial type3 ;
  run ;
title1 "Marginal test for trend by naevi" ;
proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = freckles vnvtot / dist=binomial type3 ;
  run ;

proc genmod  data = mel  descending;
  class freckles gnvtot ;
  model casecon = vfreck vnvtot / dist=binomial type3 ;
  run ;

/* Q9a
```

```
   Interaction between acuterea and skin */
title1 "Acuterea and skin" ;
proc tabulate  data=mel  missing noseps formchar=" ----   ---" ;
  class casecon acuterea skin ;
  table skin all,
        ( acuterea all ) * casecon="Ca/Co" * f=5.
        / rts = 15 ;
run ;
/* Without missing */
proc tabulate  data=mel  noseps formchar=" ----   ---" ;
  class casecon acuterea skin ;
  table skin all,
        ( acuterea all ) * casecon="Ca/Co" * f=5.
        / rts = 15 ;
run ;

/* A variable to use as quantitrative interaction:
   Skin is coded so that increasing values 0-2 corresponsds to increasing risk,
   Acuterea is coded so that DEcreasing values 4-1 corresponsds to increasing risk
   */
data mel ;
  set mel;
  acsk = (5-acuterea)*(skin+1) ;
run ;

proc genmod  data = mel  descending;
  class skin acuterea;
  model casecon = skin * acuterea / dist=binomial ;
  estimate " dark,mild" skin * acuterea 0 0 1 -1  0 0 0 0  0 0 0 0 / exp ;
  estimate " dark,pain" skin * acuterea 0 1 0 -1  0 0 0 0  0 0 0 0 / exp ;
  estimate " dark,blis" skin * acuterea 1 0 0 -1  0 0 0 0  0 0 0 0 / exp ;
  estimate "  med,none" skin * acuterea 0 0 0 -1  0 0 0 1  0 0 0 0 / exp ;
  estimate "  med,mild" skin * acuterea 0 0 0 -1  0 0 1 0  0 0 0 0 / exp ;
  estimate "  med,pain" skin * acuterea 0 0 0 -1  0 1 0 0  0 0 0 0 / exp ;
  estimate "  med,blis" skin * acuterea 0 0 0 -1  1 0 0 0  0 0 0 0 / exp ;
  estimate "light,none" skin * acuterea 0 0 0 -1  0 0 0 0  0 0 0 1 / exp ;
  estimate "light,mild" skin * acuterea 0 0 0 -1  0 0 0 0  0 0 1 0 / exp ;
  estimate "light,pain" skin * acuterea 0 0 0 -1  0 0 0 0  0 1 0 0 / exp ;
  estimate "light,blis" skin * acuterea 0 0 0 -1  0 0 0 0  1 0 0 0 / exp ;
  run ;
/* Same model, but differently parametrized to make the type3
   likelihood ratio test for interaction meaningful */
title1 "Skin x acuterea interaction -  ?" ;
proc genmod  data = mel  descending;
  class skin acuterea ;
  model casecon = skin acuterea skin*acuterea / dist=binomial type3 ;
  run ;
/* Likelihood ratio test for quantitative interaction by defining a
   quantitative interaction variable, a product of skin and acuterea */
title1 "Skin x acuterea interaction - is it linear ?" ;
proc genmod  data = mel  descending;
  class skin acuterea ;
  model casecon = skin acuterea acsk skin*acuterea / dist=binomial type3 ;
  run ;
title1 "Skin x acuterea interaction - is it significant GIVEN linear ?" ;
proc genmod  data = mel  descending;
  class skin acuterea ;
  model casecon = skin acuterea acsk  / dist=binomial type3 ;
  run ;

/* Q10
   Analysis skin as categorical / linear with and without agroup
   Very small differences
   */
title1 "skin without agroup" ;
proc genmod  data = mel  descending;
  class skin agroup ;
  model casecon = skin / dist=binomial type3 ;
run ;
title1 "skin with agroup" ;
proc genmod  data = mel  descending;
  class skin agroup ;
  model casecon = skin agroup / dist=binomial type3 ;
run ;
title1 "Trend test for skin without agroup" ;
proc genmod  data = mel  descending;
  class skin agroup ;
  model casecon = vskin / dist=binomial type3 ;
run ;
title1 "Trend test for skin with agroup" ;
proc genmod  data = mel  descending;
  class skin agroup ;
  model casecon = vskin agroup / dist=binomial type3 ;
run ;
title1 "Trend test for skin without agroup but other vars" ;
proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea agroup ;
  model casecon = sex vskin hair eyes freckles acuterea chronrea / dist=binomial type3 ;
run ;
```

```
title1 "Trend test for skin with agroup and other vars" ;
proc genmod  data = mel  descending;
  class sex skin hair eyes freckles acuterea chronrea agroup ;
  model casecon = sex vskin hair eyes freckles acuterea chronrea agroup / dist=binomial type3 ;
run ;

title1 "brbin without agroup" ;
proc genmod  data = mel  descending;
  class brbin agroup ;
  model casecon = brbin / dist=binomial type3 ;
run ;
title1 "brbin with agroup" ;
proc genmod  data = mel  descending;
  class brbin agroup ;
  model casecon = brbin agroup / dist=binomial type3 ;
run ;
```

3.6 Testicular cancer risk and maternal parity.

This exercise deals with the article "Testicular cancer risk and maternal parity: a population-based cohort study", by T. Westergaard, P.K. Andersen, J.B. Pedersen, M. Frisch, J.H. Olsen, M. Melbye. *Br. J. Cancer*, **77**,pp. 1180-1185 (1998). [4].

3.6.1 Practical exercises

The file `testis.txt`, available at `www` contains for each (non-empty) combination of the factors `SON_AGE, SON_KOH, MOTH_AGE, PARITY` the number of person-years at risk `PYRS`, the numbers of non-seminomas and seminomas, respectively `NONSEMI SEMI`, and the total number of testis cancer cases `CASES`. The first line of the file contains the variable names.

The SAS-program `Testis-sol.sas` listed below reads the data and does the analyses required to answer the practical questions.

10. Compute the crude rate ratio for testis cancer for parity 2+ versus parity 1. Compare with 5. above. SAS-users may use the SAS program `testis.sas` (and `PROC GENMOD`).

11. Reconstruct the estimates for "parity of mother at birth of son" from the top of Table 1 in the article both for all testis cancers and for non-seminomas.

12. Reconstruct the estimates from Table 2 in the article concerning mother's age (for all testis cancers). Is there an interaction between parity and mother's age?

13. Same question for birth cohort of the son.

```
data twe;
 filename testfile url "http://www.biostat.ku.dk/~pka/epidata/testis.txt";
   infile testfile firstobs=2;
   input SON_AGE SON_KOH MOTH_AGE PARITY PYRS NONSEMI SEMI CASES;
   lpyrs=log(pyrs);
   par2 = (parity>=2);
   * Merging the the first two cohorts ;
   if son_koh=1973 then son_koh=1968;
   * Grouping cohort to 3 levels ;
        if son_koh=1950 or son_koh=1958 then son_kohny=1;
   else if son_koh=1963                 then son_kohny=2;
   else if son_koh=1968 or son_koh=1973 then son_kohny=3;
run;

/* Q10
   */
title 'Q10  Crude RR for parity 2+ vs 1';
proc genmod data=twe;
   class par2;
   model cases = PAR2 / offset=lpyrs dist=poisson type3 ;
   estimate "par 2+ vs par 1" PAR2 -1 1 / exp ;
run;
```

```
/* Q11
   */
title 'Q11a  All testis cancers';
proc genmod data=twe ;
  class SON_AGE SON_KOH MOTH_AGE parity;
  model cases = SON_AGE SON_KOH MOTH_AGE PARity
               / offset=lpyrs dist=poi type3;
  estimate 'par 2  vs. par 1' parity -1 1 0 0 /exp;
  estimate 'par 3  vs. par 1' parity -1 0 1 0 /exp;
  estimate 'par 4+ vs. par 1' parity -1 0 0 1 /exp;
run;

title 'Q11b  Non seminoma testis cancers';
proc genmod data=twe;
  class SON_AGE SON_KOH MOTH_AGE parity;
  model nonsemi = SON_AGE SON_KOH MOTH_AGE PARity
                / offset=lpyrs dist=poi type3;
  estimate 'par 2  vs. par 1' parity -1 1 0 0 /exp;
  estimate 'par 3  vs. par 1' parity -1 0 1 0 /exp;
  estimate 'par 4+ vs. par 1' parity -1 0 0 1 /exp;
run;

/* Q12
   Estimates concerning Mother's age at first birth
   */
title "Q12: Mother's age at boy's birth - test of interaction" ;
proc genmod data=twe;
  class SON_AGE SON_KOH MOTH_AGE par2;
  model cases = SON_AGE SON_KOH MOTH_AGE par2 moth_age*par2
               / offset=lpyrs dist=poi type3 ;
run;

title "Q12: Mother's age at boy's birth - interaction effects" ;
proc genmod data=twe;
  class SON_AGE SON_KOH MOTH_AGE par2;
  model cases = SON_AGE SON_KOH MOTH_AGE  moth_age*par2
               / offset=lpyrs dist=poi type3;
estimate 'par2+ vs. 1,   <20' moth_age*par2 -1 1  0 0  0 0  0 0 /exp;
estimate 'par2+ vs. 1, 20-24' moth_age*par2  0 0 -1 1  0 0  0 0 /exp;
estimate 'par2+ vs. 1, 25-29' moth_age*par2  0 0  0 0 -1 1  0 0 /exp;
estimate 'par2+ vs. 1, 30+  ' moth_age*par2  0 0  0 0  0 0 -1 1 /exp;
run;

/* Q13
   Estimates concerning Birth cohort of the son
   */
title 'Q13  Birth cohort of the son - test of interaction' ;
proc genmod data=twe ;
  class SON_AGE MOTH_AGE son_kohny son_koh par2;
  model cases=SON_AGE MOTH_AGE  son_koh par2 son_kohny*par2
                  /offset=lpyrs dist=poi type3;
run;

title 'Q13  Birth cohort of the son - interaction effects' ;
proc genmod data=twe ;
  class SON_AGE MOTH_AGE son_kohny son_koh par2;
  model cases=SON_AGE MOTH_AGE  son_koh son_kohny*par2
                  /offset=lpyrs dist=poi type3;
  estimate 'par 2+ vs. 1, 1950-63' son_kohny*par2 -1 1  0 0  0 0 /exp;
  estimate 'par 2+ vs. 1, 1963-67' son_kohny*par2  0 0 -1 1  0 0 /exp;
  estimate 'par 2+ vs. 1, 1968-92' son_kohny*par2  0 0  0 0 -1 1 /exp;
run;
```

Chapter 4

Solutions with **Stata**

The Stata-programs are available on the course web site in the folder http://BendixCarstensen.com/EpiF2013/stata. There is also a link to this on the website.

Stata

Stata is a commercial statistical package which is renowned for it speed. It is not as expensive as SAS. It has good capabilities for documenting analyses through log-files (derived from `do`-files) that provides good and readable documentation of analyses in a readable format.

4.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

4.1.1 A single risk, odds and rate

4.1.2 A single risk, odds and rate

Tabulate the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naïve 95% confidence inerval for this is:

     $$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

     but a better one is the formula:

     $$\frac{p}{p + (1-p) \overset{\times}{\div} \text{erf}}, \quad \text{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\text{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

  2. The overall odds of death is simply:

     $$\frac{222}{5274 - 222} = 0.0439$$

     and the s.e. on the log-scale is is used to compute the 95% c.i.:

     $$\text{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

     so we get:

     $$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0502)$$

  3. The overall *rate* of death (per year) is

     $$222/2409.8 = 0.0921$$

     and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

     $$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

- Using your statistical package, you get (almost) the same confidence intervals, the programs are:

The Stata-program is in http://BendixCarstensen.com/EpiF2013/stataas `bissau-sol0.do`.

The following `Stata`-programs do all the calculations required in the questions for the other exercsies on the Bissau-data:

4.1.3 Rates, risks and odds

4.1.4 Rate ratio, risk ratio, odds ratio

```
use "E:\Epidemiologi\bissau.dta", replace

*The risk of death

* TO USE THE EPITAB COMMANDS WE NEED EXPOSURE VARIABLES CODED AS 0 or 1 AND
* OUTCOME CODED AS 0 or 1

gen bcg01=bcg
replace bcg01=0 if bcg==2

*QUESTION 2.1.1. 1 and 2
ci dead if bcg==1, binomial wald

ci dead if bcg==2, binomial wald


*QUESTION 2.1.1.3

tabodds dead bcg01

*QUESTION 2.1.1.4

ci dead if bcg==1, exposure(fuptime)
ci dead if bcg==2, exposure(fuptime)

*QUESTION 2.1.1.5
gen year=fuptime/365.25

ci dead if bcg==1, exposure(year)
ci dead if bcg==2, exposure(year)

*QUESTION 2.1.1.6
gen dtp_any=1 if dtp>0
replace dtp_any=0 if dtp==0

*RISK
ci dead if dtp_any==1, binomial wald

ci dead if dtp_any==0, binomial wald

*ODDS
tabodds dead dtp_any

*RATES

ci dead if dtp_any==1, exposure(fuptime)
ci dead if dtp_any==0, exposure(fuptime)

ci dead if dtp_any==1, exposure(year)
ci dead if dtp_any==0, exposure(year)


**************************
* 2.1.2                  *
**************************

*RISK RATIO
*VARIABLES MUST BE 0/1

cs dead bcg01

cc dead bcg01

*QUESTION 2.1.2.2

cs dead dtp_any

cc dead dtp_any

*QUESTION 2.1.2.3

tab bcg01 dtp_any, chi2 expected

*QUESTION 2.1.2.4

cs dead dtp_any if bcg01==1
cs dead dtp_any if bcg01==0

cc dead dtp_any if bcg01==1
cc dead dtp_any if bcg01==0

cc dead dtp_any, by(bcg01)
mhodds dead dtp_any bcg01
```

```
*QUESTION 2.1.2.5

ir dead bcg01 fuptime

ir dead dtp_any fuptime

ir dead dtp_any  fuptime if bcg01==1
ir dead dtp_any  fuptime if bcg01==0
```

4.1.5 Confounder control: stratified analysis of odds ratio and risk ratio.

4.1.6 Survival analysis of childhood mortality in Guinea-Bissau

The Stata program `bissau-cox.do` in http://BendixCarstensen.com/EpiF2013/stata reads the data from the web and defines / recodes relevant variables and fits Cox regression models with follow-up time as the time variable and including `bcg` and `agem` as categorical covariates.

```
use "C:\ewan\Epidemiologi\Data\bissau.dta", clear


*QUESTION 1
*GETTING READY TO DO A SURVLVAL ANALYSIS

stset fuptime, failure(dead==1)
*NEW VARIABLES ARE CREATED

*COX MODEL

stcox i.bcg i.agem

*THE HAZARD OF DEATH IS 42% HIGHER IN SUBJECTS WITHOUT A BCG VACCINE COMPARED TO
*SUBJECTS WITH THE VACCINE ADJUSTED FOR AGE.


*GETTING THE PARAMETRISATION FROM THE SLIDES:

stcox b2.bcg b6.agem, nohr

stcox b2.bcg b6.agem

*QUESTION 2
gen dtpany=0 if dtp==0
replace dtpany=1 if dtp>0

stcox i.dtpany i.agem

*QUESTION 3
*ADJUST FOR BCG

stcox i.dtpany i.agem b2.bcg
*THE EFFECT OF DTAANY INCREASES

*QUESTION 4
*INTERACTION BETWEEN BCG AND DTPANY?

stcox i.dtpany i.agem b2.bcg i.dtpany#b2.bcg
*THE WALD TEST FOR THE INTERACTION IS NOT SIGNIFICANT
est store m1

stcox i.dtpany i.agem b2.bcg
est store m2
lrtest m1 m2
*LIKELIHOOD RATIO NOT SIGNIFICANT

*QUESTION 5
*USE AGE AS TIME SCALE


gen outage=age+fuptime

stset outage, failure(dead==1) enter(age)

stcox i.dtpany b2.bcg

*QUESTION 6
*REPEAT ANALYSES FROM LECTURES
```

```
stset fuptime, failure(dead==1)
stcox b2.bcg i.agem
poisson dead b2.bcg i.agem, exposure(fuptime) ir
logistic dead b2.bcg i.agem
```

4.2 Case-control study of renal cancer and trichloroethene

The exercise is based on Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichloroethane. J Cancer Res Clin Oncol, 1998, pp 374–382. [1].

The following points were addressed:

1. What is the primary aim of the study?

   The primary aim as stated is to assess the effect of thricholorethene (C_2HCl_3) and tetrachloroethene (C_2Cl_4) on the occurrence of kidney cancer. This is based on a described possible biological mechanism.

2. How were the cases sampled?

   Cases in the study were patients who underwent surgery between 1987-12-01 and 1992-05-01, a period of 3.5 years. Of the 78 patients 62 responded (or their next of kin, since 4 of these 62 were dead)

3. How were the controls sampled?

   Controls were sampled from the accident wards of three *other* hospitals in the same area, form the period 1993-01-01 to 1993-12-31.

4. Are they comparable; i.e. what assumptions are needed?

   In order to deem cases and controls comparable we must assume:

   (a) In principle we must assume that the controls would have been enrolled as cases *if they had had a diagnosis of kidney cancer.*

   In practice it will suffice that the controls comes from a comparable populations, that is even though they might have been subjected to surgery at a different hospital in case of kidney cancer, if this were a similar type of hospital it would have been sufficient.

   (b) Exposure distribution in the population has not changed between the case-selection period (1987–1992) and the control sampling period (1993).

   It is of course difficult to assess this without further knowledge about the industrial development in Nordrhein-Westphalia in the period 1987–1993.

   (c) The attendance of the accident ward is unrelated to the exposure of interest.

   At face value it seems so, bar of course accidents related to trichloroethene exposure itself, but we can safely assume that they are very rare compared to

other. However, the attendance at the accident ward is hardly unrelated to age, and neither is exposure to to trichloroethene, so the attendance *is* related to the exposure, albeit not directly.

(d) Recall of exposure is the same among cases and controls.

It is not stated anywhere whether the purpose of the study was revealed to cases and controls, but in the likely event that it were, cases may be more prone to recall exposure to the thricholorethene in order to get an explanation of their (severe) disease.

As we see there are a few potential biases, some tend to *increase* the risk estimate; the problems with age is difficult to assess.

5. What is the (actual) study base?

   There is no definite answer to this, but if we define the study base to be the state of Nordrhein-Westphalia, the validity of the study hinges on assumptions that the hospitals where cases and controls are sampled are representative of the population — with respect to the exposure of interest.

   Of course, precisely the same argument applies if we define the study base to some subset of Nordrhein-Westphalia, which includes the uptake-areas of the 4 hospitals, and incidentally we might define the study base as the uptake area of these 4 hospitals.

6. What study base is the intended (for generalization)?

   The intended study base is presumably all industrialized counties comparable to Germany.

7. Is the sampling scheme incidence density sampling?

   The sampling scheme is certainly not incidence density sampling, that would require that cases were selected among those at risk at the time of the case-occurrence, ad that is not the case.

8. Can the age-effect on the occurrence of renal cancer be estimated?

   The age-effect cannot be estimated; that would require that the controls were a representative sample (w.r.t. age) of the study base. Sampling persons from the accident wards makes this a far-fetched assumption.

9. Is age a confounder?

   Most likely so; it is definitely associated with both the exposure and the outcome.

10. Key in the numbers in table 6 (p.380), and verify the analysis using Stata.

    When we enter the data from table 6 there are a few things to be aware of:

    - Unlike previously, we are entering *grouped* or `tabulated` data from a table, not individual records. This means that every line in the input dataset represents as (usually large) number of individuals.

- This in turn means that we shall need one line in the data set per entry in the table. In addition to the number, the line must then contain variables that tells us whether it is cases/controls, exposed or not, and what age.
- Note that the entries in table 6 in the paper are not the number of cases and controls, but the number of cases and *total* number in the strata. So we need to do a bit of subtraction to get the numbers right.
- Since tabular data are often quite short, it is most convenient to include the data in the program code itself. This is done by listing the data in the program starting with `input` and terminating data with an `end`.
- In the analysis we need the use the option `[fweight=n]` (frequenct `weights`) in order to tell Stata that the variable `n` represents the number of observations.

The Stata-program, `Renal.do` can be found in the folder http://bendixcarstensen.com/EpiE2012/stata/, and the log is printed at the end of this section.

11. Is there any evidence of heterogeneity of the odds-ratio across age-classes? (*Hint:* Use the Breslow-Day-test.)

    There is no evidence of age-heterogeneity, but the age-classes are quite crude (10-years).

12. In particular, how does the odds-ratio estimate given by Vamvakas *et al.* compare to the Mantel-Haenszel estimate based on the same data?

    The estimate given in the paper is somewhat smaller than the MH-estimate, but the substantial message is pretty much the same: definitely an effect, but unclear how large.

13. What is the main result (in plain words)?

    Based on this study alone, there seems to be an excess risk of kidney cancer associated with trichloroethene exposure, but its cannot really be determined whether the OR is 2.5 or 25. This is of course due to the rather small sample size.

```
*FIRST WE NEED TO ENTER THE DATA
* ONE OBSERVATION PER LINE
* AGE GROUP, EXPOSURE, CASE/CONTROL AND N
* CASE MUST BE 1 and CONTROL 0
* EXPOSED=1 AND UNEXPOSED=0
clear
  input age tri ck n
   30   1   1   2
   40   1   1   2
   50   1   1  10
   60   1   1   1
   70   1   1   4
   30   0   1   0
   40   0   1   1
   50   0   1  12
   60   0   1  17
   70   0   1   9
   30   1   0   1
   40   1   0   4
   50   1   0   2
   60   1   0   0
   70   1   0   0
   30   0   0  21
   40   0   0  11
   50   0   0  25
   60   0   0  14
   70   0   0   6
 end
```

```
 *CRUDE ODDS RATIO NOT ADJUSTING FOR AGE
 * MUST REMEMBER THAT EACH LINE REPRESENTS N INDIVIDUALS

cc ck tri [fweight=n]

 *ADJUSTING FOR AGE GROUP

cc ck tri [fweight=n], by(age) bd

*STILL MH ODDS SLIGHTLY DIFFERENT METHOD FOR CONFIDENCE INTERVALS
mhodds ck tri [fweight=n], by(age)
```

4.3 IHD data from Clayton & Hills.

The study is described by Clayton & Hills, Ch. 13. The tabulated data set of counts of IHD cases and person-years is avaliable from `www` in the file `ihd-tab.txt`.

1. Fit the model from Clayton & Hills Tables 22.7-8 (p.222) and perform the tests from exercises 24.1 and 24.2 (pp.237–238).
2. Fit the model with interaction and re-find results from Clayton & Hills Table 24.5 (p.242) and the test for no interaction.

These two tasks are completed in the Stata-program listed below:

```
use "http://www.biostat.ku.dk/~pka/spss-stata-data/ihd-tab.dta", clear

*FIT MODEL FROM CLAYTON AND HILLS P. 222

*ON THE LOG SCALE
poisson cases i.exposure i.age, exposure(pyrs)

*REPORT RATE RAIOS
poisson cases i.exposure i.age, exposure(pyrs) irr

*WALD TEST FOR NO EFFECT OF EXPOSURE ON IHD
testparm i.exposure
*THE EFFECT OF EXPOSURE IS STATISTICALLY SIGNIFICANT

*LIKELIHOOD RATIO TEST FOR THE EFFECT OF EXPOSURE ON IHD
*THE MODEL WITH MOST PARAMETERS HAS HIGHEST LOG L
*WE DO NOT GET THE SAME LOG L AS IN THE BOOK BUT THE SAME DIFFERENCE SO
*THERE IS A CONSTANT LEFT OUT.

poisson cases i.exposure i.age, exposure(pyrs)
est store mod1

poisson cases i.age, exposure(pyrs)
est store mod2

lrtest mod1 mod2

*INTERACTIONS
poisson cases i.exposure i.age i.exposure#i.age, exposure(pyrs)
est store mod3

*TEST INTERACTION
lrtest mod1 mod3
*INTERACTION NOT SIGNIFICANT

*EFFECT OF EXPOSURE IN AGE GROUPS
poisson cases i.age i.exposure#i.age, exposure(pyrs)
```

4.3.1 Using continuous variables

The IHD-data contains energy consumption as a continuous variable, `energy`, among other variables.

The questions were:

1. Read the individual diet data records from the file.
2. Create variables for the person-years, by subtracting entry date from date of exit. Also create a variable with the log-person-years.
3. Use CHD as outcome variable in a Poisson-analysis with the log-person-years as offset, using energy as a linear explanatory variable. Is there an effect on mortality?
4. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (approx. the quartiles)
5. Same question for weight and BMI (the latter you have to calculate yourself as weight/height^2).

— and they are addressed in the following Stata-program:

4.3.2 Splitting the follow-up of the IHD-data

The exercise gives quite a thorough introduction to the exercise, sor there is no need to repeat that here. The analyses can be done using the Stata-program `ihd-Lexis-sol.do` which is listed here:

```
clear

input eksp agr pyrs cases
1 0  346.87  2
1 1  979.34 12
1 2  699.14 14
0 0  560.13  4
0 1 1127.70  6
0 2  794.15  8
end

*QUESTION 1
*FIT POISSON MODEL WITH EXPOSURE AND AGE EFFECT

*ON THE LOG SCALE
poisson cases i.eksp i.agr , exposure(pyrs)

*QUESTION 2
*READ INDIVIDUAL DATA

infile id str10 doe str10 dox chd str10 dob job month energy height weight fat fibre /*
*/ in 2/L using "C:\ewan\Epidemiologi\data\diet.txt", clear

*GET THE DATES INTO DATE FORMAT

gen date_entry= date(doe,"MDY")
gen date_exit= date(dox, "MDY")
gen date_birth= date(dob, "MDY")

format date_entry date_exit date_birth %td

*QUESTION 3
*TELL STATA THAT THIS IS SURVIVAL DATA

stset date_exit, failure(chd==1) entry(date_entry) origin(date_birth) /*
*/ scale(365.25) id(id)

*HOW MANY OBSERVATIONS ?
display _N

*QUESTION 4
*SPLIT THE FOLLOW UP INTO AGE GROUPS 40-50, 50-60, 60-70
*GENERATE A NEW VARIABLE CALLED CURRENT AGE

stsplit current_age, at(40(10)70) after(date_birth)

*QUESTION 5
*HOW MANY OBSERVATIONS?

display _N

*QUESTION 6
```

```
*HAVE A LOOK AT THE DATA

browse

*QUESTION 7
*TABULATE IHD CASES AND PERSON YEARS BY EXPOSURE AND AGE GROUP

gen pyrs=_t-_t0
gen exposure=1 if energy < 2.75
replace exposure=0 if energy >= 2.75 & energy!=.

*ONLY COUNT CHD CASES ONCE
gen event=_d

table exposure current_age, c(sum event sum pyrs) format(%9.2f)
*CURRENT_AGE=0 IS ALL FOLLOW UP BEFORE AGE 40

*QUESTION 8
*ESTIMATE EFFECT OF CURRENT_AGE AND EXPOSURE
*DROP FOLLOW_UP BEFORE AGE 40

keep if current_age>0

poisson event i.exposure i.current_age, exposure(pyrs)
*THE SAME RESULT AS WITH THE TABULATED DATA

*QUESTION 9
*INTERACTION BETWEEN AGE AND EXPOSURE
poisson event i.exposure i.current_age i.exposure#i.current_age, exposure(pyrs)
testparm i.exposure#i.current_age
est store m1

poisson event i.exposure i.current_age , exposure(pyrs)
est store m2
lrtest m1 m2

*THE GROUPED DATA
collapse (sum) pyrs event , by(exposure current_age)
poisson event i.exposure i.current_age , exposure(pyrs)
```

4.4 Case-control study of BCG vaccination and leprosy.

The study is decribed by Clayton & Hills, p.156. In short, 260 cases of leprosy among individuals aged less than 35 years were ascertained in a study area in Malawi. Subjects were grouped into 7 age intervals and according to absence or presence of a scar after BCG vaccination. Three sets of controls were studied:

1. a population survey of 80,622 persons
2. a random sample of 1000 persons
3. a 4 to 1 age-matched sample

1. Fit the model from Clayton & Hills Table 23.5 (p.232).
2. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference.
3. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference.
4. Test the hypothesis of no interaction between `age` and `scar`.

5. Analyse the data set with only 1000 controls (i.e., use the controls `con1000`: Table 23.6, p.233) and compare the precision of the estimate for `scar` with that based on the entire sample.

6. Analyse the matched data set (i.e., use the controls `conmatch`: Table 23.6, p.233) and compare with the results from Table 23.7.

7. Try (erroneously) to drop `age` from the analysis of the matched data and study the consequences for the estimate of `scar`.

The following Stata program reads the data and produces the solutions to these tasks:

4.5 Case-control study of malignant melanoma.

Anne Østerlind conducted in the middle of the 80's a case-control study of risk factors for malignant melanoma in Denmark.

The review paper "Malignant melanoma in Denmark" from *Acta Oncologica*, 1990[3] is from Anne Østerlind's thesis and gives an overview of the results from the study which included 1400 interviewed persons, 474 cases and 926 controls, cf. table 5 in the article.

In the article incidence changes between 1943 and 1982 are also discussed; that part of the paper will not be touched upon in this exercise.

4.5.1 Discussion of the article.

1. Explain the design, the data base and data collection, particularly how the matching was conducted.

   The study is a group-matched (stratified) case-control study. Based on knowledge of the age- and sex-distribution of melanoma cases in Denmark a sample of the population with the same age- and sex-distribution was requested from the CPR.

   The data base (study base) is persons living in the eastern part of Denmark.

   Data were collected by personal interview.

2. How were interviews planned to minimize bias?

   Interviewers were blinded to the case/control status of the persons they interviewed for the study.

3. Explain the drop-out, particularly the analyses in Tables 5-7. What are the consequences of these results for the subsequent analyses?

   The drop-outs are differential between age-classes, older seem to be less likely to respond. It would have been more informative to have had the response probabilities for each subcategory, instead of the relative distribution separately for responders and non-responders.

4. How are the analyses carried out? Are all variables included in one step or are the analyses conducted in smaller steps? How are the matching variables accounted for? Comments?

Analyses are conducted first with one variable at a time, giving marginal results, and then jointly to give mutually controlled results.

Sex, but not age, is included in the controlled analyses, so strictly speaking there is a potential bias. Particularly if confounding is expected, that is if the distribution of the risk factors are different across age-classes. This is however not particularly likely.

5. Explain the analyses presented in Table 9. How many logistic regression models are fitted here?

   Table 9 contains 6 analyses of one variable and 1 analysis with all variables, that is a total of 7 analyses.

6. What is the conclusion from the analyses in the table?

   Table 9 seems to indicate that freckles and to some extent skin and hair colour are the major (non-modifiable) risk factors for melanoma of the skin.

7. What is the purpose of Table 11?

   Table 11 shows the joint effect of freckles and naevi on the risk of melanoma. Clearly the risk increases by both, but there is no assessment of whether an interaction is present or not. Incidentally, as you will see, there is none.

8. Which modifiable factors seem to affect the melanoma risk?

   Number of sunburns before age 15. So one would in general expect that sunlight exposure increases the occurrence of melanoma. However detailed assessment of exposure history for each piece of skin for each study participant is of course impossible. So this seems to be pretty much as good as it gets.

4.5.2 Melanoma data

The Stata-program listed below reads data and does the analyses requested in the questions. Comments explains what question the code relates to.

4.5.3 Simple tabulation analysis

1. Make the two by two table showing the association between case-control status and whether or not the person experienced *any* sunburns before the age of 15. `SAS`-users may use the program `melanom.sas` to read in the data from `www`. Estimate the odds ratio with associated 95% confidence limits and test for no association between the risk factor and case-control status.

2. Conduct similar analyses for the factors `sex, hair, eyes, freckles, acuterea, chronrea`. Compare with Table 9 in the article.

3. The case control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables. Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

4. Same question for age.

These analyses can be accomplished using the following SAS program:

```
. use "C:\ewan\Epidemiologi\melanom.dta", clear
.
. *QUESTION 2
. *ANY SUNBURNS BEFORE AGE 15
.
. gen Any_burn15=1 if burn15>0 & burn15!=.
. replace Any_burn15=0 if burn15==0
.
. *OR OF CASE IF ANY BURN BEFORE AGE 15
. cc case Any_burn15
.
. *OR=2.07 (1.50; 2.85) IS SIGNIFICANT
.
. *QUESTION 3
. *TO USE cc WE MUST HAVE EXPOSED=1 AND UNEXPOSD=0
.
. gen woman=sex-1
.
. cc case woman

. * NOT SIGNIFICANT EFFECT OF SEX
.
. logistic case i.hair
.
. logistic case i.eyes
.
. *I WANT GROUP 3=NONE AS REFERENCE
. logistic case b3.freckles
.
. *GROUP 4=NO SUNBURN AS REFERENCE
. logistic case b4.acuterea
.
. logistic case i.chronrea
.
. *QUESTION 4
. *ADJUSTING FOR SEX
.
. logistic case i.Any_burn15 i.sex

. *STILL A SIGNIFICANT EFFECT OF SUNBURN
.
. *QUESTION 5
. *ADJUSTING FOR AGE
.
. logistic case i.Any_burn15 i.agroup

. *STILL A SIGNIFICANT EFFECT OF SUNBURN
```

4.5.4 Introductory analyses.

1. Estimate (log-)odds ratios for the variable `skin` (see top left in AØ's Table 9). SAS-users may use the program `melanom.sas`.
2. Estimate also odds ratios (in SAS: use `ESTIMATE` statements).
3. Conduct the other analyses in AØ's Table 9 (*left* part) where the factors `hair, eyes, freckles, acuterea, chronrea` are studied one at a time.
4. Conduct the analysis corresponding to Table 9 (*right* part) where several variables are included simultaneously (see the table footnote).
5. Reconstruct the results from AØ's Table 10 concerning number of raised naevi.

   NB: new variables must be defined from the original variables `nvtot, nvsmall, nvlarge`.

4.5.5 Trend tests and interactions.

6. In the analyses so far all variables have been considered as categorical ('`class`' in SAS) variables while all tests in Tables 9 and 10 are trend tests. Conduct the analyses which give the P-values in Table 9 (right part) for the variables `skin` and `freckles`.

7. May `freckles` be scored linearly (1, 2, 3), when this variable is studied separately? (Conduct a test for linearity/departures from trend).

8. In AØ's Table 11 `freckles` and the total number of naevi (suitably grouped) are studied. Conduct this analysis. Is there any interaction between these two variables?

9. Study, in a similar vein, interactions between `acuterea` and `skin` and between the grouped version of `nnvtot` from question 5. and `agroup`.

10. All of AØ's analyses are conducted without accounting for the match variable age (`agroup`) (in spite of warnings given by Clayton & Hills!). Repeat some of the previous analyses adjusting for `agroup`. Are there any substantial differences? Explain!

These analyses can be accomplished using the following SAS program:

```
use "C:\ewan\Epidemiologi\data\melanom.dta", clear

*QUESTION 1
* LOG-ODDS FOR SKIN

logit case i.skin

*QUESTION 2
* ODDS FOR SKIN

logistic case i.skin

*QUESTION 3

logistic case i.hair

logistic case i.eyes

*I WANT GROUP 3=NONE AS REFERENCE
logistic case b3.freckles

*GROUP 4=NO SUNBURN AS REFERENCE
logistic case b4.acuterea

logistic case i.chronrea

*QUESTION 4
*SEVERAL VARIBALES

logistic case i.skin i.hair i.eyes b3.freckles b4.acuterea i.chronrea i.sex

*QUESTION 5

gen nvtot_grp=0 if nvtot==0
replace nvtot_grp=1 if nvtot==1
replace nvtot_grp=2 if nvtot>=2 & nvtot<=4
replace nvtot_grp=3 if nvtot>=5 & nvtot!=.

gen nvsmall_grp=0 if nvsmall==0
replace nvsmall_grp=1 if nvsmall==1
replace nvsmall_grp=2 if nvsmall>=2 & nvsmall<=4
replace nvsmall_grp=3 if nvsmall>=5 & nvsmall!=.

gen nvlarge_grp=0 if nvlarge==0
replace nvlarge_grp=1 if nvlarge==1
replace nvlarge_grp=2 if nvlarge>=2 & nvlarge!=.

logistic case i.nvtot_grp

logistic case i.nvsmall_grp

logistic case i.nvlarge_grp

logistic case i.nvtot_grp i.sex i.freckles i.hair i.skin

logistic case i.nvsmall_grp i.sex i.freckles i.hair i.skin

logistic case i.nvlarge_grp i.sex i.freckles i.hair i.skin


*************************************
* 2.5.5 TREND TESTS AND INTERACTIONS  *
```

```
**************************************

*QUESTION 6
*TREND TEST

logistic case skin
*WALD TEST p=0.002

logistic case freckles

*QUESTION 7
*LINEAR MODEL FOR FRECKLES

logistic case b3.freckles
est store m1

logistic case freckles
est store m2
lrtest m1 m2
*ACCEPTED

*QUESTION 8
*INTERACTION BEWTEEN FRECKLES AND nvtot_grp

logistic case b3.freckles i.nvtot_grp b3.freckles#i.nvtot_grp
est store m1

logistic case b3.freckles i.nvtot_grp
est store m2
lrtest m1 m2
*INTERACTION NOT SIGNIFICANT p=0.5071

*GETTING THE ESTIMATES FROM TABLE 11

logistic case b3.freckles#b0.nvtot_grp , base

*OR FOR FRECKLES ADJUSTED FOR NVTOT AND OR FOR NVTOT ADJUSTED FOR FRECKLES
logistic case b3.freckles b0.nvtot_grp

*TREND TEST FOR FRECKLES (WALD)
logistic case freckles b0..nvtot_grp

*TREND TEST FOR NVTOT_GRP (WALD)
logistic case b3.freckles nvtot_grp

*NUMBER OF CASES AND CONTROLS BY FRECKLES AND NVTOT_GRP
gen control=1-casecon
table freckles nvtot_grp, c(sum casecon sum control)

table freckles if nvtot_grp!=., c(sum casecon sum control)

table nvtot_grp if freckles!=., c(sum casecon sum control)

*QUESTION 9

logistic case b4.acuterea i.skin b4.acuterea#i.skin
est store m1

logistic case b4.acuterea i.skin
est store m2
lrtest m1 m2
*THE INTERACTION IS STATISTICALLY SIGNIFICANT p=0.0112

logistic case  b4.acuterea#b0.skin, base

*OR FOR ACUTREA ADJUSTED FOR SKIN AND OR FOR SKIN ADJUSTED FOR ACUTREA
logistic case b4.acuterea i.skin , base

*NUMBER OF CASES AND CONTROLS BY ACUTREA AND SKIN
table acuterea skin, c(sum casecon sum control)
*NOT MANY IN GROUP WITH ACUTREA=1

table acuterea if skin!=., c(sum casecon sum control)

table  skin if acuterea!=., c(sum casecon sum control)

* NVTOT AND AGROUP
logistic case i.nvtot_grp i.agroup i.nvtot_grp#i.agroup
testparm i.nvtot_grp#i.agroup
est store m1

*SIX OBSERVATIONS NOT USED WHEN INTERACTION LRTEST NEEDS SAME DATA
logistic case i.nvtot_grp i.agroup if e(sample)
est store m2
lrtest m1 m2
*NOT SIGNIFICANT


*QUESTION 10
```

```
*ADJUST FOR AGE

logistic case i.skin i.agroup

logistic case b4.acuterea i.agroup

*DOES NOT MAKE MUCH DIFFERENCE
*AGE DOES NOT INFLUENCE SKIN COLOUR OR ACUT REACTION SO NOT CONFOUNDER
```

4.6 Testicular cancer risk and maternal parity.

This exercise deals with the article "Testicular cancer risk and maternal parity: a population-based cohort study", by T. Westergaard, P.K. Andersen, J.B. Pedersen, M. Frisch, J.H. Olsen, M. Melbye. *Br. J. Cancer*, **77**,pp. 1180-1185 (1998). [4].

4.6.1 Discussion of the article.

1. What is the authors' argument for the existence of an effect of maternal parity on the risk of testicular cancer in the son?

   Mainly the existing literature on occurrence of testis cancer.

2. Describe the design of the study:

   - a. which "sons" are included in the study?

     Sons of mothers born after 1935; that is essentially only boys born after 1950. But not all such boys.

   - b. when are they followed?

     They are followed from the start of CPR, 1.1.1968 till the end of 1992.

   - c. how are cases defined and ascertained?

     Cases are defined as cases reported to the Danish Cancer registry with certain diagnoses.

3. Concentrating on all testicular cancers, what do you consider to be the main result reported in Table 1?

   The effect of parity 2+, `and` the remarkable small effect of any other variable.

4. Explain in words the interpretation of the value RR=0.80 for parity 2+.

   This means that the incidence rate of TC among boys born as the 2nd or later sibling of a woman is 0.8 of the rate among first-born.

5. Compare this value with the corresponding crude RR (and 95 % CI) obtained without any adjustment. Explain the differences between the two results.

   The crude RR was 0.57, but adjustment for the age at follow-up (called `SON_AGE` in the dataset), makes this effect go to to 0.8. This is because of confounding — incidence increases by age, and first born are likely to be older in a dataset collected as follow-up of cohorts.

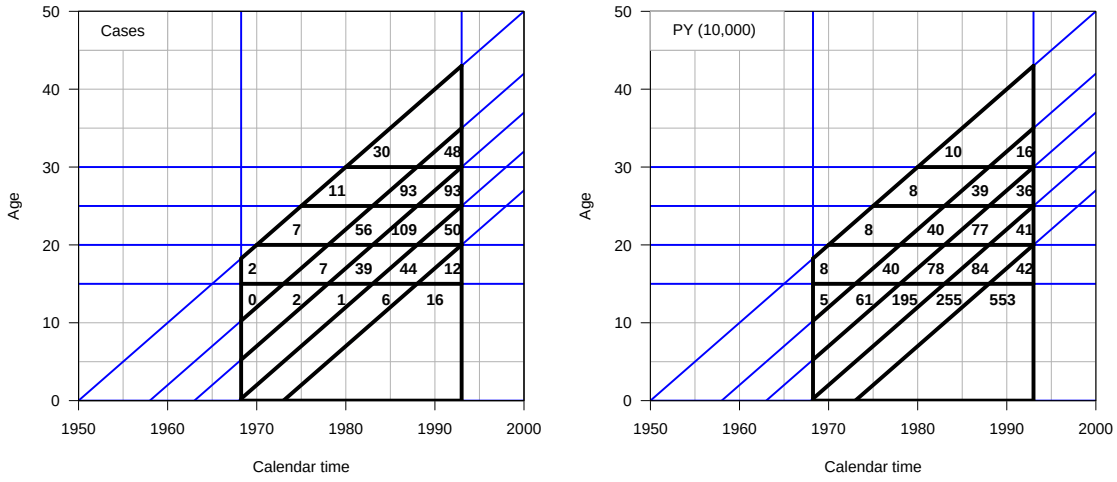


Figure 4.1: *Lexis diagrams showing the number of cases and person-years by age and date of follow-up. The black outline indicates the sampling frame for the study.*

6. Draw a Lexis diagram to illustrate the combinations of age and calendar period which contribute person-years to the study. This is shown in the figure.

   In the section with solutions in R is a small section illustration how to draw the Lexis diagrams with lines indicating age-classes and birth cohorts.

7. Explain the meaning of the estimates for "Interval from ..." in the lower part of Table 1.

   This interval basically gives the age-difference to the closest older sibling. This is meant as a proxy fro some sort of infection pressure in infancy.

8. What type of analysis is reported in Table 2?

   Table 2 reports *interactions*, investigating whether the effect of being 1st born is constant across other factor available in this study.

9. Discuss how, alternatively, a case-control design could have been conducted to address the same question as the cohort study reported in the article.

   A case-control study would include as cases the same cases as this cohort study, namely those ascertained from the Cancer Registry. As there is a substantial variation by age and birth-cohort in testis cancer incidence, which is well known, it would be natural to match each case to a number of controls born on the same day and who survived till the date of diagnosis of the case (for example all).

   The maternal parameters would then be ascertained from the various registers for both cases and controls.

   Even though an individual matching is done this way, the analysis needed not be done as matched (conditional) logistic regression, the matching variables, date of birth and age are perfectly quantifiable and could just be included as covariates in an ordinary logistic regression analysis.

The case-control study would not be able to assess the incidence rates, but in so far the aim of the analysis was to assess the effect of maternal parameters, this would not be of interest. The level and age-dependence of TC incidence rates are well known from descriptive studies based on the cancer registry data anyway.

4.6.2 Practical exercises

The file `testis.txt`, available at `www` contains for each (non-empty) combination of the factors `SON_AGE, SON_KOH, MOTH_AGE, PARITY` the number of person-years at risk `PYRS`, the numbers of non-seminomas and seminomas, respectively `NONSEMI SEMI`, and the total number of testis cancer cases `CASES`. The first line of the file contains the variable names.

The Stata-program `Testis-sol.do` listed below reads the data and does the analyses required to answer the practical questions.

10. Compute the crude rate ratio for testis cancer for parity 2+ versus parity 1. Compare with 5. above.

11. Reconstruct the estimates for "parity of mother at birth of son" from the top of Table 1 in the article both for all testis cancers and for non-seminomas.

12. Reconstruct the estimates from Table 2 in the article concerning mother's age (for all testis cancers). Is there an interaction between parity and mother's age?

13. Same question for birth cohort of the son.

```
use "C:\ewan\Epidemiologi\Data\testis.dta", clear

* CRUDE RATE RATE RATIO FOR PARITY 2+

*QUESTION 10
poisson cases i.par2, exposure(pyrs) irr

*COMPARE TO QUESTION 5
*FROM TABLE 2 OF PAPER WE GET PERSON YEARS AND CASES
*228 EXPOSED CASES 8009691 PYRS
*398 UNEXPOSED CASES 7972276 PYRS

iri 228 398  8009691 7972276
*THE SAME RATE RATIO AS ABOVE

*QUESTION 11
*ESTIMATES FOR PARITY OF MOTHER

*ALL CASES
table parity, c(sum cases sum pyrs) format(%9.0f)

poisson cases i.parity i.moth_age i.son_koh i.son_age, exposure(pyrs) irr

*TREND TEST
poisson cases parity i.moth_age i.son_koh i.son_age, exposure(pyrs) irr
est store m1

poisson cases i.moth_age i.son_koh i.son_age, exposure(pyrs) irr
est store m2
lrtest m1 m2

*NON-SEMINOMAS
table parity, c(sum nonsemi sum pyrs) format(%9.0f)

poisson nonsemi i.parity i.moth_age i.son_koh i.son_age, exposure(pyrs) irr

*TREND TEST
poisson nonsemi parity i.moth_age i.son_koh i.son_age, exposure(pyrs) irr
est store m1

poisson nonsemi i.moth_age i.son_koh i.son_age, exposure(pyrs) irr
est store m2
lrtest m1 m2
```

```
*QUESTION 12
* AGE OF MOTHER AND INTERACTION WITH PARITY

table moth_age, c(sum cases sum pyrs) format(%9.0f)

poisson cases b20.moth_age i.son_koh i.son_age i.parity, exposure(pyrs) irr base
est store m1

poisson cases b20.moth_age i.son_koh i.son_age i.parity  b20.moth_age#i.parity/*
*/, exposure(pyrs) irr base
est store m2
lrtest m2 m1
*INTERACTION NOT STATISTICALLY SIGNIFICANT

*QUESTION 13
*INTERACTION BETWEEN MOTHER AGE AND BIRTH COHORT OF SON
 poisson cases b20.moth_age i.son_koh i.son_age i.parity  b20.moth_age#i.son_koh/*
*/, exposure(pyrs) irr base
est store m2
lrtest m2 m1
*INTERACTION NOT STATISTICALLY SIGNIFICANT
```

Chapter 5

Solutions with R

The R-programs are available on the course web site in the folder
http://BendixCarstensen.com/EpiF2013/r. There is also a link to this on the website.

R

R is a free statistics package, which has become the default computing tool a large part of the statisticians of the world. It is dominant in bioinformatics. It is particularly useful for its excellent and versatile graphics. As can be seen from the first few solutions, it can also be used as a mere desk top calculatior.

R comes with a versatil documentation system for analyses and results, Rweave, which is used for these solutions. It provides a way of documention reproducible research, which is widely used, particularly in bioinformatics.

R can be expanded by downloading additional packages, of which there are currently about 3000. The relevant site is CRAN, the Comprehensive R Archive Network, http://cran.r-project.org/.

5.1 Vaccinations and childhood mortality in Guinea-Bissau

In rural Guinea-Bissau, 5274 children under 7 months of age were visited two times at home with an interval of six months (Kristensen et al., *BMJ*, 2000, [2]). Information about vaccination (BCG, DTP, measles vaccine) was collected at each visit, and at the second visit, death during follow-up was registered. Some children moved away during follow-up, some survived until the next visit. The following variables are found in the data set `bissau.txt`:

`id`	Id number
`fuptime`	Follow-up time in days
`dead`	0 = censored, 1 = dead
`bcg`	1 = Yes, 2 = No
`dtp`	Number of DTP doses (0,1,2,3)
`age`	Age at first visit in days
`agem`	Age at first visit in months

5.1.1 A single risk, odds and rate

The R-program is in http://BendixCarstensen.com/EpiF2013/sasas `bissau-sol0.R`.

5.1.2 A single risk, odds and rate

Reading and Tabulating the dataset:

```
> bis <- read.table( "../data/bissau.txt", header=TRUE )
> N <- nrow( bis )
> D <- sum( bis$dead )
> Y <- sum( bis$fuptime/365.25 )
> cbind( N, D, Y )
```

```
        N   D        Y
[1,] 5274 222 2409.821
```

shows the number of children is 5274, the number of deaths 222 and the number of person-years 2409.8 (namely 880187 days)

- Following the lectures we get

  1. The overall risk of death is 222/5274=4.21%. A naive 95% confidence interval for this is:

     $$p \pm 1.96\sqrt{p \times (1-p)/n} = 0.0421 \pm 1.96\sqrt{0.0421 \times 0.9579/5274} = (0.0367; 0.0475),$$

     ```
     > p <- D/N
     > se <- sqrt( p*(1-p)/N )
     > round( cbind( p, lo=p-1.96*se, hi=p+1.96*se ), 4 )
     ```

     ```
               p     lo     hi
     [1,] 0.0421 0.0367 0.0475
     ```

     But a better one is the formula:

     $$\frac{p}{p + (1-p) \overset{\times}{\div} \mathrm{erf}}, \quad \mathrm{erf} = \exp\left(1.96\sqrt{1/x + 1/(n-x)}\right)$$

     Which gives:

     $$\mathrm{erf} = \exp(1.96\sqrt{1/222 + 1/5052} = 1.144$$

     and so the c.i.:

     $$\frac{0.0421}{0.0421 + 0.9579 \overset{\times}{\div} 1.144} = (0.0370; 0.0479)$$

     ```
     > erf <- exp( 1.96*sqrt(1/D+1/(N-D)) )
     > round( cbind( p, lo=p/(p+(1-p)*erf),
     +                  hi=p/(p+(1-p)/erf) ), 4 )
     ```

     ```
               p    lo     hi
     [1,] 0.0421 0.037 0.0479
     ```

2. The overall odds of death is simply:

$$\frac{222}{5274-222} = 0.0439$$

and the s.e. on the log-scale is is used to compute the 95% c.i., it is the same error factor as before:

$$\text{erf} = \exp\left(1.96\sqrt{1/222 + 1/5052}\right) = 1.144$$

so we get:

$$0.0439 \overset{\times}{\div} 1.144 = c(0.0384, 0.0503)$$

```
> odds <- D/(N-D)
> round( cbind( odds, lo=odds/erf, hi=odds*erf ), 4 )

       odds     lo     hi
[1,] 0.0439 0.0384 0.0503
```

3. The overall *rate* of death (per year) is

$$222/2409.8 = 0.0921$$

and the error factor is $\exp(1.96/sqrtD) = 1.141$ (with $D = 222$), so the confidence interval is:

$$0.0921 \overset{\times}{\div} 1.141 = (0.0807, 0.1050)$$

```
> rate <- D/Y
> erf <- exp(1.96/sqrt(D))
> round( cbind( rate, lo=rate/erf, hi=rate*erf ), 4 )

       rate     lo     hi
[1,] 0.0921 0.0808 0.1051
```

- Using the modelling we can get the same. Although it seems like bringing coal to Newcastle, there is sense in this, because we get some code which is generalizable:

1. A single proportion can be modelled in a binomial model with log-link and subsequently using `ci.exp` to fish out and exponentiate the result. We can either use the tabulated numbers or the entire data set. Note that when we use tabulated data we must put the response in as a two-column matrix with dead and non-dead, using `cbind`:

```
> library( Epi )
> summary( m0 <-glm( cbind(D,N-D) ~ 1, family=binomial(link=log) ) )

Call:
glm(formula = cbind(D, N - D) ~ 1, family = binomial(link = log))

Deviance Residuals:
[1]  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.16787    0.06569  -48.23   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 0  degrees of freedom
Residual deviance: 9.8588e-14  on 0  degrees of freedom
AIC: 9.1983

Number of Fisher Scoring iterations: 3
```

```
> round( ci.exp( m0 ), 4 )

            exp(Est.)  2.5%  97.5%
(Intercept)    0.0421 0.037 0.0479

> summary( l0 <-glm( dead ~ 1, family=binomial(link=log), data=bis ) )

Call:
glm(formula = dead ~ 1, family = binomial(link = log), data = bis)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2933  -0.2933  -0.2933  -0.2933   2.5171

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.16787    0.06568  -48.23   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1841.1  on 5273  degrees of freedom
Residual deviance: 1841.1  on 5273  degrees of freedom
AIC: 1843.1

Number of Fisher Scoring iterations: 6

> round( ci.exp( l0 ), 4 )

            exp(Est.)  2.5%  97.5%
(Intercept)    0.0421 0.037 0.0479
```

2. The same goes for the odds, now we just use the default link function which is the logit, and so exponentiation of the estimate (the intercept) will be the odds:

```
> summary( m1 <-glm( cbind(D,N-D) ~ 1, family=binomial ) )

Call:
glm(formula = cbind(D, N - D) ~ 1, family = binomial)

Deviance Residuals:
[1]  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.12486    0.06857  -45.57   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 0  degrees of freedom
Residual deviance: 4.3165e-13  on 0  degrees of freedom
AIC: 9.1983

Number of Fisher Scoring iterations: 2

> round( ci.exp( m1 ), 4 )

            exp(Est.)   2.5%  97.5%
(Intercept)    0.0439 0.0384 0.0503

> summary( l1 <-glm( dead ~ 1, family=binomial, data=bis ) )

Call:
glm(formula = dead ~ 1, family = binomial, data = bis)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2933  -0.2933  -0.2933  -0.2933   2.5171
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.12486    0.06857  -45.57   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1841.1  on 5273  degrees of freedom
Residual deviance: 1841.1  on 5273  degrees of freedom
AIC: 1843.1

Number of Fisher Scoring iterations: 6

> round( ci.exp( l1 ), 4 )

            exp(Est.)   2.5%  97.5%
(Intercept)    0.0439 0.0384 0.0503
```

3. The likelihood for a constant rate looks like a likelihood for a poisson variate, so we can use the Poisson family to estimate a single rate:

```
> summary( m2 <-glm( D ~ 1, family=poisson, offset=log(Y) ) )

Call:
glm(formula = D ~ 1, family = poisson, offset = log(Y))

Deviance Residuals:
[1]  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38463    0.06712  -35.53   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.3767e-14  on 0  degrees of freedom
Residual deviance: 1.3767e-14  on 0  degrees of freedom
AIC: 9.2413

Number of Fisher Scoring iterations: 2

> round( ci.exp( m2 ), 4 )

            exp(Est.)   2.5%  97.5%
(Intercept)    0.0921 0.0808 0.1051

> summary( l2 <-glm( dead ~ 1, family=poisson,
+                              offset=log(fuptime/365.25), data=bis ) )

Call:
glm(formula = dead ~ 1, family = poisson, data = bis, offset = log(fuptime/365.25))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3038  -0.3038  -0.3030  -0.2850   3.3151

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.38463    0.06712  -35.53   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1759.2  on 5273  degrees of freedom
Residual deviance: 1759.2  on 5273  degrees of freedom
AIC: 2205.2

Number of Fisher Scoring iterations: 6

> round( ci.exp( l2 ), 4 )
```

```
             exp(Est.)   2.5%  97.5%
(Intercept)    0.0921 0.0808 0.1051
```

So we see we get the same by using the `glm` function and the classical formulae. But the `glm` machinery is easier to generalize than the classical formulae.

In the following is shown the R-commands to do all the calculations required in the questions.

5.1.3 Rates, risks and odds

1. For convenience we first load the Epi package, and the read the data — including the variable names from the first line:

```
> library( Epi )
> library( epitools )
> bissau <- read.table( "../data/bissau.txt",
+                       header=TRUE )
> str( bissau )
```

```
'data.frame':        5274 obs. of  7 variables:
 $ id     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ fuptime: int  65 161 166 166 161 161 166 166 166 166 ...
 $ dead   : int  1 0 0 0 0 0 0 0 0 0 ...
 $ bcg    : int  1 1 2 1 1 1 1 1 1 1 ...
 $ dtp    : int  1 2 0 0 0 0 2 1 2 2 ...
 $ age    : int  182 125 69 96 131 26 129 90 119 146 ...
 $ agem   : int  5 4 2 3 4 0 4 2 3 4 ...
```

Then we need the no. onservation, no. of deaths and no. person-years for the two groups defined by `bcg`:

```
> with( subset( bissau, bcg==1 ), c(sum(dead),sum(fuptime),length(dead)) )
```

```
[1]    125 554929   3301
```

```
> with( subset( bissau, bcg==2 ), c(sum(dead),sum(fuptime),length(dead)) )
```

```
[1]     97 325258   1973
```

Alternatively — some would say simpler, some would say more convoluted — we could use `xtabs`:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ bcg, data=bissau ) )
```

```
bcg   dead fuptime      n
  1    125  554929   3301
  2     97  325258   1973
```

```
> str( xx )
```

```
 xtabs [1:2, 1:3] 125 97 554929 325258 3301 ...
 - attr(*, "dimnames")=List of 2
  ..$ bcg: chr [1:2] "1" "2"
  ..$    : chr [1:3] "dead" "fuptime" "n"
 - attr(*, "class")= chr [1:2] "xtabs" "table"
 - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ bcg, data = bissau)
```

2. We then compute the fraction of dead and the confidence interval:

```
> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/n)
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )
```

```
          pi      pilow       piup
1 0.03786731 0.03135578 0.04437884
2 0.04916371 0.03962328 0.05870414
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )
```

```
    pi pilow piup
1 3.79  3.14 4.44
2 4.92  3.96 5.87
```

3. We now compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )
```

```
  omega omegalow omegaup
1 0.039    0.033   0.047
2 0.052    0.042   0.063
```

```
> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )
```

```
    pi pilow piup pilow_2 piup_2
1 3.79  3.14 4.44    3.19   4.49
2 4.92  3.96 5.87    4.05   5.96
```

4. Now we compute rate per day - note that we added fuptime in `xtabs` above:

```
> y <- xx[,"fuptime"]
> lambda         <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low     <- lambda / errorfact_rate;
> lambda_up      <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )
```

```
        lambda   lambda_low    lambda_up
1 0.0002252540 0.0001890329 0.0002684156
2 0.0002982248 0.0002444082 0.0003638914
```

5. Rates per day are not interesting, so we convert them to rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )
```

```
  lambda lambda_low lambda_up
1  0.082      0.069     0.098
2  0.109      0.089     0.133
```

6. We now repeat everything using the indicator of whether any DTP was received:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ bcg, data=bissau ) )
```

```
bcg   dead fuptime      n
  1    125  554929   3301
  2     97  325258   1973
```

```
> str( xx )
```

```
 xtabs [1:2, 1:3] 125 97 554929 325258 3301 ...
 - attr(*, "dimnames")=List of 2
  ..$ bcg: chr [1:2] "1" "2"
  ..$    : chr [1:3] "dead" "fuptime" "n"
 - attr(*, "class")= chr [1:2] "xtabs" "table"
 - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ bcg, data = bissau)
```

7. We then compute the fraction of dead and the confidence interval:

```
> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/n)
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )
```

```
          pi      pilow       piup
1 0.03786731 0.03135578 0.04437884
2 0.04916371 0.03962328 0.05870414
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )
```

```
   pi pilow piup
1 3.79  3.14 4.44
2 4.92  3.96 5.87
```

8. We now compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )
```

```
  omega omegalow omegaup
1 0.039    0.033   0.047
2 0.052    0.042   0.063
```

```
> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )
```

```
    pi pilow piup pilow_2 piup_2
1 3.79  3.14 4.44    3.19   4.49
2 4.92  3.96 5.87    4.05   5.96
```

9. Now we compute rate per day - note that we added fuptime in `xtabs` above:

```
> y <- xx[,"fuptime"]
> lambda         <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low     <- lambda / errorfact_rate;
> lambda_up      <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )
```

```
        lambda   lambda_low    lambda_up
1 0.0002252540 0.0001890329 0.0002684156
2 0.0002982248 0.0002444082 0.0003638914
```

10. Rates per day are not interesting, we convert them to rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )
```

```
  lambda lambda_low lambda_up
1  0.082      0.069     0.098
2  0.109      0.089     0.133
```

11. Finally we repeat it al but now subdividing persons by wheter they have received any DTP dose or not:

```
> ( xx <- xtabs( cbind(dead,fuptime,n=1) ~ (dtp>0), data=bissau ) )
```

```
dtp > 0   dead fuptime      n
  FALSE    128  516175   3101
  TRUE      94  364012   2173
```

```
> str( xx )
```

```
 xtabs [1:2, 1:3] 128 94 516175 364012 3101 ...
 - attr(*, "dimnames")=List of 2
  ..$ dtp > 0: chr [1:2] "FALSE" "TRUE"
  ..$        : chr [1:3] "dead" "fuptime" "n"
 - attr(*, "class")= chr [1:2] "xtabs" "table"
 - attr(*, "call")= language xtabs(formula = cbind(dead, fuptime, n = 1) ~ (dtp > 0), data = b
```

```
> d <- xx[,"dead"]
> n <- xx[,"n"]
> pi    <- d / n
> sdpi  <- sqrt(pi*(1-pi)/xx[,"n"])
> pilow <- pi - 1.96*sdpi
> piup  <- pi + 1.96*sdpi
> cbind( pi, pilow, piup )
```

```
              pi      pilow       piup
FALSE 0.04127701 0.03427527 0.04827875
TRUE  0.04325817 0.03470440 0.05181194
```

If we prefer the result in % and rounded, we just do:

```
> round( cbind( pi, pilow, piup )*100, 2 )
```

```
        pi pilow piup
FALSE 4.13  3.43 4.83
TRUE  4.33  3.47 5.18
```

Compute odds with c.i. and backtransform to improved c.i. for the proportions

```
> omega      <- pi/(1-pi)
> sdlogomega <- sqrt(1/d+1/(n-d))
> errorfact  <- exp(1.96*sdlogomega)
> omegalow   <- omega / errorfact
> omegaup    <- omega * errorfact
> pilow_2    <- omegalow/(1+omegalow)
> piup_2     <- omegaup /(1+omegaup )
> round( cbind( omega, omegalow, omegaup ), 3 )
```

```
      omega omegalow omegaup
FALSE 0.043    0.036   0.051
TRUE  0.045    0.037   0.056
```

```
> round( cbind( pi, pilow, piup, pilow_2, piup_2 )*100, 2 )
```

```
        pi pilow piup pilow_2 piup_2
FALSE 4.13  3.43 4.83    3.48   4.89
TRUE  4.33  3.47 5.18    3.55   5.27
```

Rate per day:

```
> y <- xx[,"fuptime"]
> lambda          <- d/y;
> errorfact_rate <- exp(1.96*sqrt(1/d));
> lambda_low      <- lambda / errorfact_rate;
> lambda_up       <- lambda * errorfact_rate;
> cbind( lambda, lambda_low, lambda_up )
```

```
            lambda   lambda_low    lambda_up
FALSE 0.0002479779 0.0002085333 0.0002948836
TRUE  0.0002582332 0.0002109676 0.0003160884
```

Rates per year:

```
> round( cbind( lambda, lambda_low, lambda_up )*365.25, 3 )
```

```
      lambda lambda_low lambda_up
FALSE  0.091      0.076     0.108
TRUE   0.094      0.077     0.115
```

5.1.4 Rate ratio, risk ratio, odds ratio

1. In order to compute odds-ratios and risk ratios, we use the `twoby2` from the `Epi` package:

```
> ( tt <- with( bissau, table(BCG=bcg,dead) ) )

   dead
BCG    0    1
  1 3176  125
  2 1876   97

> twoby2( tt )

2 by 2 table analysis:
------------------------------------------------------
Outcome   : 0
Comparing : 1 vs. 2

     0   1    P(0) 95% conf. interval
1 3176 125  0.9621    0.9551   0.9681
2 1876  97  0.9508    0.9404   0.9595

                                   95% conf. interval
             Relative Risk: 1.0119    0.9997   1.0242
         Sample Odds Ratio: 1.3137    1.0016   1.7232
Conditional MLE Odds Ratio: 1.3136    0.9908   1.7376
    Probability difference: 0.0113    0.0001   0.0233

             Exact P-value: 0.0555
        Asymptotic P-value: 0.0487
------------------------------------------------------
```

— but this is the wrong way round, so we swap the outcome categories:

```
> twoby2( tt[,2:1] )

2 by 2 table analysis:
------------------------------------------------------
Outcome   : 1
Comparing : 1 vs. 2

    1    0    P(1) 95% conf. interval
1 125 3176  0.0379    0.0319   0.0449
2  97 1876  0.0492    0.0405   0.0596

                                   95% conf. interval
             Relative Risk:  0.7702    0.5943   0.9982
         Sample Odds Ratio:  0.7612    0.5803   0.9984
Conditional MLE Odds Ratio:  0.7612    0.5755   1.0093
    Probability difference: -0.0113   -0.0233  -0.0001

             Exact P-value: 0.0555
        Asymptotic P-value: 0.0487
------------------------------------------------------
```

So we see that the mortality is smaller among those BCG vaccinated.

But it is always more convenient to annotate variables correctly, so we turn `bcg dtpany` and `Dead` into factors. Note that we let the first level of the factor be the exposed:

```
> bissau <- transform( bissau, bcg = factor(bcg,levels=1:2,
+                                           labels=c("BCG","no BCG")),
+                              dtpany = factor(dtp>0,levels=c(TRUE,FALSE),
+                                              labels=c("1+ DTP","no DTP")),
+                                Dead = factor(dead,levels=0:1,
+                                              labels=c("Alive","Dead") )
+                      )
```

2. The same analysis for DTP (any dose):

```
> ( tt <- with( bissau, table(dtpany,Dead) ) )
```

```
        Dead
dtpany   Alive Dead
  1+ DTP  2079   94
  no DTP  2973  128
```

```
> twoby2( tt )
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Alive
Comparing : 1+ DTP vs. no DTP

       Alive Dead    P(Alive) 95% conf. interval
1+ DTP  2079   94      0.9567    0.9473   0.9645
no DTP  2973  128      0.9587    0.9511   0.9652

                                  95% conf. interval
            Relative Risk:  0.9979    0.9865   1.0095
         Sample Odds Ratio:  0.9522    0.7254   1.2500
Conditional MLE Odds Ratio:  0.9523    0.7195   1.2640
    Probability difference: -0.0020   -0.0134   0.0089

            Exact P-value: 0.7281
        Asymptotic P-value: 0.7244
------------------------------------------------------
```

```
> twoby2( tt[,2:1] )
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Dead
Comparing : 1+ DTP vs. no DTP

       Dead Alive    P(Dead) 95% conf. interval
1+ DTP   94  2079     0.0433    0.0355   0.0527
no DTP  128  2973     0.0413    0.0348   0.0489

                                  95% conf. interval
            Relative Risk: 1.0480    0.8076   1.3599
         Sample Odds Ratio: 1.0502    0.8000   1.3785
Conditional MLE Odds Ratio: 1.0501    0.7911   1.3899
    Probability difference: 0.0020   -0.0089   0.0134

            Exact P-value: 0.7281
        Asymptotic P-value: 0.7244
------------------------------------------------------
```

We see that there is no effect of DTP on mortality; the RR is 1.05 and the c.i. is reasonably narrow: (0.81,1.36).

3. Now we look at the association of the two exposures:

```
> with( bissau, table(dtpany,bcg) )
```

```
        bcg
dtpany    BCG no BCG
  1+ DTP 2142     31
  no DTP 1159   1942
```

We see that DTP vaccination is largely confined to those who are BCG-vaccinated. Thus it is only relevant to evaluate the DTP effect among those BCG-vaccinated, because there is no information on the DTP-effect among the non-BCG-vaccinated:

```
> ( tt <- with( subset(bissau,bcg=="no BCG"), table(dtpany,Dead) ) )
```

```
        Dead
dtpany   Alive Dead
  1+ DTP    29    2
  no DTP  1847   95
```

```
> twoby2( tt[,2:1] )
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Dead
Comparing : 1+ DTP vs. no DTP

       Dead Alive    P(Dead) 95% conf. interval
1+ DTP    2    29     0.0645    0.0162   0.2242
no DTP   95  1847     0.0489    0.0402   0.0595

                                  95% conf. interval
             Relative Risk: 1.3188    0.3403   5.1114
         Sample Odds Ratio: 1.3408    0.3153   5.7028
Conditional MLE Odds Ratio: 1.3406    0.1528   5.4347
    Probability difference: 0.0156   -0.0322   0.1585

             Exact P-value: 0.6628
        Asymptotic P-value: 0.6913
------------------------------------------------------
```

But among those with a BCG-vaccination there is information:

```
> ( tt <- with( subset(bissau,bcg=="BCG"), table(dtpany,Dead) ) )
```

```
        Dead
dtpany   Alive Dead
  1+ DTP  2050   92
  no DTP  1126   33
```

```
> twoby2( tt[,2:1] )
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Dead
Comparing : 1+ DTP vs. no DTP

       Dead Alive    P(Dead) 95% conf. interval
1+ DTP   92  2050     0.0430    0.0351   0.0524
no DTP   33  1126     0.0285    0.0203   0.0398

                                  95% conf. interval
            Relative Risk: 1.5085    1.0201   2.2307
         Sample Odds Ratio: 1.5313    1.0221   2.2942
Conditional MLE Odds Ratio: 1.5311    1.0110   2.3697
    Probability difference: 0.0145    0.0008   0.0269

            Exact P-value: 0.0444
        Asymptotic P-value: 0.0389
------------------------------------------------------
```

and we see that is a borderline significant RR=1.5 associated with DTP.

```
> ( ff <- with( bissau, ftable( dtpany, bcg, dead ) ) )
```

```
              dead    0    1
dtpany bcg
1+ DTP BCG         2050   92
       no BCG        29    2
no DTP BCG         1126   33
       no BCG      1847   95
```

```
> round( cbind( ff, ff[,2]/ff[,1]*100 ), 1 )
```

```
     [,1] [,2] [,3]
[1,] 2050   92  4.5
[2,]   29    2  6.9
[3,] 1126   33  2.9
[4,] 1847   95  5.1
```

This table shows that BCG alone is protective, but that either absence of BCG or addition of DTP increases mortality.

This analysis can be made (but only for the OR) by the `effx` function:

```
> effx( dead, type="binary", exposure=dtpany, strata=bcg, data=bissau )
```

```
------------------------------------------------------------------------------
response      :  dead
type          :  binary
exposure      :  dtpany
stratified by :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  1+ DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on dead
stratified by bcg

number of observations  5274
```

```
                                      Effect  2.5% 97.5%
strata BCG level no DTP vs 1+ DTP      0.653 0.436 0.978
strata no BCG level no DTP vs 1+ DTP   0.746 0.175 3.170

Test for effect modification on 1 df: p-value= 0.86
```

So we see there is no evidence that DTP has differential effect, so we could BCG as a confounder instead (controlling for it:

```
> effx( dead, type="binary", exposure=dtpany, control=bcg, data=bissau )


------------------------------------------------------------------------------
response      :  dead
type          :  binary
exposure      :  dtpany
control vars  :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  1+ DTP
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on dead
controlled for bcg

number of observations  5274

Effect   2.5%  97.5%
 0.660  0.448  0.971

Test for no effects of exposure on 1 df: p-value= 0.0313
```

4. The `effx` function allows calculation of the rate-ratios etc. very easily:

```
> effx( dead, type="failure", fup=fuptime/365.25, exposure=bcg, data=bissau )


------------------------------------------------------------------------------
response      :  dead
type          :  failure
exposure      :  bcg

bcg is a factor with levels: BCG / no BCG
baseline is  BCG
effects are measured as rate ratios
------------------------------------------------------------------------------

effect of bcg on dead
number of observations  5274

Effect   2.5%  97.5%
  1.32   1.02   1.73

Test for no effects of exposure on 1 df: p-value= 0.0395


> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany, base=2, data=bissau )


------------------------------------------------------------------------------
response      :  dead
type          :  failure
exposure      :  dtpany
```

```
dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as rate ratios
------------------------------------------------------------------------------

effect of dtpany on dead
number of observations  5274

Effect   2.5%  97.5%
 1.040  0.798  1.360

Test for no effects of exposure on 1 df: p-value= 0.766


> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany,
+             strata=bcg, base=2, data=bissau )


------------------------------------------------------------------------------
response      :  dead
type          :  failure
exposure      :  dtpany
stratified by :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as rate ratios
------------------------------------------------------------------------------

effect of dtpany on dead
stratified by bcg

number of observations  5274

                                     Effect  2.5% 97.5%
strata BCG level 1+ DTP vs no DTP      1.53 1.030  2.27
strata no BCG level 1+ DTP vs no DTP   1.24 0.305  5.02

Test for effect modification on 5272 df: p-value= 0.685


> effx( dead, type="failure", fup=fuptime/365.25, exposure=dtpany,
+             control=bcg, base=2, data=bissau )


------------------------------------------------------------------------------
response      :  dead
type          :  failure
exposure      :  dtpany
control vars  :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as rate ratios
------------------------------------------------------------------------------

effect of dtpany on dead
controlled for bcg

number of observations  5274

Effect   2.5%  97.5%
  1.50   1.03   2.19

Test for no effects of exposure on 1 df: p-value= 0.0316
```

The results are virtually identical to those for the relative risk, where we ignored the follow-up time.

5.1.5 Confounder control: stratified analysis of odds ratio and risk ratio

We will continue using the data from Guinea-Bissau for this third part of the exercise.

1. Revisit the analysis from previously, using just death (dead) as outcome, and estimate the DTP effect for each level of BCG.

   When we use only dead/alive as outcome in the analysis-function `effx`, we must use "Dead" as outcome:

```
> effx( (Dead=="Dead")*1, type="binary", exposure=dtpany, strata=bcg, data=bissau )

------------------------------------------------------------------------------
response      :  (Dead == "Dead") * 1
type          :  binary
exposure      :  dtpany
stratified by :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  1+ DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on (Dead == "Dead") * 1
stratified by bcg

number of observations  5274

                                      Effect  2.5% 97.5%
strata BCG level no DTP vs 1+ DTP     0.653 0.436 0.978
strata no BCG level no DTP vs 1+ DTP  0.746 0.175 3.170

Test for effect modification on 1 df: p-value= 0.86
```

   We see that there is no interaction by the test for effct modification (the likelihood-ratio counterpart of the Breslow-Day-test). So we conclude that there is the same effect of DTP for both levels of BCG vaccination. It is clear that because of the vary sparse data, the effect of DTP in the "no BCG" stratum is largely undetermined.

   But we also see that the reference levels used is those exposed to dtp, so we must use the `base` argument to get the right comparison:

```
> effx( (Dead=="Dead")*1,
+       type = "binary",
+   exposure = dtpany,
+       base = "no DTP",
+     strata = bcg,
+       data = bissau )

------------------------------------------------------------------------------
response      :  (Dead == "Dead") * 1
type          :  binary
exposure      :  dtpany
stratified by :  bcg
```

```
dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
bcg is a factor with levels: BCG/no BCG
effects are measured as odds ratios
------------------------------------------------------------------------

effect of dtpany on (Dead == "Dead") * 1
stratified by bcg

number of observations  5274

                                      Effect  2.5% 97.5%
strata BCG level 1+ DTP vs no DTP       1.53 1.020  2.29
strata no BCG level 1+ DTP vs no DTP    1.34 0.315  5.70

Test for effect modification on 1 df: p-value= 0.86
```

2. Use the BCG as a potentially confounding variable and obtain the MH-estimate for the OR and RR. What are they?

   We can use `effx` with a slight modification to compute the common effect, by simply replacing `strata=` with `control=`:

```
> effx( (Dead=="Dead")*1, type="binary", exposure=dtpany, base="no DTP",
+       control=bcg, data=bissau )


------------------------------------------------------------------------
response      :  (Dead == "Dead") * 1
type          :  binary
exposure      :  dtpany
control vars  :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as odds ratios
------------------------------------------------------------------------

effect of dtpany on (Dead == "Dead") * 1
controlled for bcg

number of observations  5274

Effect   2.5%  97.5%
  1.52   1.03   2.23

Test for no effects of exposure on 1 df: p-value= 0.0313
```

   It is also possible to estimate the relative risk, using the argument `eff="RR"` — but only from version 1.1.40 of the Epi package, where it is also possible to use just a logical as a binary response. You can check your version of the `Epi`-package by:

```
> installed.packages()["Epi",c("Version","Built"),drop=FALSE]


    Version  Built
Epi "1.1.40" "2.15.1"


> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP",
+       control=bcg, data=bissau, eff="RR" )
```

```
------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  bcg

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as relative risk
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for bcg

number of observations  5274

Effect   2.5%  97.5%
  1.49   1.03   2.17

Test for no effects of exposure on 1 df: p-value= 0.0314
```

3. Do the same, using age in months (`agem`) as control variable in the analysis. Is there any DTP effect?

   We first get an overview of how the data are distributed by age, `agem`:

```
> ( tt <- with( bissau, table(agem,dtp) ) )

    dtp
agem   0   1   2   3
   0 867   7   0   0
   1 808  81   0   0
   2 559 326  32   2
   3 339 328 122  18
   4 256 267 160  76
   5 196 209 181 109
   6  76 100  96  59

> pctab( tt )

    dtp
agem     0     1     2     3   All     N
   0  99.2   0.8   0.0   0.0 100.0 874.0
   1  90.9   9.1   0.0   0.0 100.0 889.0
   2  60.8  35.5   3.5   0.2 100.0 919.0
   3  42.0  40.6  15.1   2.2 100.0 807.0
   4  33.7  35.2  21.1  10.0 100.0 759.0
   5  28.2  30.1  26.0  15.7 100.0 695.0
   6  23.0  30.2  29.0  17.8 100.0 331.0

> ( tt <- with( bissau, table(agem,bcg) ) )

    bcg
agem BCG no BCG
   0 237    637
   1 468    421
   2 598    321
   3 589    218
   4 581    178
   5 554    141
   6 274     57
```

```
> pctab( tt )
```

```
    bcg
agem   BCG no BCG   All     N
   0  27.1   72.9 100.0 874.0
   1  52.6   47.4 100.0 889.0
   2  65.1   34.9 100.0 919.0
   3  73.0   27.0 100.0 807.0
   4  76.5   23.5 100.0 759.0
   5  79.7   20.3 100.0 695.0
   6  82.8   17.2 100.0 331.0
```

We see that the distribution of DTP and BCG vaccinations are highly dependent on age. So we should perhaps expect hat some of the effect would disappar when we control for age. Note that we can control for age in two different ways; we can either include it as a continuous variable (with a linear effect) or as a factor:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=agem, data=bissau
```

```
------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  agem

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for agem

number of observations  5274

Effect   2.5%  97.5%
 1.010  0.731  1.400

Test for no effects of exposure on 1 df: p-value= 0.947
```

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=factor(agem), dat
```

```
------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  agem

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for agem

number of observations  5274

Effect   2.5%  97.5%
 1.020  0.729  1.420

Test for no effects of exposure on 1 df: p-value= 0.915
```

We see that with this control there is no effect of DTP, irrespective of how we control for age.

4. Do the same, but now using both `agem` and `bcg` (that is, the cross-classification) as control variables in the analysis. Is there any DTP effect?

   If we control for both, it means that we insert both variables as confounders, in the `effx` function this means that we should insert the two in a `list` used as the `control` argument:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=list(bcg,factor(a


------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  bcg factor(agem)

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for bcg factor(agem)

number of observations  5274

Effect   2.5%  97.5%
 1.470  0.954  2.280

Test for no effects of exposure on 1 df: p-value= 0.0754


> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", control=list(bcg,agem), d


------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  bcg agem

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for bcg agem

number of observations  5274

Effect   2.5%  97.5%
 1.460  0.954  2.230

Test for no effects of exposure on 1 df: p-value= 0.077
```

   Again we see that the addition of `bcg` as a confounding variable reveals some effect of DTP, regardless of whether we use `agem` as a quantitative variable (assuming that the effect of age is linear on the log-odds-scale).

Formally speaking the effect is non-significant as opposed to what it was when we only controlled for BCG and not age. However, the difference between a p-value of 3 and one of 7% is very small, the actual effect is in both cases an odd-ratio of 1.46, and depending on how we control we have and lower limit of the confidnece interval for the OR of 1.03 (signif.) or of 0.95 (non-sign.).

However in both cases we see an indication of elevated risk of about 50%, but we cannot really tell wheter it is a few percent or a doubling of mortality.

Since the absolute mortality is so small, it does not matter wheter we use OR or RR, the two measures are virtually identical when the probabilities of putcome are small:

```
> effx( Dead=="Dead", type="binary", exposure=dtpany, base="no DTP", eff="RR",
+       control=list(bcg,agem), data=bissau )


------------------------------------------------------------------------------
response      :  Dead == "Dead"
type          :  binary
exposure      :  dtpany
control vars  :  bcg agem

dtpany is a factor with levels: 1+ DTP / no DTP
baseline is  no DTP
effects are measured as relative risk
------------------------------------------------------------------------------

effect of dtpany on Dead == "Dead"
controlled for bcg agem

number of observations  5274

Effect   2.5%  97.5%
 1.440  0.956  2.160

Test for no effects of exposure on 1 df: p-value= 0.0761
```

5.1.6 Survival analysis of childhood mortality in Guinea-Bissau

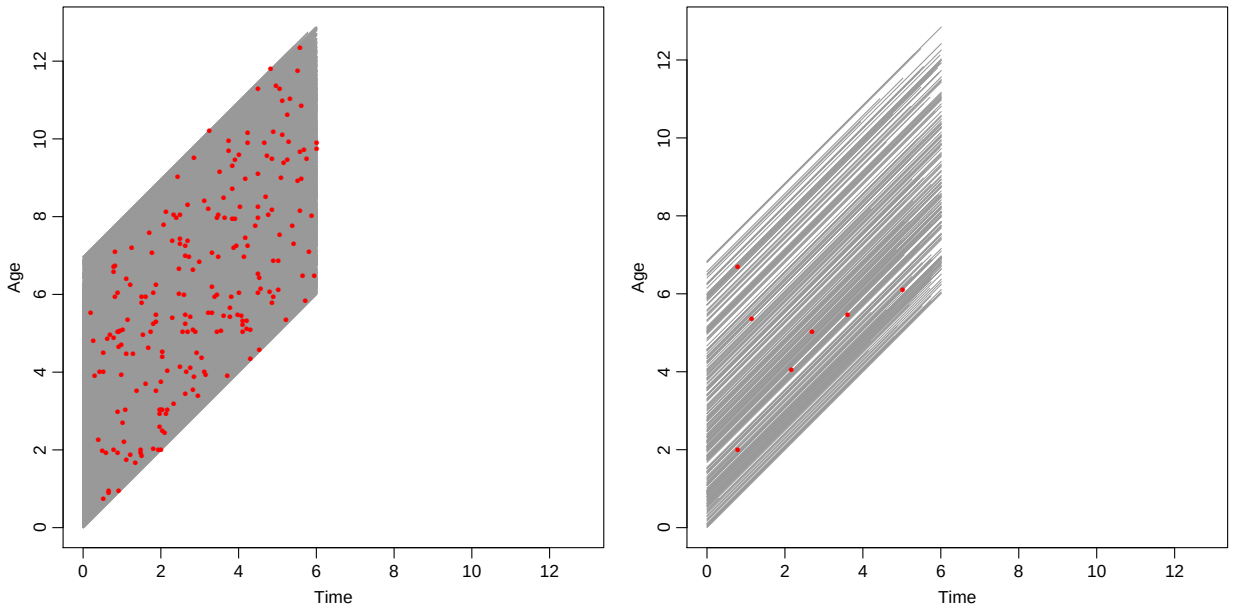
1. We start by reading the data and transforming data as before In order to do survival analysis we set up data as a `Lexis` object with two time-scales, time since visit and current age (time since birth), note we enter the time-scales in months:

```
> Lb <- Lexis( entry = list( Time = 0,
+                            Age = age/(365.25/12) ),
+              exit = list( Time = fuptime/(365.25/12) ),
+       exit.status = Dead,
+              data = bissau )


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


> summary( Lb )


Transitions:
     To
From    Alive Dead  Records:  Events: Risk time:  Persons:
  Alive  5052  222      5274      222   28917.85      5274

Rates:
```

```
     To
From    Alive Dead Total
  Alive     0 0.01  0.01
```

We can show the follow-up in a Lexis-diagram, both for all and for a 5% random sample:

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plot( Lb, col=gray(0.6) )
> points( Lb, pch=c(NA,16)[Lb$"lex.Xst"], col="red", cex=0.6 )
> wh <- runif(nrow(Lb))<0.05
> plot( Lb[wh,], col=gray(0.6) )
> points( Lb[wh,], pch=c(NA,16)[Lb[wh,"lex.Xst"]], col="red", cex=0.6 )
```



Note that both `age` and `fuptime` are measured in days, so the time-variables we have in `Lb` are measured in months. A simple Cox-model with bcg and age at entry (in months) at entry is set up using time since first visit (`Time`) as time-scale:

```
> library( survival )
> m1 <- coxph( Surv( Time, Time+lex.dur, lex.Xst=="Dead" ) ~
+              bcg + factor(agem),
+              data=Lb )
> ci.exp( m1 )
```

```
                exp(Est.)      2.5%    97.5%
bcgno BCG       1.4150942 1.0628313 1.884110
factor(agem)1   1.1218691 0.7118992 1.767933
factor(agem)2   0.7734721 0.4659289 1.284014
factor(agem)3   1.2201058 0.7574272 1.965414
factor(agem)4   1.3944726 0.8680841 2.240052
factor(agem)5   1.3918874 0.8534351 2.270062
factor(agem)6   0.9895328 0.4950160 1.978068
```

We see that persons without BCG-vaccination have a higher mortality.

2. We can evaluate the effect of DTP vaccination by changing the covariate `bcg` wirth `dtpany`

```
> m2 <- update( m1, . ~ . - bcg + dtpany )
> ci.exp( m2 )
```

```
               exp(Est.)      2.5%    97.5%
factor(agem)1 1.0293268 0.6558760 1.615418
factor(agem)2 0.6788407 0.4068749 1.132595
factor(agem)3 1.0393116 0.6327681 1.707053
factor(agem)4 1.1729539 0.7104716 1.936490
factor(agem)5 1.1584998 0.6882826 1.949958
factor(agem)6 0.8117511 0.3957607 1.664996
dtpanyno DTP  0.9979789 0.7197838 1.383696
```

and we see there is no mrginal effect of DTP om mortality.

3. But if we enter both variables we see an effect of both:

```
> m3 <- update( m1, . ~ . + dtpany )
> ci.exp( m3 )
```

```
               exp(Est.)      2.5%    97.5%
bcgno BCG     1.7376437 1.1871318 2.543446
factor(agem)1 1.1420990 0.7243778 1.800704
factor(agem)2 0.7302230 0.4365875 1.221349
factor(agem)3 1.0978956 0.6678785 1.804781
factor(agem)4 1.2286509 0.7444014 2.027915
factor(agem)5 1.2090204 0.7186231 2.034071
factor(agem)6 0.8495861 0.4143648 1.741935
dtpanyno DTP  0.6915217 0.4524274 1.056970
```

we see a protective effect of BCD, but potential harmful effect of DTP.

4. We can then try to insert the interaction:

```
> m4 <- update( m3, . ~ . + dtpany:bcg )
> ci.exp( m4 )
```

```
                        exp(Est.)      2.5%    97.5%
bcgno BCG               1.4305868 0.3522739 5.809624
factor(agem)1           1.1453270 0.7261539 1.806468
factor(agem)2           0.7319001 0.4374948 1.224421
factor(agem)3           1.0989575 0.6684984 1.806598
factor(agem)4           1.2299067 0.7451548 2.030008
factor(agem)5           1.2120327 0.7202963 2.039471
factor(agem)6           0.8509784 0.4150298 1.744849
dtpanyno DTP            0.6796851 0.4369438 1.057280
bcgno BCG:dtpanyno DTP 1.2359351 0.2877469 5.308609
```

```
> anova( m3, m4, test="Chisq" )
```

```
Analysis of Deviance Table
 Cox model: response is  Surv(Time, Time + lex.dur, lex.Xst == "Dead")
 Model 1: ~ bcg + factor(agem) + dtpany
 Model 2: ~ bcg + factor(agem) + dtpany + bcg:dtpany
   loglik Chisq Df P(>|Chi|)
1 -1875.2
2 -1875.2 0.086  1     0.7693
```

which, although not significant, is not very informative; we want the RR for each combination of the two variables:

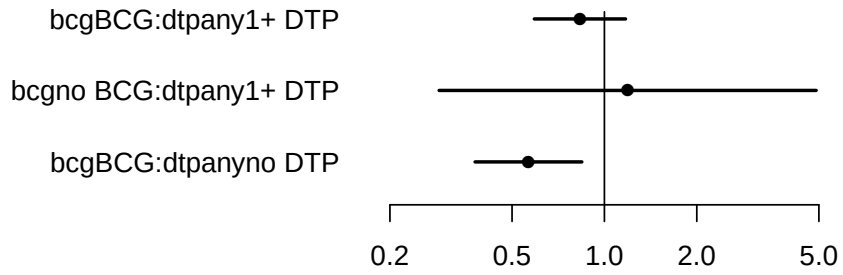
```
> m5 <- update( m1, . ~ . - bcg + dtpany:bcg )
> ci.exp( m5 )
```

```
                          exp(Est.)      2.5%     97.5%
factor(agem)1           1.1453270 0.7261539 1.8064679
factor(agem)2           0.7319001 0.4374948 1.2244210
factor(agem)3           1.0989575 0.6684984 1.8065977
factor(agem)4           1.2299067 0.7451548 2.0300083
factor(agem)5           1.2120327 0.7202963 2.0394708
factor(agem)6           0.8509784 0.4150298 1.7448486
bcgBCG:dtpany1+ DTP     0.8321130 0.5908951 1.1718020
bcgno BCG:dtpany1+ DTP  1.1904099 0.2893258 4.8978542
bcgBCG:dtpanyno DTP     0.5655749 0.3788463 0.8443395
```

`coxph` produces a warning, because the interaction generated contains the intercept (namely the sum of the 4 columns), and so automatically exclude the last one.

We can also see the estimates graphically

```
> plotEst( ci.exp(m5,subset="bcg"), xlog=T, vref=1 )
```

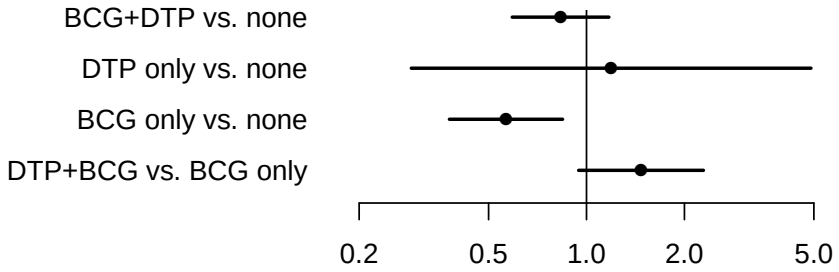


The reference is the no BCG, no DTP group, so we see that the protective effect is smaller in the DTP-vaccinated group than in the non-DTP-vaccinated group. It is of course also of interest to see the DTP-effect within the BCG-group, and that can be teased out:

```
> CM <- rbind( diag(3), c(1,0,-1) )
> rownames( CM ) <- c("BCG+DTP vs. none",
+                     "DTP only vs. none",
+                     "BCG only vs. none",
+                     "DTP+BCG vs. BCG only")
> CM
```

```
                    [,1] [,2] [,3]
BCG+DTP vs. none       1    0    0
DTP only vs. none      0    1    0
BCG only vs. none      0    0    1
DTP+BCG vs. BCG only   1    0   -1
```

```
> plotEst( ci.exp(m5,subset="bcg",ctr.mat=CM), xlog=T, vref=1 )
```



There are very few persons and

```
> ftable( xtabs( cbind(dead,N=1) ~ bcg + dtpany, data=Lb ) )
```

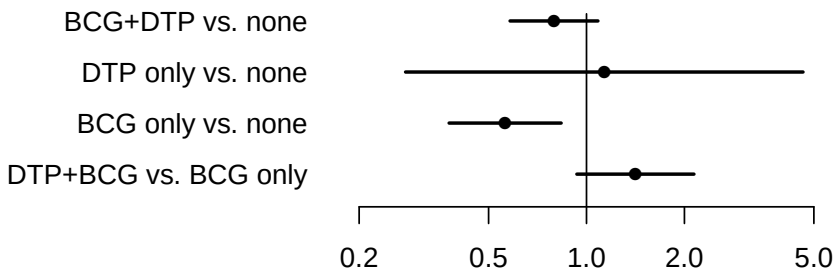
```
               dead    N
bcg    dtpany
BCG    1+ DTP    92 2142
       no DTP    33 1159
no BCG 1+ DTP     2   31
       no DTP    95 1942
```

so essentially we can only estimate a BCG-effect among non-DTP vaccinated and a DTP-effect among BCG-vaccinated, which are the effects we see in the main effects model.

5. If we instead use current age as time-scale, we just change the outcome variable using `update`, and then re-use the code to generate the graph:

```
> m5a <- update( m5, Surv(Age,Age+lex.dur,lex.Xst=="Dead") ~ . - factor(agem) )

> plotEst( ci.exp(m5a,subset="bcg",ctr.mat=CM), xlog=T, vref=1 )
```

We see that the analysis with current age as time scale gives pretty much the same estimates as the analysis with time since entry and age at entry:



6. Finally we compare the results from a Poisson model where we assume constant rates, and a logistic regression model where we altogether ignore censoring:

```
> ci.exp(m3,subset="no")

                exp(Est.)      2.5%    97.5%
bcgno BCG     1.7376437 1.1871318 2.543446
dtpanyno DTP  0.6915217 0.4524274 1.056970


> p3 <- glm( (lex.Xst=="Dead") ~ bcg + dtpany + factor(agem),
+            offset=log(fuptime), family=poisson, data=Lb )
> ci.exp(p3,subset="no")

                exp(Est.)      2.5%    97.5%
bcgno BCG     1.7349813 1.1851585 2.539880
dtpanyno DTP  0.6912244 0.4521804 1.056638


> l3 <- glm( (lex.Xst=="Dead") ~ bcg + dtpany + factor(agem),
+            family=binomial, data=Lb )
> ci.exp(l3,subset="no")

                exp(Est.)      2.5%    97.5%
bcgno BCG     1.7397389 1.1772134 2.571064
dtpanyno DTP  0.6784886 0.4393344 1.047828
```

   We see that the three approaches produce virtually identical results. For the Poisson model it is because the mortality varies very little with age and follow-up, and the Poisson model *is* a model that assumes constant mortality. For the logistic model it is because the amount of censring is quite limited

5.2 Case-control study of renal cancer and trichloroethene

The exercise is based on Vamvakas *et al.*: Renal cell cancer correlated with occupational exposure to trichloroethane. J Cancer Res Clin Oncol, 1998, pp 374–382. [1].

The following points were addressed:

1. What is the primary aim of the study?

   The primary aim as stated is to assess the effect of thricholorethene (C_2HCl_3) and tetrachloroethene (C_2Cl_4) on the occurrence of kidney cancer. This is based on a described possible biological mechanism.

2. How were the cases sampled?

   Cases in the study were patients who underwent surgery between 1987-12-01 and 1992-05-01, a period of 3.5 years. Of the 78 patients 62 responded (or their next of kin, since 4 of these 62 were dead)

3. How were the controls sampled?

   Controls were sampled from the accident wards of three *other* hospitals in the same area, form the period 1993-01-01 to 1993-12-31.

4. Are they comparable; i.e. what assumptions are needed?

   In order to deem cases and controls comparable we must assume:

   (a) In principle we must assume that the controls would have been enrolled as cases *if they had had a diagnosis of kidney cancer*.

   In practice it will suffice that the controls comes from a comparable populations, that is even though they might have been subjected to surgery at a different hospital in case of kidney cancer, if this were a similar type of hospital it would have been sufficient.

   (b) Exposure distribution in the population has not changed between the case-selection period (1987–1992) and the control sampling period (1993).

   It is of course difficult to assess this without further knowledge about the industrial development in Nordrhein-Westphalia in the period 1987–1993.

   (c) The attendance of the accident ward is unrelated to the exposure of interest.

   At face value it seems so, bar of course accidents related to trichloroethene exposure itself, but we can safely assume that they are very rare compared to other. However, the attendance at the accident ward is hardly unrelated to age, and neither is exposure to to trichloroethene, so the attendance *is* related to the exposure, albeit not directly.

   (d) Recall of exposure is the same among cases and controls.

   It is not stated anywhere whether the purpose of the study was revealed to cases and controls, but in the likely event that it were, cases may be more prone to recall exposure to the thricholorethene in order to get an explanation of their (severe) disease.

   As we see there are a few potential biases, some tend to *increase* the risk estimate; the problems with age is difficult to assess.

5. What is the (actual) study base?

   There is no definite answer to this, but if we define the study base to be the state of Nordrhein-Westphalia, the validity of the study hinges on assumptions that the hospitals where cases and controls are sampled are representative of the population — with respect to the exposure of interest.

   Of course, precisely the same argument applies if we define the study base to some subset of Nordrhein-Westphalia, which includes the uptake-areas of the 4 hospitals, and incidentally we might define the study base as the uptake area of these 4 hospitals.

6. What study base is the intended (for generalization)?

   The intended study base is presumably all industrialized counties comparable to Germany.

7. Is the sampling scheme incidence density sampling?

   The sampling scheme is certainly not incidence density sampling, that would require that cases were selected among those at risk at the time of the case-occurrence, ad that is not the case.

8. Can the age-effect on the occurrence of renal cancer be estimated?

   The age-effect cannot be estimated; that would require that the controls were a representative sample (w.r.t. age) of the study base. Sampling persons from the accident wards makes this a far-fetched assumption.

9. Is age a confounder?

   Most likely so; it is definitely associated with both the exposure and the outcome.

10. Key in the numbers in table 6 (p.380), and verify the analysis using SAS `proc freq`.

    When we enter the data from table 6 there are a few things to be aware of:

    - Unlike previously, we are entering *grouped* or `tabulated` data from a table, not individual records. This means that every line in the input dataset represents as (usually large) number of individuals.
    - This in turn means that we shall need one line in the data set per entry in the table. In addition to the number, the line must then contain variables that tells us whether it is cases/controls, exposed or not, and what age.
    - Note that the entries in table 6 in the paper are not the number of cases and controls, but the number of cases and *total* number in the strata. So we need to do a bit of subtraction to get the numbers right.
    - Unlike SAS there is no elegant way to include a small dataset as this in the program text, so you must key it in in a separate file and read it from that.

    The data looks like this:

```
 age tri ck n
  30   1   1  2
  40   1   1  2
  50   1   1 10
  60   1   1  1
  70   1   1  4
  30   0   1  0
  40   0   1  1
  50   0   1 12
  60   0   1 17
  70   0   1  9
  30   1   0  1
  40   1   0  4
  50   1   0  2
  60   1   0  0
  70   1   0  0
  30   0   0 21
  40   0   0 11
  50   0   0 25
  60   0   0 14
  70   0   0  6
```

    It is stored in the file `renal.txt` and is read with

```
> # renal <- read.table(url("http://bendixcarstensen.com/EpiE2012/data/renal.txt"),header=TRUE
> renal <- read.table("../data/renal.txt",header=TRUE)
> renal

   age tri ck  n
1   30   1  1  2
2   40   1  1  2
```

```
3   50   1  1 10
4   60   1  1  1
5   70   1  1  4
6   30   0  1  0
7   40   0  1  1
8   50   0  1 12
9   60   0  1 17
10  70   0  1  9
11  30   1  0  1
12  40   1  0  4
13  50   1  0  2
14  60   1  0  0
15  70   1  0  0
16  30   0  0 21
17  40   0  0 11
18  50   0  0 25
19  60   0  0 14
20  70   0  0  6
```

The R-program can be found in the folder http://bendixcarstensen.com/EpiE2012/R/, and the entire log file is a part of this section.

Once we have the data we use the `effx` function from the `Epi` package to make an overall analysis:

```
> library(Epi)
> effx( response=ck, type="bin", exp=tri, weight=n, data=renal )


------------------------------------------------------------------------------
response      :  ck
type          :  binary
exposure      :  tri

tri is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of an increase of 1 unit in tri on ck
number of observations  17

Effect   2.5%  97.5%
  5.36   2.08  13.80

Test for no effects of exposure on 1 df: p-value= 0.000221
```

But we want to control for age as a confounder; this is done by adding the argument `control=`. Note that we must enter `age` as `factor(age)` in order for the variable to be treated as a class variable (which is called a factor in R):

```
> effx( response=ck, type="bin", exp=tri, control=factor(age), weight=n, data=renal )


------------------------------------------------------------------------------
response      :  ck
type          :  binary
exposure      :  tri
control vars  :  age

tri is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------
```

```
effect of an increase of 1 unit in tri on ck
controlled for age

number of observations  17

Effect   2.5%  97.5%
 16.00   4.05  63.10

Test for no effects of exposure on 1 df: p-value= 2.09e-06
```

11. Is there any evidence of heterogeneity of the odds-ratio across age-classes? (*Hint:* Use the Breslow-Day-test.)

    We can test for heterogeneity of `tri`-exposure by using the argument `strata=` instead of `control=`:

```
> effx( response=ck, type="bin", exp=tri, strata=factor(age), weight=n, data=renal )


------------------------------------------------------------------------------
response      :  ck
type          :  binary
exposure      :  tri
stratified by :  factor(age)

tri is numeric
factor(age) is a factor with levels: 30/40/50/60/70
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of an increase of 1 unit in tri on ck
stratified by factor(age)

number of observations  17

            Effect  2.5% 97.5%
strata 30 2.92e+08 0.000   Inf
strata 40 5.50e+00 0.385  78.6
strata 50 1.04e+01 1.970  55.2
strata 60 1.29e+07 0.000   Inf
strata 70 2.32e+07 0.000   Inf

Test for effect modification on 4 df: p-value= 0.434
```

    So we see there is no evidence of age-heterogeneity, but the age-classes are quite crude (10-years). Thus it seems that the analysis where we just include age as a controlling factor gives an adequate description of data.

12. In particular, how does the odds-ratio estimate given by Vamvakas *et al.* compare to the Mantel-Haenszel estimate based on the same data?

    The estimate given in the paper is somewhat smaller than the MH-estimate, but the substantial message is pretty much the same: definitely an effect, but unclear how large.

13. What is the main result (in plain words)?

    Based on this study alone, there seems to be an excess risk of kidney cancer associated with trichloroethene exposure, but its cannot really be determined whether the OR is 2.5 or 25. This is of course due to the rather small sample size.

5.3 IHD data from Clayton & Hills.

The study is described by Clayton & Hills, Ch. 13. The tabulated data set of counts of IHD cases and person-years is avaliable from `www` in the file `ihd-tab.txt`.

1. Fit the model from Clayton & Hills Tables 22.7-8 (p.222) and perform the tests from exercises 24.1 and 24.2 (pp.237–238).

   First we load the `Epi` package and read the grouped IHD-data from the file `ihd-tab.txt` from the data folder `"http://BendixCarstensen.com/EpiE2012/data"`:

```
> options( width=90 )
> library( Epi )
> library( foreign )
> ihdt <- read.table("http://BendixCarstensen.com/EpiE2012/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
```

   Then we 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.lin( mt, E=T ), 3 )
```

```
             Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)    -5.418  0.442 -12.256 0.000     0.004 0.002 0.011
factor(age)1    0.129  0.475   0.271 0.786     1.138 0.448 2.888
factor(age)2    0.692  0.461   1.500 0.134     1.998 0.809 4.935
exposure        0.870  0.308   2.823 0.005     2.386 1.305 4.364
```

   We see that the results are pretty well in agreement with the results from tables 22.8 and 24.1 in Clayton & Hills.

2. Fit the model with interaction and re-find results from Clayton & Hills Table 24.5 (p.242) and the test for no interaction.

```
> mi <- update( mt, . ~ . + factor(age):exposure )
> round( ci.lin( mi, E=T ), 4 )
```

```
                     Estimate StdErr        z      P exp(Est.)   2.5%   97.5%
(Intercept)           -5.0237 0.5000 -10.0474 0.0000    0.0066 0.0025  0.0175
factor(age)1          -0.5153 0.6708  -0.7681 0.4424    0.5973 0.1604  2.2245
factor(age)2           0.3132 0.6124   0.5114 0.6091    1.3678 0.4119  4.5422
exposure              -0.0258 0.8660  -0.0298 0.9762    0.9745 0.1785  5.3205
factor(age)1:exposure  1.2720 1.0165   1.2513 0.2108    3.5678 0.4866 26.1622
factor(age)2:exposure  0.8719 0.9728   0.8962 0.3701    2.3914 0.3553 16.0968
```

   The test for no interaction is obtained by comparing the two models using `anova`. For the sake of completeness we also fit the model without age-effect showing the deviances illustrated in figure 24.1 in C&H:

```
> me <- update( mt, . ~ . - factor(age) )
> anova( mi, mt, me, test="Chisq" )


Analysis of Deviance Table

Model 1: cases ~ factor(age) + exposure + factor(age):exposure
Model 2: cases ~ factor(age) + exposure
Model 3: cases ~ exposure
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0     0.0000
2         2     1.6727 -2  -1.6727   0.4333
3         4     5.6891 -2  -4.0164   0.1342
```

5.3.1 Using continuous variables

The IHD-data contains energy consumption as a continuous variable, `energy`, among other variables, and this exercise we will discuss how to model these if the continuous effect is not linear:

1. Read the individual diet data records from the file.
2. Create variables for the person-years, by subtracting entry date from date of exit. Also create a variable with the log-person-years.
3. Use CHD as outcome variable in a Poisson-analysis with the log-person-years as offset, using energy as a linear explanatory variable. Is there an effect on mortality?
4. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (approx. the quartiles)
5. Same question for weight and BMI (the latter you have to calculate yourself as weight/height^2).

5.3.2 Splitting the follow-up of the IHD-data

1. First we (again) load the `Epi` package and read the (modified) grouped IHD-data from the file `ihd-xtab.dta` from the data folder "http://BendixCarstensen.com/EpiE2012/data"

```
> options( width=90 )
> library( Epi )
> library( foreign )
> ihdt <- read.table("http://BendixCarstensen.com/EpiE2012/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
```

   Then we 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.lin( mt, E=T ), 3 )
```

```
             Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)    -5.418  0.442 -12.256 0.000     0.004 0.002 0.011
factor(age)1    0.129  0.475   0.271 0.786     1.138 0.448 2.888
factor(age)2    0.692  0.461   1.500 0.134     1.998 0.809 4.935
exposure        0.870  0.308   2.823 0.005     2.386 1.305 4.364
```

We see that the results are pretty well in agreement with the results from table 24.1 in Clayton & Hills.

2. Next, we read the individual records from the file `diet.txt`; remembering to specify how missing values is coded, and using `as.is=TRUE` to prevent character variebles to be automatically converted to factors:

```
> ihdi <- read.table( # "../data/diet.txt",
+                     "http://BendixCarstensen.com/EpiE2012/data/diet.txt",
+                     header=TRUE, na.strings=".", as.is=TRUE )
> 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   :Classes 'cal.yr', 'numeric'  num [1:337] 1965 1965 1966 1966 1966 ...
 $ dox   :Classes 'cal.yr', 'numeric'  num [1:337] 1977 1977 1977 1977 1976 ...
 $ chd   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dob   :Classes 'cal.yr', 'numeric'  num [1:337] 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
```

3. Now we 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, `per`iod) 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. Id not `entry.status` is given, 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 )
```

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

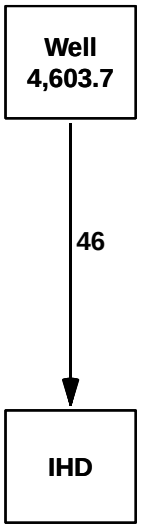
```
> summary( Lx )
```

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

Rates:
     To
From   Well  IHD Total
  Well    0 0.01  0.01
```

[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.

There is a method for plotting the follow-up in boxes. Noe despeartely exciting but capturing the essence:

```
> boxes( Lx, boxpos=TRUE )
```



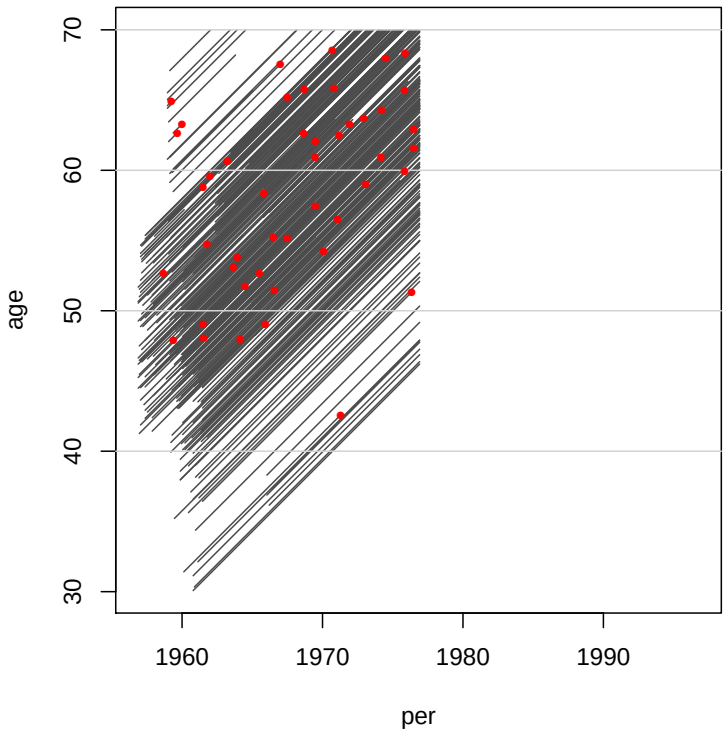
4. The time-splitting is now done by the function `splitLexis`. To use the function we must specify which timescale to split the data on. In this case we want to split along the scale "current age", i.e. time since date of birth, here names `age`. We then specify the intervals where we want the follow-up 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( Ls )

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

Rates:
     To
From   Well  IHD Total
  Well    0 0.01  0.01


> head( Ls )

  lex.id      per      age   lex.dur lex.Cst lex.Xst id      doe      dox chd      dob
1      1 1964.623 49.61533  0.384668    Well    Well  1 1964.623 1976.916   0 1915.008
2      1 1965.008 50.00000 10.000000    Well    Well  1 1964.623 1976.916   0 1915.008
3      1 1975.008 60.00000  1.908282    Well    Well  1 1964.623 1976.916   0 1915.008
4      2 1964.957 50.53799  9.462012    Well    Well  2 1964.957 1976.916   0 1914.419
5      2 1974.419 60.00000  2.496920    Well    Well  2 1964.957 1976.916   0 1914.419
6      3 1965.874 58.78439  1.215606    Well    Well  3 1965.874 1976.916   0 1907.090
  job month  energy   height   weight    fat fibre
1   0     8 2.87395 175.3870 71.48737 141.71 17.83
2   0     8 2.87395 175.3870 71.48737 141.71 17.83
3   0     8 2.87395 175.3870 71.48737 141.71 17.83
```

```
4   0    12 1.98234 164.2872 70.08120  85.77  9.49
5   0    12 1.98234 164.2872 70.08120  85.77  9.49
6   0    11 2.66858 169.3926 71.89560 107.67 15.99
```

For the fun of it you can try the default `plot` and `points` methods for a `Lexis` object. Note that gridlines 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 )
```



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`) is easily computed

```
> nrow( Ls )
```

```
[1] 755
```

6. We now list the first 20 records:

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

```
   lex.id      per      age    lex.dur lex.Cst lex.Xst id      doe      dox chd      dob
1       1 1964.623 49.61533  0.3846680    Well    Well  1 1964.623 1976.916   0 1915.008
2       1 1965.008 50.00000 10.0000000    Well    Well  1 1964.623 1976.916   0 1915.008
3       1 1975.008 60.00000  1.9082820    Well    Well  1 1964.623 1976.916   0 1915.008
4       2 1964.957 50.53799  9.4620123    Well    Well  2 1964.957 1976.916   0 1914.419
5       2 1974.419 60.00000  2.4969199    Well    Well  2 1964.957 1976.916   0 1914.419
```

```
6       3 1965.874 58.78439  1.2156057    Well    Well  3 1965.874 1976.916   0 1907.090
7       3 1967.090 60.00000  9.8261465    Well    Well  3 1965.874 1976.916   0 1907.090
8       4 1965.707 58.72690  1.2731006    Well    Well  4 1965.707 1976.916   0 1906.980
9       4 1966.980 60.00000  9.9356605    Well    Well  4 1965.707 1976.916   0 1906.980
10      5 1965.707 59.46064  0.5393566    Well    Well  5 1965.707 1976.245   0 1906.246
11      5 1966.246 60.00000  9.9986311    Well    Well  5 1965.707 1976.245   0 1906.246
12      6 1965.203 50.98152  3.4606434    Well    Well  6 1965.203 1968.664   0 1914.222
13      7 1958.873 45.13895  4.8610541    Well    Well  7 1958.873 1976.916   0 1913.734
14      7 1963.734 50.00000 10.0000000    Well    Well  7 1958.873 1976.916   0 1913.734
15      7 1973.734 60.00000  3.1813826    Well    Well  7 1958.873 1976.916   0 1913.734
16      8 1965.370 50.42847  9.5715264    Well    Well  8 1965.370 1976.916   0 1914.942
17      8 1974.942 60.00000  1.9739904    Well    Well  8 1965.370 1976.916   0 1914.942
18      9 1959.125 67.09651  2.8993840    Well    Well  9 1959.125 1962.025   0 1892.029
19     10 1964.538 60.16701  9.8316222    Well    Well 10 1964.538 1974.370   0 1904.371
20     11 1964.790 60.52293  9.4757016    Well    Well 11 1964.790 1974.266   0 1904.267
   job month  energy   height   weight    fat fibre
1    0     8 2.87395 175.3870 71.48737 141.71 17.83
2    0     8 2.87395 175.3870 71.48737 141.71 17.83
3    0     8 2.87395 175.3870 71.48737 141.71 17.83
4    0    12 1.98234 164.2872 70.08120  85.77  9.49
5    0    12 1.98234 164.2872 70.08120  85.77  9.49
6    0    11 2.66858 169.3926 71.89560 107.67 15.99
7    0    11 2.66858 169.3926 71.89560 107.67 15.99
8    0     9 2.83669 167.0050 74.88937 132.17 17.04
9    0     9 2.83669 167.0050 74.88937 132.17 17.04
10   0     9 2.94150 174.4980 78.38208 126.35 14.54
11   0     9 2.94150 174.4980 78.38208 126.35 14.54
12   0     3 2.47351 176.5046 72.39456 103.10 12.49
13   0    11 2.55554 168.9100 64.18440 111.54 16.35
14   0    11 2.55554 168.9100 64.18440 111.54 16.35
15   0    11 2.55554 168.9100 64.18440 111.54 16.35
16   0     5 2.98756 165.9890 73.80072 159.53 16.09
17   0     5 2.98756 165.9890 73.80072 159.53 16.09
18   0     2 2.31124 165.7096 49.07952 115.68 15.44
19   0     7 3.12495 181.2036 78.29137 114.43 18.89
20   0    10 2.16144 174.1932 63.91224 111.16 11.66
```

7. In order to reproduce the table of events and person-years in in Clayton & Hills we first use the function `timeBand` to produce a factor with one level for each of the age-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  40 50 60 70
 - attr(*, "time.scales")= chr  "per" "age"
```

Then we can 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 )
```

```
          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
```

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

Now we 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/EpiE2012/data/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( cases ~ factor(age) + exposure,
+            offset = log(pyrs), family=poisson, data=ihdx )
> round( ci.lin( mt, E=T ), 3 )
```

```
             Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)    -5.303  0.439 -12.074 0.000     0.005 0.002 0.012
factor(age)1    0.204  0.472   0.432 0.666     1.226 0.486 3.092
factor(age)2    0.747  0.461   1.618 0.106     2.110 0.854 5.211
exposure        0.622  0.303   2.054 0.040     1.862 1.029 3.370
```

8. We can now also estimate the effect of age and exposure from the split dataset. We exclude follow-uptime before age 40 — as we saw from the table above, there is some risk time (but no events) before age 40.

```
> levels( Ls$agr )
```

```
[1] "(-Inf,40]" "(40,50]"   "(50,60]"   "(60,70]"   "(70,Inf]"
```

```
> 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
1      1 1964.623 49.61533  0.384668    Well    Well  1 1964.623 1976.916   0 1915.008
2      1 1965.008 50.00000 10.000000    Well    Well  1 1964.623 1976.916   0 1915.008
3      1 1975.008 60.00000  1.908282    Well    Well  1 1964.623 1976.916   0 1915.008
4      2 1964.957 50.53799  9.462012    Well    Well  2 1964.957 1976.916   0 1914.419
5      2 1974.419 60.00000  2.496920    Well    Well  2 1964.957 1976.916   0 1914.419
6      3 1965.874 58.78439  1.215606    Well    Well  3 1965.874 1976.916   0 1907.090
  job month  energy   height   weight    fat fibre     agr eksp
1   0     8 2.87395 175.3870 71.48737 141.71 17.83 (40,50] High
2   0     8 2.87395 175.3870 71.48737 141.71 17.83 (50,60] High
3   0     8 2.87395 175.3870 71.48737 141.71 17.83 (60,70] High
4   0    12 1.98234 164.2872 70.08120  85.77  9.49 (50,60]  Low
5   0    12 1.98234 164.2872 70.08120  85.77  9.49 (60,70]  Low
6   0    11 2.66858 169.3926 71.89560 107.67 15.99 (50,60]  Low
```

With this restriction we can now fit the model to the individual-record time-split data, and verify that we get the same:

```
> mi <- glm( (lex.Xst=="IHD") ~ factor(agr) + eksp,
+            offset = log(lex.dur), family=poisson, data=Ls )
> round( ci.lin( mi, E=T ), 3 )


                  Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)         -5.303  0.439 -12.074 0.000     0.005 0.002 0.012
factor(agr)(50,60]   0.204  0.472   0.432 0.666     1.226 0.486 3.092
factor(agr)(60,70]   0.747  0.461   1.619 0.106     2.110 0.854 5.210
ekspLow              0.622  0.303   2.054 0.040     1.862 1.029 3.370


> round( ci.lin( mt, E=T ), 3 )


            Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)   -5.303  0.439 -12.074 0.000     0.005 0.002 0.012
factor(age)1   0.204  0.472   0.432 0.666     1.226 0.486 3.092
factor(age)2   0.747  0.461   1.618 0.106     2.110 0.854 5.211
exposure       0.622  0.303   2.054 0.040     1.862 1.029 3.370


> ci.lin( mi, E=T ) / ci.lin( mt, E=T )


                  Estimate    StdErr        z         P exp(Est.)     2.5%     97.5%
(Intercept)       1.000002 0.9999537 1.000048       NaN 0.9999897 1.000030 0.9999499
factor(agr)(50,60] 1.000037 0.9999574 1.000080 0.9999625 1.0000076 1.000047 0.9999682
factor(agr)(60,70] 1.000011 0.9999515 1.000059 0.9998040 1.0000081 1.000052 0.9999643
ekspLow           1.000005 0.9999720 1.000033 0.9998334 1.0000034 1.000020 0.9999868
```

The point of using the individual data is that it the is possible to include individual-level variables in a model too.

9. We now an interaction between age and exposure and see that we 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: (lex.Xst == "IHD") ~ factor(agr) + eksp
Model 2: (lex.Xst == "IHD") ~ 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: cases ~ factor(age) + exposure
Model 2: cases ~ 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
```

10. The likelihood ratio statistic (`Chi-square`) is the same as the residual deviance in model without interaction for the grouped data, because the deviance is 0 for the interaction model for the grouped data, which is seen from the above output from the `anova` function.

5.4 Case-control study of BCG vaccination and leprosy.

The study is decribed by Clayton & Hills, p.156. In short, 260 cases of leprosy among individuals aged less than 35 years were ascertained in a study area in Malawi. Subjects were grouped into 7 age intervals and according to absence or presence of a scar after BCG vaccination. Three sets of controls were studied:

1. a population survey of 80,622 persons
2. a random sample of 1000 persons
3. a 4 to 1 age-matched sample

1. Fit the model from Clayton & Hills Table 23.5 (p.232).

   First we read the data:

```
> library( Epi )
> bcg <- read.table( "http://www.biostat.ku.dk/~pka/epidata/bcgalldata.txt",
+                    header=TRUE )
> summary( bcg )
```

```
      age         scar          status         n
 Min.   :1   Min.   :0.0   case    :14   Min.   :    1.00
 1st Qu.:2   1st Qu.:0.0   con1000 :14   1st Qu.:   24.75
 Median :4   Median :0.5   conall  :14   Median :   65.00
 Mean   :4   Mean   :0.5   conmatch:14   Mean   : 1480.75
 3rd Qu.:6   3rd Qu.:1.0                 3rd Qu.:  439.00
 Max.   :7   Max.   :1.0                 Max.   :11719.00
```

Then we can fit the logistic regression model with `age` and `scar`:

```
> m1 <- glm( status=="case" ~ factor(age) + factor(scar),
+            weight=n, family=binomial,
+            data=subset( bcg, status %in% c("case","conall") ) )
> round( ci.exp( m1 ), 3 )
```

```
               exp(Est.)   2.5%   97.5%
(Intercept)        0.000  0.000   0.001
factor(age)2      13.784  3.264  58.205
factor(age)3      35.985  8.754 147.921
factor(age)4      45.793 11.086 189.155
factor(age)5      49.410 11.925 204.725
factor(age)6      63.792 15.455 263.314
factor(age)7      63.920 15.521 263.243
factor(scar)1      0.579  0.439   0.763
```

The refernce group is the youngest (for the age-effect) and those without BCG-scar.

2. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference groups. If we wanted, say the last age-group as reference for the age-effects, we can change the reference level of the age-factor using `relevel`.

   We use the `update`-facility that takes a already fitted model (in this cas `m1`) as input, and allows us to change singe components of it. The "`.`"s mean "the same response" and "the same covariates" as in the original model.

```
> m1a <- update( m1, . ~ . - factor(age) + relevel(factor(age),7) )
> round( ci.exp( m1a ), 3 )
```

```
                         exp(Est.)  2.5% 97.5%
(Intercept)                  0.009 0.007 0.012
factor(scar)1                0.579 0.439 0.763
relevel(factor(age), 7)1     0.016 0.004 0.064
relevel(factor(age), 7)2     0.216 0.133 0.350
relevel(factor(age), 7)3     0.563 0.378 0.838
relevel(factor(age), 7)4     0.716 0.466 1.101
relevel(factor(age), 7)5     0.773 0.501 1.192
relevel(factor(age), 7)6     0.998 0.673 1.481
```

3. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference. This is just the 5th age-class:

```
> m1a <- update( m1, . ~ . - factor(age) + relevel(factor(age),5) )
> round( ci.exp( m1a ), 3 )
```

```
                         exp(Est.)  2.5% 97.5%
(Intercept)                  0.007 0.005 0.010
factor(scar)1                0.579 0.439 0.763
relevel(factor(age), 5)1     0.020 0.005 0.084
relevel(factor(age), 5)2     0.279 0.169 0.462
relevel(factor(age), 5)3     0.728 0.478 1.110
```

```
relevel(factor(age), 5)4     0.927 0.601 1.429
relevel(factor(age), 5)6     1.291 0.832 2.004
relevel(factor(age), 5)7     1.294 0.839 1.995
```

We see that the estiamte of the scar effect is the same.

4. Test the hypothesis of no interaction between `age` and `scar`.

   Again this is just using the `update` to add an interaction:

```
> mi <- update( m1, . ~ . + factor(age):factor(scar) )
> anova( m1, mi, test="Chisq" )


Analysis of Deviance Table

Model 1: status == "case" ~ factor(age) + factor(scar)
Model 2: status == "case" ~ factor(age) + factor(scar) + factor(age):factor(scar)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        20     3288.0
2        14     3284.4  6   3.6002   0.7306
```

   We see there is no sign of interaction.

5. Analyse the data set with only 1000 controls (i.e., use the controls `con1000`: Table 23.6, p.233) and compare the precision of the estimate for `scar` with that based on the entire sample.

   Once mode we can use `update` to change the dataset used to fit the model; the rest is unchanged:

```
> m1000 <- update( m1, data=subset( bcg, status %in% c("case","con1000") ) )
> round( rbind( ci.lin( m1    , subset="scar", E=T ),
+               ci.lin( m1000, subset="scar", E=T ) ), 3 )


              Estimate StdErr      z     P exp(Est.)  2.5% 97.5%
factor(scar)1   -0.547  0.141 -3.882 0.000     0.579 0.439 0.763
factor(scar)1   -0.548  0.160 -3.414 0.001     0.578 0.422 0.792
```

   We see that ths.e. of the log-odds-ratio increases from 0.14 to 0.16, some 15%.

6. Analyse the matched data set (i.e., use the controls `conmatch`: Table 23.6, p.233) and compare with the results from Table 23.7.

   The machinery is the sam as before:

```
> mmatch <- update( m1, data=subset( bcg, status %in% c("case","conmatch") ) )
> round( rbind( ci.lin( m1    , subset="scar", E=T ),
+               ci.lin( m1000 , subset="scar", E=T ),
+               ci.lin( mmatch, subset="scar", E=T ) ), 3 )


              Estimate StdErr      z     P exp(Est.)  2.5% 97.5%
factor(scar)1   -0.547  0.141 -3.882 0.000     0.579 0.439 0.763
factor(scar)1   -0.548  0.160 -3.414 0.001     0.578 0.422 0.792
factor(scar)1   -0.572  0.155 -3.699 0.000     0.564 0.417 0.764
```

   — and we see a small improvement in the s.e. of the log-OR even if the number of cases is (approximately) the same:

```
> xtabs( n ~ status, data=bcg )
```

```
status
    case  con1000   conall conmatch
     260     1000    80622     1040
```

7. Try (erroneously) to drop `age` from the analysis of the matched data and study the consequences for the estimate of `scar`.

   Again this is very simply obtained by update and repeating the code from before:

```
> msimp <- update( mmatch, . ~ . - factor(age) )
> round( rbind( ci.lin( mmatch, subset="scar", E=T ),
+               ci.lin( msimp , subset="scar", E=T ) ), 3 )

              Estimate StdErr      z     P exp(Est.)  2.5% 97.5%
factor(scar)1   -0.572  0.155 -3.699 0.000     0.564 0.417 0.764
factor(scar)1   -0.477  0.142 -3.369 0.001     0.621 0.470 0.819
```

   We see that the s.e. of the log-OR decreases, but also that the OR is now biased, the log-OR shrinks numerically by some 15%; the bias is about 2/3 of the estimated s.e. — quite substantial!

5.5 Case-control study of malignant melanoma.

Anne Østerlind conducted in the middle of the 80's a case-control study of risk factors for malignant melanoma in Denmark.

The review paper "Malignant melanoma in Denmark" from *Acta Oncologica*, 1990[3] is from Anne Østerlind's thesis and gives an overview of the results from the study which included 1400 interviewed persons, 474 cases and 926 controls, cf. table 5 in the article.

In the article incidence changes between 1943 and 1982 are also discussed; that part of the paper will not be touched upon in this exercise.

5.5.1 Discussion of the article.

1. Explain the design, the data base and data collection, particularly how the matching was conducted.

   The study is a group-matched (stratified) case-control study. Based on knowledge of the age- and sex-distribution of melanoma cases in Denmark a sample of the population with the same age- and sex-distribution was requested from the CPR.

   The data base (study base) is persons living in the eastern part of Denmark.

   Data were collected by personal interview.

2. How were interviews planned to minimize bias?

   Interviewers were blinded to the case/control status of the persons they interviewed for the study.

3. Explain the drop-out, particularly the analyses in Tables 5-7. What are the consequences of these results for the subsequent analyses?

   The drop-outs are differential between age-classes, older seem to be less likely to respond. It would have been more informative to have had the response probabilities

for each subcategory, instead of the relative distribution separately for responders and non-responders.

4. How are the analyses carried out? Are all variables included in one step or are the analyses conducted in smaller steps? How are the matching variables accounted for? Comments?

   Analyses are conducted first with one variable at a time, giving marginal results, and then jointly to give mutually controlled results.

   Sex, but not age, is included in the controlled analyses, so strictly speaking there is a potential bias. Particularly if confounding is expected, that is if the distribution of the risk factors are different across age-classes. This is however not particularly likely.

5. Explain the analyses presented in Table 9. How many logistic regression models are fitted here?

   Table 9 contains 6 analyses of one variable and 1 analysis with all variables, that is a total of 7 analyses.

6. What is the conclusion from the analyses in the table?

   Table 9 seems to indicate that freckles and to some extent skin and hair colour are the major (non-modifiable) risk factors for melanoma of the skin.

7. What is the purpose of Table 11?

   Table 11 shows the joint effect of freckles and naevi on the risk of melanoma. Clearly the risk increases by both, but there is no assessment of whether an interaction is present or not. Incidentally, as you will see, there is none.

8. Which modifiable factors seem to affect the melanoma risk?

   Number of sunburns before age 15. So one would in general expect that sunlight exposure increases the occurrence of melanoma. However detailed assessment of exposure history for each piece of skin for each study participant is of course impossible. So this seems to be pretty much as good as it gets.

5.5.2 Melanoma data

5.5.3 Simple tabulation analysis

First we read the data. Remember to specify how the missing values are coded.

```
> mel <- read.table( "http://www.biostat.ku.dk/~pka/epidata/melanom.txt",
+                    header=TRUE, na.strings="." )
> str( mel )
```

```
'data.frame':        1400 obs. of  14 variables:
 $ casecon : int  1 1 1 0 1 0 0 0 0 1 ...
 $ sex     : int  2 1 2 2 2 2 2 1 2 2 ...
 $ ageint  : int  71 68 42 66 36 68 68 39 75 49 ...
 $ agroup  : int  70 60 40 60 30 60 60 30 70 40 ...
 $ 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 ...
$ acuterea: int  3 3 4 4 4 3 4 4 2 1 ...
$ chronrea: int  NA 2 2 1 2 2 2 2 4 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 ...
$ nvtot   : int  2 3 23 0 1 0 0 3 5 6 ...
$ burn15  : int  1 NA NA NA NA NA 0 0 1 4 ...
```

```
> library( Epi )
```

1. Make the two by two table showing the association between case-control status and whether or not the person experienced *any* sunburns before the age of 15.

```
> bb <- with ( mel, table(casecon,sunburn=burn15>0) )
> bb
```

```
       sunburn
casecon FALSE TRUE
      0   277  236
      1    93  164
```

Estimate the odds ratio with associated 95% confidence limits and test for no association between the risk factor and case-control status.

```
> twoby2( bb[2:1,2:1] )
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : TRUE
Comparing : 1 vs. 0

  TRUE FALSE    P(TRUE) 95% conf. interval
1  164    93     0.6381    0.5776   0.6946
0  236   277     0.4600    0.4173   0.5034

                                  95% conf. interval
            Relative Risk: 1.3871    1.2163   1.5819
         Sample Odds Ratio: 2.0698    1.5211   2.8164
Conditional MLE Odds Ratio: 2.0678    1.5043   2.8532
    Probability difference: 0.1781    0.1038   0.2488

            Exact P-value: 0
        Asymptotic P-value: 0
------------------------------------------------------
```

We see that the p-value for the hypothesis of no association is effectively 0, the OR is 2.1 (1.5;2.8).

A little more intuitive machinery is the `effx`-function:

```
> effx( casecon, "binary", exposure=(burn15>0), data=mel )
```

```
------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  (burn15 > 0)

(burn15 > 0) is numeric
effects are measured as odds ratios
```

```
------------------------------------------------------------------------------

effect of an increase of 1 unit in (burn15 > 0) on casecon
number of observations  770

Effect   2.5%  97.5%
  2.07   1.52   2.82

Test for no effects of exposure on 1 df: p-value= 2.75e-06
```

2. Conduct similar analyses for the factors `sex, hair, eyes, freckles, acuterea, chronrea`. Compare with Table 9 in the article.

   The simple approach to all analyses would then be:

```
> effx( casecon, "binary", exposure=factor(sex), data=mel )
> effx( casecon, "binary", exposure=factor(hair), data=mel )
> effx( casecon, "binary", exposure=factor(freckles), data=mel )
> effx( casecon, "binary", exposure=factor(acuterea), data=mel )
> effx( casecon, "binary", exposure=factor(chronrea), data=mel )
```

   — but it is a bit more readable if we actually code the factors first (we do it here for all variables of interest):

```
> mel <- transform( mel, sex = factor(sex,labels=c("M","F")),
+                        skin = factor(skin,labels=c("dark","medium","light")),
+                        hair = factor(hair,labels=c("dark","light","blond","red")),
+                        eyes = factor(eyes,labels=c("brown","green/gray","blue")),
+                   freckles = factor(freckles,labels=c("many","some","none")),
+                   acuterea = factor(acuterea,labels=c("blisters","painful","mild","none")),
+                   chronrea = factor(chronrea,labels=c("deep","moderate","mild","none")) )
```

   and then just give the reference level we want in text (must be one of the defined levels):

```
> effx( casecon, "binary", data=mel, exposure=sex, base="F" )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  sex

sex is a factor with levels: M / F
baseline is  F
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of sex on casecon
number of observations  1400

Effect   2.5%  97.5%
 0.952  0.761  1.190

Test for no effects of exposure on 1 df: p-value= 0.669


> effx( casecon, "binary", data=mel, exposure=hair, base="dark" )
```

```
------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  hair

hair is a factor with levels: dark / light / blond / red
baseline is  dark
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of hair on casecon
number of observations  1400

              Effect  2.5% 97.5%
light vs dark   1.49 1.170  1.89
blond vs dark   1.70 0.995  2.91
red vs dark     1.74 1.140  2.68

Test for no effects of exposure on 3 df: p-value= 0.0017


> effx( casecon, "binary", data=mel, exposure=eyes, base="brown" )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  eyes

eyes is a factor with levels: brown / green/gray / blue
baseline is  brown
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of eyes on casecon
number of observations  1394

                    Effect  2.5% 97.5%
green/gray vs brown   0.85 0.592  1.22
blue vs brown         1.06 0.756  1.48

Test for no effects of exposure on 2 df: p-value= 0.22


> effx( casecon, "binary", data=mel, exposure=freckles, base="none" )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  freckles

freckles is a factor with levels: many / some / none
baseline is  none
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of freckles on casecon
number of observations  1396

             Effect 2.5% 97.5%
many vs none   3.01 2.21  4.11
some vs none   1.49 1.16  1.92

Test for no effects of exposure on 2 df: p-value= 2.15e-11
```

```
> effx( casecon, "binary", data=mel, exposure=acuterea, base="none" )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  acuterea

acuterea is a factor with levels: blisters / painful / mild / none
baseline is  none
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of acuterea on casecon
number of observations  1388

                 Effect  2.5% 97.5%
blisters vs none   2.16 0.929  5.02
painful vs none    1.64 1.050  2.56
mild vs none       1.27 0.987  1.63

Test for no effects of exposure on 3 df: p-value= 0.0497


> effx( casecon, "binary", data=mel, exposure=chronrea, base="deep" )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  chronrea

chronrea is a factor with levels: deep / moderate / mild / none
baseline is  deep
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of chronrea on casecon
number of observations  1394

                 Effect 2.5% 97.5%
moderate vs deep   1.39 1.07  1.80
mild vs deep       1.84 1.33  2.56
none vs deep       1.96 1.03  3.73

Test for no effects of exposure on 3 df: p-value= 0.00142
```

3. The case control study was matched for sex and age and, therefore, analyses of any risk factor should be adjusted for these two variables. Study how much the association between the risk factor "any sunburns before the age of 15" and case-control status is affected by adjustment for sex.

   This can also be done by using the `effx` machine:

```
> effx( casecon, "binary", exposure=(burn15>0), data=mel )


------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  (burn15 > 0)

(burn15 > 0) is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------
```

```
effect of an increase of 1 unit in (burn15 > 0) on casecon
number of observations  770

Effect   2.5%  97.5%
  2.07   1.52   2.82

Test for no effects of exposure on 1 df: p-value= 2.75e-06
```

```
> effx( casecon, "binary", exposure=(burn15>0), data=mel, control=sex )
```

```
------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  (burn15 > 0)
control vars  :  sex

(burn15 > 0) is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of an increase of 1 unit in (burn15 > 0) on casecon
controlled for sex

number of observations  770

Effect   2.5%  97.5%
  2.06   1.52   2.81

Test for no effects of exposure on 1 df: p-value= 3.24e-06
```

4. Same question for age.

```
> effx( casecon, "binary", exposure=(burn15>0), data=mel, control=agroup )
```

```
------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  (burn15 > 0)
control vars  :  agroup

(burn15 > 0) is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of an increase of 1 unit in (burn15 > 0) on casecon
controlled for agroup

number of observations  770

Effect   2.5%  97.5%
  2.06   1.51   2.83

Test for no effects of exposure on 1 df: p-value= 5.06e-06
```

```
> effx( casecon, "binary", exposure=(burn15>0), data=mel, control=factor(agroup) )
```

```
------------------------------------------------------------------------------
response      :  casecon
type          :  binary
exposure      :  (burn15 > 0)
control vars  :  agroup

(burn15 > 0) is numeric
effects are measured as odds ratios
------------------------------------------------------------------------------

effect of an increase of 1 unit in (burn15 > 0) on casecon
controlled for agroup

number of observations  770

Effect   2.5%  97.5%
  2.04   1.49   2.80

Test for no effects of exposure on 1 df: p-value= 7.94e-06
```

5.5.4 Introductory analyses.

1. Estimate (log-)odds ratios for the variable `skin` (see top left in AØ's Table 9).

   The log-odds-ratio is what we get from a logistic regression model:

```
> mskin <- glm( casecon ~ skin, family=binomial, data=mel )
> summary( mskin )


Call:
glm(formula = casecon ~ skin, family = binomial, data = mel)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9790  -0.9145  -0.7941   1.3897   1.6172

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9924     0.1262  -7.861 3.82e-15
skinmedium    0.3369     0.1530   2.201  0.02772
skinlight     0.5060     0.1575   3.213  0.00132

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1779.9  on 1389  degrees of freedom
Residual deviance: 1769.3  on 1387  degrees of freedom
  (10 observations deleted due to missingness)
AIC: 1775.3

Number of Fisher Scoring iterations: 4
```

2. Estimate also odds ratios.

   We can get both log-OR and OR using `ci.lin`:

```
> round( ci.lin( mskin, E=T ), 3 )


            Estimate StdErr      z     P exp(Est.)  2.5% 97.5%
(Intercept)   -0.992  0.126 -7.861 0.000     0.371 0.289 0.475
skinmedium     0.337  0.153  2.201 0.028     1.401 1.038 1.891
skinlight      0.506  0.158  3.213 0.001     1.659 1.218 2.259
```

3. Conduct the other analyses in AØ's Table 9 (*left* part) where the factors `hair, eyes, freckles, acuterea, chronrea` are studied one at a time.

   This is just a series of logistic regression models as above, but we must remember to relevel the factors where the desired reference is not the first. We use `Relevel` (with capital "R"), which allows a reordering of levels

```
> mhair <- glm( casecon ~ hair, family=binomial, data=mel )
> meyes <- glm( casecon ~ eyes, family=binomial, data=mel )
> mfrek <- glm( casecon ~ Relevel(freckles,3:1), family=binomial, data=mel )
> macut <- glm( casecon ~ Relevel(acuterea,4:1), family=binomial, data=mel )
> mchro <- glm( casecon ~ chronrea, family=binomial, data=mel )
> round( rbind( ci.lin(mhair,Exp=TRUE)[-1,],
+               ci.lin(meyes,Exp=TRUE)[-1,],
+               ci.lin(mfrek,Exp=TRUE)[-1,],
+               ci.lin(macut,Exp=TRUE)[-1,],
+               ci.lin(mchro,Exp=TRUE)[-1,] ), 3 )
```

```
                                Estimate StdErr      z     P exp(Est.)  2.5%
hairlight                          0.397  0.122  3.262 0.001     1.487 1.172
hairblond                          0.531  0.274  1.943 0.052     1.701 0.995
hairred                            0.556  0.219  2.544 0.011     1.744 1.136
eyesgreen/gray                    -0.162  0.185 -0.878 0.380     0.850 0.592
eyesblue                           0.058  0.172  0.336 0.737     1.059 0.756
Relevel(freckles, 3:1)some         0.398  0.129  3.097 0.002     1.489 1.157
Relevel(freckles, 3:1)many         1.102  0.158  6.966 0.000     3.011 2.208
Relevel(acuterea, 4:1)mild         0.237  0.128  1.856 0.063     1.268 0.987
Relevel(acuterea, 4:1)painful      0.493  0.229  2.155 0.031     1.638 1.046
Relevel(acuterea, 4:1)blisters     0.770  0.430  1.789 0.074     2.160 0.929
chronreamoderate                   0.326  0.133  2.448 0.014     1.386 1.067
chronreamild                       0.611  0.167  3.653 0.000     1.842 1.327
chronreanone                       0.671  0.329  2.039 0.041     1.956 1.026
                                97.5%
hairlight                       1.888
hairblond                       2.908
hairred                         2.677
eyesgreen/gray                  1.221
eyesblue                        1.484
Relevel(freckles, 3:1)some      1.915
Relevel(freckles, 3:1)many      4.105
Relevel(acuterea, 4:1)mild      1.628
Relevel(acuterea, 4:1)painful   2.565
Relevel(acuterea, 4:1)blisters  5.022
chronreamoderate                1.800
chronreamild                    2.556
chronreanone                    3.728
```

4. Conduct the analysis corresponding to Table 9 (*right* part) where several variables are included simultaneously (see the table footnote).

   This just mean that we include all variables plus sex in the model:

```
> m.all <- glm( casecon ~ skin +
+                         hair +
+                         eyes +
+                         Relevel(freckles,3:1) +
+                         Relevel(acuterea,4:1) +
+                         chronrea, family=binomial, data=mel )
> round( ci.exp( m.all )[-1,], 2 )
```

```
                           exp(Est.) 2.5% 97.5%
skinmedium                      1.32 0.96  1.80
```

```
skinlight                          1.32 0.93  1.87
hairlight                          1.47 1.14  1.90
hairblond                          1.59 0.90  2.81
hairred                            1.25 0.79  1.99
eyesgreen/gray                     0.74 0.50  1.09
eyesblue                           0.89 0.62  1.28
Relevel(freckles, 3:1)some         1.49 1.15  1.93
Relevel(freckles, 3:1)many         2.92 2.11  4.04
Relevel(acuterea, 4:1)mild         1.10 0.85  1.44
Relevel(acuterea, 4:1)painful      1.37 0.85  2.19
Relevel(acuterea, 4:1)blisters     1.68 0.70  4.04
chronreamoderate                   1.21 0.91  1.60
chronreamild                       1.40 0.96  2.02
chronreanone                       1.15 0.54  2.42
```

5. Reconstruct the results from AØ's Table 10 concerning number of raised naevi.

   We first define new variables, even if we could define them on the fly, it would be too cumbersome. There is a `cut` function in R, which groups numerical variables; it needs a bit of care. Note that it is useful that R known the number `Inf`: ∞.

```
> mel <- transform( mel, gnvsmall = cut(nvsmall,breaks=c(0,1,2,5,Inf),right=FALSE),
+                        gnvlarge = cut(nvlarge,breaks=c(0,1,2,  Inf),right=FALSE),
+                        gnvtot   = cut(nvtot  ,breaks=c(0,1,2,5,Inf),right=FALSE) )
```

   We then fit the models behind the numbers in table 10, first the crude model, then the adjusted:

```
> c.small <- glm( casecon ~ gnvsmall, family=binomial, data=mel )
> c.large <- glm( casecon ~ gnvlarge, family=binomial, data=mel )
> c.tot   <- glm( casecon ~ gnvtot  , family=binomial, data=mel )
> a.small <- update( c.small, . ~ . + sex + freckles + hair + skin )
> a.large <- update( c.large, . ~ . + sex + freckles + hair + skin )
> a.tot   <- update( c.tot  , . ~ . + sex + freckles + hair + skin )
```

   It is then reasonably straightforward to assemble the estimated odds-ratio in the same format as in table 10:

```
> round( cbind( rbind( ci.exp( c.tot  , subset="gnv" ),
+                      ci.exp( c.small, subset="gnv" ),
+                      ci.exp( c.large, subset="gnv" ) ),
+               rbind( ci.exp( a.tot  , subset="gnv" ),
+                      ci.exp( a.small, subset="gnv" ),
+                      ci.exp( a.large, subset="gnv" ) ) ), 2 )
```

```
               exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5%
gnvtot[1,2)         1.51 1.10  2.07      1.49 1.08  2.06
gnvtot[2,5)         2.26 1.64  3.11      2.22 1.59  3.09
gnvtot[5,Inf)       5.36 3.54  8.11      4.91 3.20  7.54
gnvsmall[1,2)       1.58 1.14  2.19      1.59 1.13  2.22
gnvsmall[2,5)       2.45 1.77  3.41      2.41 1.72  3.40
gnvsmall[5,Inf)     5.00 3.25  7.69      4.73 3.03  7.38
gnvlarge[1,2)       1.82 1.20  2.77      1.65 1.07  2.54
gnvlarge[2,Inf)     3.57 1.78  7.16      2.66 1.30  5.46
```

   We see that the number of naevi seems to be a predictor of melanoma risk.

5.5.5 Trend tests and interactions.

6. In the analyses so far all variables have been considered as categorical (factors) while all tests in Tables 9 and 10 in the paper are trend tests.

   We now do the analysis behind the trend test for freckles in the joint model by replacing the factor version of freckles with a continuous, scored variable. The *numbers* of the factor levels is achieved in R by `as.integer`, so the model is obtained by:

```
> lfrec <- update( m.all, . ~ . - Relevel(freckles,3:1) + as.integer(freckles) )
```

   The trend test is simply the test for whether the linear of `freckles` is 0:

```
> round( ci.lin( lfrec, E=T ), 3 )
```

```
                              Estimate StdErr      z     P exp(Est.)  2.5%
(Intercept)                     -0.013  0.284 -0.046 0.963     0.987 0.566
skinmedium                       0.274  0.161  1.701 0.089     1.315 0.959
skinlight                        0.284  0.177  1.611 0.107     1.329 0.940
hairlight                        0.387  0.130  2.967 0.003     1.472 1.140
hairblond                        0.443  0.288  1.541 0.123     1.558 0.886
hairred                          0.238  0.235  1.014 0.311     1.269 0.801
eyesgreen/gray                  -0.308  0.196 -1.570 0.116     0.735 0.500
eyesblue                        -0.124  0.185 -0.672 0.501     0.883 0.615
Relevel(acuterea, 4:1)mild       0.103  0.135  0.767 0.443     1.109 0.851
Relevel(acuterea, 4:1)painful    0.307  0.241  1.274 0.203     1.360 0.847
Relevel(acuterea, 4:1)blisters   0.490  0.446  1.098 0.272     1.632 0.681
chronreamoderate                 0.188  0.144  1.306 0.191     1.206 0.910
chronreamild                     0.340  0.189  1.799 0.072     1.404 0.970
chronreanone                     0.158  0.379  0.417 0.676     1.171 0.558
as.integer(freckles)            -0.514  0.081 -6.378 0.000     0.598 0.510
                              97.5%
(Intercept)                   1.721
skinmedium                    1.803
skinlight                     1.878
hairlight                     1.900
hairblond                     2.739
hairred                       2.013
eyesgreen/gray                1.079
eyesblue                      1.269
Relevel(acuterea, 4:1)mild    1.445
Relevel(acuterea, 4:1)painful 2.183
Relevel(acuterea, 4:1)blisters 3.914
chronreamoderate              1.598
chronreamild                  2.034
chronreanone                  2.460
as.integer(freckles)          0.700
```

   We see that there is a strong linear effect of `freckles`.

   For skin we obtain:

```
> lskin <- update( m.all, . ~ . - skin + as.integer(skin) )
> round( ci.lin( lskin, E=T ), 3 )
```

```
                              Estimate StdErr      z     P exp(Est.)  2.5%
(Intercept)                     -1.564  0.256 -6.102 0.000     0.209 0.127
hairlight                        0.392  0.130  3.012 0.003     1.480 1.147
hairblond                        0.461  0.289  1.599 0.110     1.586 0.901
hairred                          0.227  0.236  0.960 0.337     1.255 0.790
eyesgreen/gray                  -0.308  0.196 -1.572 0.116     0.735 0.500
```

```
eyesblue                          -0.127  0.185 -0.687 0.492     0.881 0.613
Relevel(freckles, 3:1)some         0.400  0.132  3.025 0.002     1.492 1.151
Relevel(freckles, 3:1)many         1.068  0.166  6.452 0.000     2.911 2.104
Relevel(acuterea, 4:1)mild         0.098  0.135  0.728 0.467     1.103 0.847
Relevel(acuterea, 4:1)painful      0.307  0.242  1.271 0.204     1.359 0.847
Relevel(acuterea, 4:1)blisters     0.514  0.448  1.145 0.252     1.671 0.694
chronreamoderate                   0.198  0.143  1.379 0.168     1.219 0.920
chronreamild                       0.328  0.189  1.734 0.083     1.388 0.958
chronreanone                       0.123  0.381  0.323 0.747     1.131 0.536
as.integer(skin)                   0.124  0.087  1.434 0.152     1.132 0.955
                                 97.5%
(Intercept)                      0.346
hairlight                        1.910
hairblond                        2.793
hairred                          1.994
eyesgreen/gray                   1.079
eyesblue                         1.266
Relevel(freckles, 3:1)some       1.934
Relevel(freckles, 3:1)many       4.026
Relevel(acuterea, 4:1)mild       1.437
Relevel(acuterea, 4:1)painful    2.182
Relevel(acuterea, 4:1)blisters   4.024
chronreamoderate                 1.614
chronreamild                     2.012
chronreanone                     2.386
as.integer(skin)                 1.341
```

so there is no significant trend across skin-categories.

7. May `freckles` be scored linearly (1, 2, 3), when this variable is studied separately?

   The test of linearity is simply comparison of the model `freckles` as factor and the model with linear effect of `freckles`:

```
> anova( lfrec, m.all, test="Chisq" )


Analysis of Deviance Table

Model 1: casecon ~ skin + hair + eyes + Relevel(acuterea, 4:1) + chronrea +
    as.integer(freckles)
Model 2: casecon ~ skin + hair + eyes + Relevel(freckles, 3:1) + Relevel(acuterea,
    4:1) + chronrea
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1353     1674.7
2      1352     1673.5  1   1.2179   0.2698
```

   The test is non-significant, that means that we can accept the null hypothesis of a linear trend across the categories of freckles. We can also make a similar test comparing models where only `freckles` enter:

```
> anova(         mfrek,
+       update( mfrek, . ~ . - Relevel(freckles,3:1) +   as.integer(freckles) ),
+       test = "Chisq" )


Analysis of Deviance Table

Model 1: casecon ~ Relevel(freckles, 3:1)
Model 2: casecon ~ as.integer(freckles)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1393     1738.5
2      1394     1740.1 -1  -1.6087   0.2047
```

Also in this case there is no evidence for departure from linearity.

8. In AØ's Table 11 `freckles` and the total number of naevi (suitably grouped) are studied.

   This is analysis of a model with interaction between `freckles` and `gnvtot`:

```
> ma <- update( mfrek, . ~ . + gnvtot )
> mi <- update( mfrek, . ~ Relevel(gnvtot,4:1):freckles )
> round( ci.exp( mi ), 2 )
```

```
                                          exp(Est.) 2.5% 97.5%
(Intercept)                                    0.27 0.21  0.34
Relevel(gnvtot, 4:1)[5,Inf):frecklesmany      13.92 6.17 31.41
Relevel(gnvtot, 4:1)[2,5):frecklesmany         6.10 3.02 12.33
Relevel(gnvtot, 4:1)[1,2):frecklesmany         4.77 2.29  9.96
Relevel(gnvtot, 4:1)[0,1):frecklesmany         2.51 1.64  3.82
Relevel(gnvtot, 4:1)[5,Inf):frecklessome       9.14 4.60 18.13
Relevel(gnvtot, 4:1)[2,5):frecklessome         2.66 1.61  4.41
Relevel(gnvtot, 4:1)[1,2):frecklessome         1.76 1.07  2.91
Relevel(gnvtot, 4:1)[0,1):frecklessome         1.48 1.06  2.08
Relevel(gnvtot, 4:1)[5,Inf):frecklesnone       2.89 1.38  6.03
Relevel(gnvtot, 4:1)[2,5):frecklesnone         2.53 1.52  4.19
Relevel(gnvtot, 4:1)[1,2):frecklesnone         1.71 1.05  2.80
Relevel(gnvtot, 4:1)[0,1):frecklesnone         1.00 1.00  1.00
```

   These are the ORs in table 10

   We can test for interaction between the two variables by comparing the interaction model with the model with main effects:

```
> anova( ma, mi, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: casecon ~ Relevel(freckles, 3:1) + gnvtot
Model 2: casecon ~ Relevel(gnvtot, 4:1):freckles
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1386     1666.3
2      1380     1661.0  6   5.2911   0.5071
```

   We see that there is no compelling indication of an interaction. However, the χ^2-test is 5.3, so there might possibly be a significant 1 df. interaction in there. So we try to add a linear interaction. But we first must make sue how the two factors are coded:

```
> levels( mel$gnvtot )
```

```
[1] "[0,1)"   "[1,2)"   "[2,5)"   "[5,Inf)"
```

```
> levels( mel$freckles )
```

```
[1] "many" "some" "none"
```

   Thus we can not just take the product of the numerical levels, because we want the product of two numerical values that both increase with increasing risk, in order to see if the risk increases more or less than predicted by the marginal effects:

```
> mel$frnv <- as.integer( mel$gnvtot ) * ( 4-as.integer(mel$freckles) )
> with( mel, tapply( frnv, list(freckles,gnvtot), median ) )
```

```
     [0,1) [1,2) [2,5) [5,Inf)
many     3     6     9      12
some     2     4     6       8
none     1     2     3       4
```

So we include the interaction and just look at the Wald-test for it

```
> round( ci.lin( update( ma, . ~ . + frnv ), E=T ), 3 )
```

```
                           Estimate StdErr       z     P exp(Est.)  2.5% 97.5%
(Intercept)                  -1.321  0.109 -12.113 0.000     0.267 0.216 0.330
Relevel(freckles, 3:1)some    0.243  0.187   1.301 0.193     1.275 0.884 1.840
Relevel(freckles, 3:1)many    0.801  0.319   2.511 0.012     2.229 1.192 4.166
gnvtot[1,2)                   0.319  0.213   1.492 0.136     1.375 0.905 2.090
gnvtot[2,5)                   0.602  0.324   1.861 0.063     1.826 0.969 3.443
gnvtot[5,Inf)                 1.253  0.492   2.548 0.011     3.502 1.335 9.185
frnv                          0.050  0.078   0.645 0.519     1.052 0.902 1.225
```

Clearly, the conclusion is that there is no interaction, and that the joint effect of freckles and total number of naevi is well described by the marginal effects:

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

```
                           exp(Est.)  2.5% 97.5%
(Intercept)                    0.273 0.223 0.335
Relevel(freckles, 3:1)some     1.389 1.072 1.799
Relevel(freckles, 3:1)many     2.660 1.929 3.668
gnvtot[1,2)                    1.502 1.090 2.070
gnvtot[2,5)                    2.184 1.579 3.022
gnvtot[5,Inf)                  4.662 3.057 7.110
```

9. Study, in a similar vein, interactions between `acuterea` and `skin` and between the grouped version of `nvtot` and `agroup`.

   The variables `skin`, `gnvtot` and `agroup` are coded so that that increasing levels correspond to increasing risk, but `acuterea` is coded the other way round, so again we need to be careful when defining numerical interactions

```
> levels( mel$skin )
```

```
[1] "dark"   "medium" "light"
```

```
> levels( mel$acuterea )
```

```
[1] "blisters" "painful"  "mild"     "none"
```

```
> levels( mel$gnvtot )
```

```
[1] "[0,1)"   "[1,2)"   "[2,5)"   "[5,Inf)"
```

```
> mel$skac <- as.integer(mel$skin) * (5-as.integer(mel$acuterea))
> mel$agnv <- mel$agroup * as.integer(mel$gnvtot)
```

We the fit the model with two variables separately, the model with the continuous interaction and the mode with the grouped interaction to see if there is a any interaction, and if it is easily interpretable.

First for `skin` and `acuterea`:

```
> msa  <- glm( casecon ~ skin + acuterea, family=binomial, data=mel )
> msal <- update( msa, . ~ . + skac )
> msai <- update( msa, . ~ . + skin:acuterea )
> anova( msa, msal, msai, test="Chisq" )


Analysis of Deviance Table

Model 1: casecon ~ skin + acuterea
Model 2: casecon ~ skin + acuterea + skac
Model 3: casecon ~ skin + acuterea + skin:acuterea
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1372     1750.9
2      1371     1747.9  1   3.0334  0.08157
3      1366     1734.4  5  13.4818  0.01926
```

We see there is a significant interaction between `skin` and `acuterea`, which is not captured by the linear term. Thus it is of course of relevance to see how it looks. To that end we re-fit the model with the interaction term alone, reparametrized, so that the dark skin and no acute reaction is the reference:

```
> msai <- update( msa, . ~ Relevel(skin,3:1):acuterea )
> round( ci.exp( msai ), 3 )


                                          exp(Est.)  2.5%  97.5%
(Intercept)                                   0.422 0.289  0.617
Relevel(skin, 3:1)light:acutereablisters      2.368 0.718  7.812
Relevel(skin, 3:1)medium:acutereablisters     1.421 0.323  6.247
Relevel(skin, 3:1)dark:acutereablisters       4.737 0.417 53.817
Relevel(skin, 3:1)light:acutereapainful       1.731 0.857  3.495
Relevel(skin, 3:1)medium:acutereapainful      1.442 0.671  3.098
Relevel(skin, 3:1)dark:acutereapainful        1.895 0.694  5.171
Relevel(skin, 3:1)light:acutereamild          1.766 1.135  2.749
Relevel(skin, 3:1)medium:acutereamild         1.235 0.797  1.913
Relevel(skin, 3:1)dark:acutereamild           0.687 0.407  1.160
Relevel(skin, 3:1)light:acutereanone          0.716 0.401  1.279
Relevel(skin, 3:1)medium:acutereanone         1.222 0.753  1.984
Relevel(skin, 3:1)dark:acutereanone           1.000 1.000  1.000
```

We can make a forest plot of this:

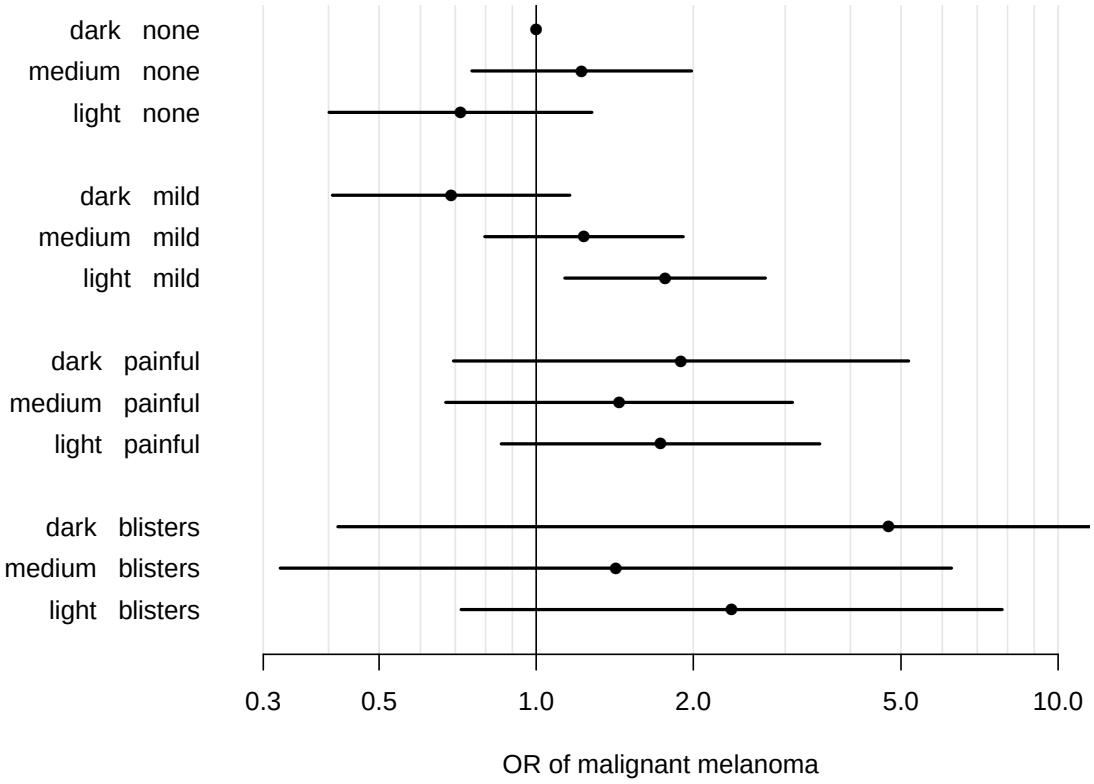
```
> plotEst( ci.exp(msai)[-1,] )
```

— which is rather uninformative, so we need to put limits on it

```
> plotEst( ci.exp(msai)[-1,], xlim=c(0.5,8), xlog=TRUE )
```

But the labeling is awful, and the arrangement of estimates can be improved:

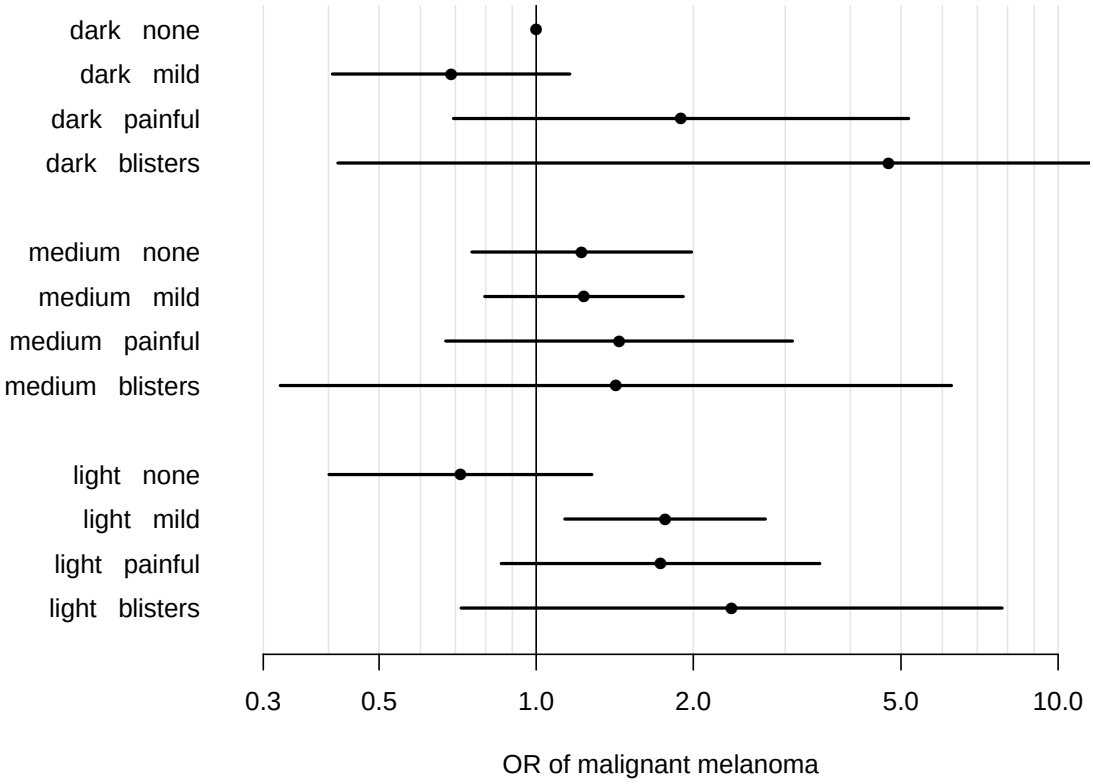
```
> ee <-  ci.exp(msai)[-1,]
> rownames( ee ) <- gsub( "Relevel\\(skin, 3:1\\)","",rownames(ee) )
> rownames( ee ) <- gsub( ":acuterea","   ",rownames(ee) )
> plotEst( ee,  xlog=TRUE, vref=1,
+          xtic=c(0.3,0.5,1,2,5,10),grid=c(3:9/10,2:10),
+          y=rep(1:3,4)+rep(c(0,4,8,12),each=3), xlab="OR of malignant melanoma" )
```



The interaction is significant, but there does not seem to any consistent pattern, except that is only among those with acute reaction mild that the skin color has an effect.

However, we can also take a look at the same estimates sorted differently:

```
> plotEst( ee,  xlog=TRUE, vref=1,
+          xtic=c(0.3,0.5,1,2,5,10),grid=c(3:9/10,2:10),
+          y=rep(1:3*5,4)+rep(1:4,each=3), xlab="OR of malignant melanoma" )
```



Looking at the interaction this way reveals that acute reaction has an effect only among light-skinned, and the difference is mainly between those with no acute reaction and those with (any).

Now we take a look at the interaction between `agroup` and `gnvtot`:

```
> mag  <- glm( casecon ~ factor(agroup) + gnvtot, family=binomial, data=mel )
> magl <- update( mag, . ~ . + agnv )
> magi <- update( mag, . ~ . + factor(agroup):gnvtot )
> anova( mag, magl, magi, test="Chisq" )

Analysis of Deviance Table

Model 1: casecon ~ factor(agroup) + gnvtot
Model 2: casecon ~ factor(agroup) + gnvtot + agnv
Model 3: casecon ~ factor(agroup) + gnvtot + factor(agroup):gnvtot
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1383     1701.1
2      1382     1698.8  1   2.3545   0.1249
3      1365     1685.0 17  13.7575   0.6842
```

And we see that there is no indication of any type of interaction (Phew!).

10. All of AØ's analyses are conducted without accounting for the match variable age (`agroup`) (in spite of warnings given by Clayton & Hills!). Repeat some of the previous analyses adjusting for `agroup`.

    We take a look at the effect of the variables skin color and freckles with and without adjustment for age-group:

```
> m0 <- glm( casecon ~ 1, family=binomial, data=mel )
> mfrek <- update( m0, .~. + Relevel(freckles,3:1) )
> mfrag <- update( mfrek, .~. + factor(agroup) )
> mskin <- update( m0, .~. + skin )
> mskag <- update( mskin, .~. + factor(agroup) )
> round( cbind( rbind( ci.exp( mfrek, subset="frec" ),
+                      ci.exp( mskin, subset="skin" ) ),
+               rbind( ci.exp( mfrag, subset="frec" ),
+                      ci.exp( mskag, subset="skin" ) ) ), 2 )
```

```
                            exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5%
Relevel(freckles, 3:1)some       1.49 1.16  1.92       1.50 1.16  1.93
Relevel(freckles, 3:1)many       3.01 2.21  4.11       3.04 2.22  4.17
skinmedium                       1.40 1.04  1.89       1.39 1.03  1.88
skinlight                        1.66 1.22  2.26       1.66 1.21  2.26
```

The first 3 columns are the unadjusted estimates, and the last three are the adjusted estimates; we see there is virtually no differences between them. This is because there is very little association between the risk factors and age.

5.6 Testicular cancer risk and maternal parity.

This exercise deals with the article "Testicular cancer risk and maternal parity: a population-based cohort study", by T. Westergaard, P.K. Andersen, J.B. Pedersen, M. Frisch, J.H. Olsen, M. Melbye. *Br. J. Cancer*, **77**,pp. 1180-1185 (1998). [4].

5.6.1 Discussion of the article.

1. What is the authors' argument for the existence of an effect of maternal parity on the risk of testicular cancer in the son?

   Mainly the existing literature on occurrence of testis cancer.

2. Describe the design of the study:

   - a. which "sons" are included in the study?

     Sons of mothers born after 1935; that is essentially only boys born after 1950. But not all such boys.

   - b. when are they followed?

     They are followed from the start of CPR, 1.1.1968 till the end of 1992.

   - c. how are cases defined and ascertained?

     Cases are defined as cases reported to the Danish Cancer registry with certain diagnoses.

3. Concentrating on all testicular cancers, what do you consider to be the main result reported in Table 1?

   The effect of parity 2+, `and` the remarkable small effect of any other variable.

4. Explain in words the interpretation of the value RR=0.80 for parity 2+.

   This means that the incidence rate of TC among boys born as the 2nd or later sibling of a woman is 0.8 of the rate among first-born.

5. Compare this value with the corresponding crude RR (and 95 % CI) obtained without any adjustment. Explain the differences between the two results.

   The crude RR was 0.57, but adjustment for the age at follow-up (called `SON_AGE` in the dataset), makes this effect go to to 0.8. This is because of confounding — incidence increases by age, and first born are likely to be older in a dataset collected as follow-up of cohorts.

6. Draw a Lexis diagram to illustrate the combinations of age and calendar period which contribute person-years to the study. This is shown in the figure.

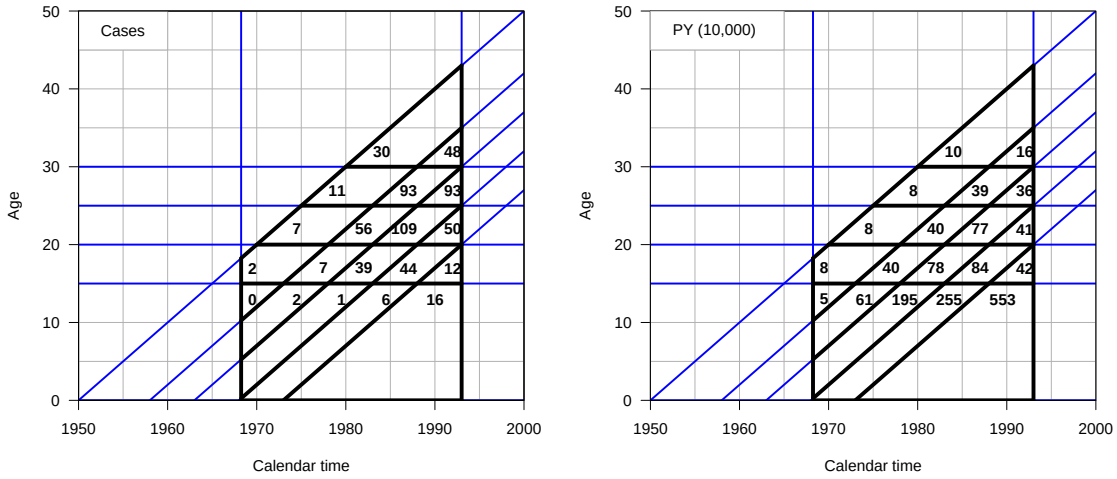


Figure 5.1: *Lexis diagrams showing the number of cases and person-years by age and date of follow-up. The black outline indicates the sampling frame for the study.*

   In the section with solutions in R is a small section illustration how to draw the Lexis diagrams with lines indicating age-classes and birth cohorts.

7. Explain the meaning of the estimates for "Interval from ..." in the lower part of Table 1.

   This interval basically gives the age-difference to the closest older sibling. This is meant as a proxy fro some sort of infection pressure in infancy.

8. What type of analysis is reported in Table 2?

   Table 2 reports *interactions*, investigating whether the effect of being 1st born is constant across other factor available in this study.

9. Discuss how, alternatively, a case-control design could have been conducted to address the same question as the cohort study reported in the article.

   A case-control study would include as cases the same cases as this cohort study, namely those ascertained from the Cancer Registry. As there is a substantial variation by age and birth-cohort in testis cancer incidence, which is well known, it would be natural to match each case to a number of controls born on the same day and who survived till the date of diagnosis of the case (for example all).

The maternal parameters would then be ascertained from the various registers for both cases and controls.

Even though an individual matching is done this way, the analysis needed not be done as matched (conditional) logistic regression, the matching variables, date of birth and age are perfectly quantifiable and could just be included as covariates in an ordinary logistic regression analysis.

The case-control study would not be able to assess the incidence rates, but in so far the aim of the analysis was to assess the effect of maternal parameters, this would not be of interest. The level and age-dependence of TC incidence rates are well known from descriptive studies based on the cancer registry data anyway.

5.6.2 The Lexis diagram

The Lexis diagram shown above is generated by the following R-code. This is somewhat outside the course but included for the completeness of documentation.

The part first just generates the empty Lexis-diagram:

```
> library( Epi )
> par( mgp=c(3,1,0)/1.5, mar=c(3,3,1,1) )
> Lexis.diagram(date=c(1950,2000),age=c(0,50))
> box()
```

The next pieces of code draws the lines showing the tabulation intervals in the dataset. Note that we need to do this repeatedly, so we stick the code into a function `draw.grid`, that we can call later:

```
> par( mgp=c(3,1,0)/1.5, mar=c(3,3,1,1) )
> Lexis.diagram(date=c(1950,2000),age=c(0,50))
> draw.grid <-
+ function()
+ {
+ alim <- c(0,15,20,25,30)
+ blim <- 1900 + c(50,58,63,68,73)
+ plim <- c( 1968.25, 1993 )
+ abline( h=alim, v=plim, col="blue", lwd=2 )
+ for (i in 1:5) abline( -blim[i], 1, col="blue", lwd=2 )
+ segments( pmax( rep( plim[1], 5), 1950+alim ), alim,
+           rep( plim[2], 5), alim, lwd=4 )
+ segments( rep( plim[1], 5 ), plim[1] - blim,
+           rep( plim[2], 5 ), plim[2] - blim, lwd=4 )
+ segments( plim, 0, plim, plim - blim[1], lwd=4 )
+ box()
+ }
> draw.grid()
```

The two next ones also puts in the number of cases and PYs in each of the areas. First we read the data, compute the no cases and PY and the midpoints in terms of son's cohort and son's age, which allows us to derive the (date,age) coordinates to be used in the diagram.

```
> # tc <- read.table( "../data/tc-testis.txt", header=T )
> tw <- read.table( "http://www.biostat.ku.dk/~pka/epidata/testis.txt",
+                   header=TRUE, na.strings="." )
> names( tw ) <- tolower( names(tw) )
> names( tw ) <- gsub( "_", ".", names(tw) )
> attach( tw )
> D <- tapply( cases, list( son.age, son.koh ), sum )
> Y <- tapply( pyrs, list( son.age, son.koh ), sum )
```

```
> x <- outer( unique( son.age )+c(17.5,rep(2.5,4)),
+             unique( son.koh )+2.5,
+             "+" )
> x[1,] <- x[1,]-c(0,3,3,3,2)
> y <- outer( c(13,17,22,27,32), rep(1,5), "*" )
```

Then we can repeat the code that generated the diagram with the lines and add to taht the no. of cases:

```
> Lexis.diagram( date=c(1950,2000),age=c(0,50) )
> draw.grid()
> text( x[!is.na(D)], y[!is.na(D)], paste( D[!is.na(D)] ),
+       col="black", adj=1, font=2 )
> rect( 1950, 45, 1960, 50, col="white", border=gray(0.7) )
> text( 1952.5, 47.5, "Cases", adj=0 )
> box()
```

— and the same adding the person-years:

```
> Lexis.diagram(date=c(1950,2000),age=c(0,50))
> draw.grid()
> text( x[!is.na(Y)], y[!is.na(Y)], paste( round( Y/10^4 )[!is.na(Y)] ),
+       col="black", adj=1, font=2 )
> rect( 1950, 45, 1965, 50, col="white", border=gray(0.7) )
> text( 1952.5, 47.5, "PY (10,000)", adj=0 )
```

5.6.3 Practical exercises

First we read in the data from the website

```
> library( Epi )
> tc <- read.table( "http://www.biostat.ku.dk/~pka/epidata/testis.txt",
+                   header=TRUE, na.strings="." )
> names( tc ) <- tolower( names(tc) )
> str( tc )
```

```
'data.frame':        237 obs. of  8 variables:
 $ son_age : int  0 0 0 0 0 0 0 0 0 0 ...
 $ son_koh : int  1950 1950 1950 1950 1950 1950 1950 1958 1958 1958 ...
 $ moth_age: int  12 12 12 20 20 20 20 12 12 12 ...
 $ parity  : int  1 2 3 1 2 3 4 1 2 3 ...
 $ pyrs    : num  25096.8 1859 64.2 21779.2 4972.1 ...
 $ nonsemi : int  0 0 0 0 0 0 0 0 0 0 ...
 $ semi    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ cases   : int  0 0 0 0 0 0 0 0 0 0 ...
```

We then make variables into factors:

```
> tc <- transform( tc, son.koh = Relevel(factor(son_koh),list(1,2,3,c(4:5))),
+                      son.age = factor( son_age ),
+                     moth.age = factor( moth_age ),
+                       parity = factor( parity ),
+                           p2 = Relevel(factor(parity),list(1,2:4)) )
```

10. Compute the crude rate ratio for testis cancer for parity 2+ versus parity 1.

    We fit a Poisson model with `cases` as outcome, log-person-years as offset and parity 2+ as only predictor:

    ```
    > m1 <- glm( cases ~ p2, offset=log(pyrs), family=poisson, data=tc )
    > round( ci.exp(m1), 2 )
    ```

```
            exp(Est.) 2.5% 97.5%
(Intercept)      0.00 0.00  0.00
p22+3+4          0.57 0.48  0.67
```

The crude estimate is smaller than the adjusted because there is a subsantial age-confounding.

11. Reconstruct the estimates for "parity of mother at birth of son" from the top of Table 1 in the article both for all testis cancers and for non-seminomas.

    We fit a model of the same structre, just with adjustement for age, birth cohort and maternal age:

```
> m2 <- glm( cases ~ parity + son.age + son.koh + moth.age,
+            offset=log(pyrs), family=poisson, data=tc )
> m2g <- update( m2, .~.-parity+p2 )
```

    Note we can fit the same models with a different outcome by just using the `update` function:

```
> s2 <- update( m2, nonsemi ~ . )
> s2g <- update( m2g, nonsemi ~ . )
```

    And finally we can show the estiamtes in the same layout as in the article:

```
> round( rbind( cbind( ci.exp(m2,subset="par"),
+                      ci.exp(s2,subset="par") ),
+               cbind( ci.exp(m2g,subset="p2"),
+                      ci.exp(s2g,subset="p2") ) ), 2 )
```

```
        exp(Est.) 2.5% 97.5% exp(Est.) 2.5% 97.5%
parity2      0.78 0.65  0.95      0.78 0.63  0.98
parity3      0.87 0.64  1.19      0.83 0.58  1.20
parity4      0.68 0.37  1.27      0.73 0.37  1.46
p22+3+4      0.80 0.67  0.95      0.79 0.64  0.98
```

    Thus we have recreated the same estimates as in the article.

12. Reconstruct the estimates from Table 2 in the article concerning mother's age (for all testis cancers). Is there an interaction between parity and mother's age?

    First we fit the model, and note that the parameters of interest are those whre there is a ":" in the name:

```
> m3 <- update( m1, . ~ moth.age + moth.age:p2 + son.age + son.koh )
> round( ci.exp(m3,subset=":"), 3 )
```

```
                    exp(Est.)  2.5% 97.5%
moth.age12:p22+3+4      0.770 0.415 1.430
moth.age20:p22+3+4      0.793 0.634 0.992
moth.age25:p22+3+4      0.841 0.585 1.207
moth.age30:p22+3+4      0.617 0.246 1.545
```

    We can test the interaction by comparing with the model without interaction:

```
> anova( m3, update( m3, .~.- moth.age:p2 + p2 ), test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: cases ~ moth.age + son.age + son.koh + moth.age:p2
Model 2: cases ~ moth.age + son.age + son.koh + p2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       222     190.15
2       225     190.53 -3 -0.38146    0.944
```

we see there is no sign whatsoever of interaction, as also suggetsed by the very simile RRs between the age-classes of the mother.

13. Same question for birth cohort of the son.

    We can literally use the same code with small changes:

```
> m4 <- update( m1, . ~ moth.age + son.age + son.koh + son.koh:p2 )
> round( ci.exp(m4,subset=":"), 3 )

                          exp(Est.)  2.5% 97.5%
son.koh1950:p22+3+4           0.321 0.078 1.322
son.koh1958:p22+3+4           0.797 0.588 1.081
son.koh1963:p22+3+4           0.939 0.722 1.221
son.koh1968+1973:p22+3+4      0.628 0.435 0.906


> anova( m4, update( m4, .~.- son.koh:p2 + p2 ), test="Chisq" )

Analysis of Deviance Table

Model 1: cases ~ moth.age + son.age + son.koh + son.koh:p2
Model 2: cases ~ moth.age + son.age + son.koh + p2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       222     185.07
2       225     190.53 -3  -5.4623   0.1409
```

But we note that we do not get the same numbers as in the table, because this analysis is done using a version of `son.koh`, where the first two levels are collapsed — but only in the interaction, which causes R to lose track of the linear dependencies, so get the the wrong RRs for parity:

```
> tc$son.coh <- Relevel(tc$son.koh,list(1:2,3,4))
> m4 <- update( m1, . ~ moth.age + son.age + son.koh + son.coh:p2 )
> round( ci.exp(m4,subset=":"), 2 )

                          exp(Est.) 2.5% 97.5%
son.coh1950+1958:p21           1.33 0.99  1.78
son.coh1963:p21                1.07 0.82  1.38
son.coh1968+1973:p21           1.59 1.10  2.30
son.coh1950+1958:p22+3+4       1.00 1.00  1.00
son.coh1963:p22+3+4            1.00 1.00  1.00
son.coh1968+1973:p22+3+4       1.00 1.00  1.00


> round( 1/ci.exp(m4,subset=":"), 2 )

                          exp(Est.) 2.5% 97.5%
son.coh1950+1958:p21           0.75 1.01  0.56
son.coh1963:p21                0.94 1.22  0.72
son.coh1968+1973:p21           0.63 0.91  0.43
son.coh1950+1958:p22+3+4       1.00 1.00  1.00
son.coh1963:p22+3+4            1.00 1.00  1.00
son.coh1968+1973:p22+3+4       1.00 1.00  1.00
```

```
> anova( m4, update( m4, .~.-son.coh:p2 + p2 ), test="Chisq" )


Analysis of Deviance Table

Model 1: cases ~ moth.age + son.age + son.koh + son.coh:p2
Model 2: cases ~ moth.age + son.age + son.koh + p2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       223     187.04
2       225     190.53 -2  -3.4899   0.1747
```

We see we get the same estimates, and also that there is no interaction here either.