Epidemiological Methods in Medical Research

Computer practicals

Department of Biostatistics
Institute of Public Health, University of Copenhagen
Spring 2019
http://BendixCarstensen.com/EpiPhD/F2019
Version 1.2

Compiled Friday 11th January, 2019, 08:31
from: /home/bendix/teach/Epi/KU-epi/pracs/pracs.tex

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

Elisabeth Wreford Andersen Statistics and pharmacoepidemiology
Danish Cancer Society Research Center
elian@cancer.dk

Per Kragh Andersen Department of Biostatistics,
Institute of Public Health, University of Copenhagen
http://publicifsv.sund.ku.dk/~pka/
pkan@sund.ku.dk

Contents

Chapter 1

Introduction

1.1 Preface

This is the collection of exercises for the course in epidemiology for PhD-students in the spring of 2019.

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. This is reflected in the solutions sections which are only for SAS, Stata and R.

1.2 Website

The course website for the statistics practicals is
http://BendixCarstensen.com/EpiPhD/F2019. Note that this web address is sensitive to upper- and lower-case letters.

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/EpiPhD/F2019.

Descriptions of the datasets are in the exercise texts.

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 SAS program `bissau.sas` reads the data from `www` — 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. Then do the following by using the formulae from the lectures:

1. What is the overall risk of death? Make a confidence interval 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 death (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.

Hand calculations

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:

4. Estimate the 6-month *risk* of death for children with or without BCG vaccination (SAS users may use the program `bissau.sas`).

5. Compute 95% confidence limits for the two risk parameters.

6. 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.

7. Estimate the *rate (per day)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.

8. Estimate the *rate (per year)* of death for children with or without BCG vaccination. Compute also 95% confidence limits for the rate parameters.

9. 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 (call it `dtpany`, for example).

Calculations using a statistical model

10. 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 afterward.

11. 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 afterward.

12. 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 afterward.

13. 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.

14. 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?

15. Do the same for DTP (any dose vs. none, i.e. as a binary exposure). What do you conclude?

16. Test the association between BCG and DTP-any dose using a Chi-square test. In this mortality is not involved, only test whether the occurrence of the two types of vaccination are related. How would you describe the relationship? What do you conclude?

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

18. 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.

19. Revisit the previous analyses of this dataset, but now using death (`dead`) as outcome, and estimate the DTP effect for each level of BCG.

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

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

22. Estimate the effects of DTP vaccination (yes vs. no) on the 6-month odds of death with and without adjustment for age at first visit (in months, `agem`).

    Does age seem to be a confounder for the effect of DTP?

23. Adjust further for BCG vaccination. How does this adjustment affect the DTP estimate?

24. Examine if there is an interaction between the effects of BCG and DTP.

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

25. 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 the lectures.

26. 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.

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

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

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

30. 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.

What do you conclude?

2.2 Case-control study of renal cancer and trichorehtene

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

The paper is available at the course homepage as
`http://BendixCarstensen.com/EpiPhD/Papers/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 your statistical software (in SAS you could use `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 available 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 Poisson 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 energy 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`	Not used
`energy`	Daily energy intake in MJ
`height`	Height in cm
`weight`	Weight in cm
`fat`	Daily fat intake (g)
`fibre`	Daily fibre intake (g)

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

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

5. 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?

6. Does this change if the effect of age is modeled with a linear spline?

7. Is there any evidence of a non-linear effect of energy, when using linear splines with knots at say 2, 2.5 and 3? (these numbers are approximately the quartiles in the energy-distribution).

8. How does the non-linear relationship look? Plot the estimated curve together with the estimated linear relationship.

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

10. The Poisson model(s) ou just fitted implicitly assumes that the rates of CHD are constant over time. Try to releax this assumption by fitting the corresponding Cox-models with time since study entry as time scale. Does the regresssion parameters change in any of the models?

11. Try to use current age as underlying time scale in the Cox-models instead.
    Hint: You must compute age at entry and age at exit as new variables and use these as input for the Cox-model.

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 handled somewhat differently in SAS, Stata and R, so the exercise is here given in three different versions, one for each programming language.

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).

   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 created age-variable should 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`).

[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.

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.

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 fibre  in 2/L /*
*/ using "http://BendixCarstensen.com/EpiPhD/F2014/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 now has 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` is 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)
   ```

Using R

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

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

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

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

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

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, `period`) and age. `exit` is also a named list, with one element with the name of one of the timescales, giving the values of the exit times on this time scale. `exit.status` gives the state that persons are in at exit from the study. If `entry.status` is not specified, it is assumed that everyone starts in the *first* state, and this is noted:

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

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

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

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

[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 <- 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 grid-lines corresponding to the breaks gets inserted:

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

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:

   ```
   > 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/EpiPhD/F2014/data/ihd-xtab.txt", heade
   > 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.

2.4.1 Simple 2×2 table

But first we want to analyse the 2 × 2 table from the lectures. Analyzing the table in SAS (or any other program for that matter) requires that we have one observation for each *cell* in the table, so in this case we need 4 observations. And for each we need 3 variables: a numeric `n` — the number in the table, and two classification variables indicating what vaccination status (`bcg`), resp. disease status (`lep`) the number refers to. In SAS small-ish datasets can be entered directly by including data between "`datalines ;`" and a "`;`" (the "`$`"s indicate that the preceding variables are character variables):

```
data a ;
  input n bcg $ lep $ ;
datalines ;
101  y  y
554  y  n
159  n  y
446  n  n
;
run ;
```

We can derive the odds-ratio (and the two different risk ratios) using: `proc freq`:

```
proc freq  data = a ;
  weight n ;
  tables bcg * lep / relrisk ;
run ;
```

This will produce a table classified by `bcg` and `lep`, where each combination has `n` observations in it. If you omit the `weight` statement you will just get a table with 4 entries of 1.

4. Key in the data as described and find out what the OR for leprosy associated with presence vs. absence of BCG scar is.

5. Repeat the analysis using the total study as presented in the lectures (you can find the numbers on the slide handouts).

   What is the effect of using the entire study in terms of the OR and the confidence interval of it?

2.4.2 MH-analysis

6. Estimate the marginal odds ratio for vaccination (i.e., without adjusting for age) together with a 95% confidence interval using *all* 80622 controls (`status=conall`). Compare with the results on BC's slides from 30 January.

7. Use the Mantel-Haenszel method to adjust for `age` to see how much the estimated odds ratio for vaccination changes.

8. Repeat the previous question, now using the random sample of 1000 controls (`status=con1000`). How is the confidence interval affected by having fewer controls?

9. Repeat the previous question, now using the age-matched sample of controls (`status=conmatch`). How is the confidence interval affected?

10. Do the analysis of the age-matched sample of controls, but now ignoring the matching. How is the estimate and the confidence interval affected?

2.4.3 Model based analysis

11. 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?

12. Estimate odds ratios and confidence intervals with non-exposed and youngest, respectively, as reference groups (in `SAS`: use '`ESTIMATE`' statements).

13. Estimate instead odds ratios and confidence intervals with the age group 20-24 as reference.

14. Test the hypothesis of no interaction between `age` and `scar`.

15. 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.

16. 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.

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

2.5 Food poisoning in Fyn

An outbreak of Salmonella *Manhattan* poisoning was observed in Fyn county, and a matched case-control study was conducted, each case being matched to a control of the same sex, age and municipality of residence. Thus the matching is on a non-quantifiable variable.

Participants were asked (by telephone) what types of food they had eaten during the last week.

The data on these replies are given for cases and matched controls in the dataset `manh.txt` in the usual data folder.

The variables in the data set are:

The food item of primary interest we shall look at will be `hamburg`, hamburgerryg, boiled and smoked pork.

- Tabulate the number of cases, resp. controls exposed and non-exposed to `hamburg`.

- Estimate the odds-ratio of infection associated with the exposure to `hamburg`.

  What is the conclusion of this analysis

- Are there other items that are of importance for the occurrence of S. *Manhattan* infection?

Table 2.1: *Variables in the dataset* `manh`. *All exposure variables are 0/1 variables indicating if a persone har replied yes to eating this type of food.*

`id`	person
`set`	matched set
`caco`	case (1) control (0)
`sex`	sex
`okskod`	beef
`svinkod`	pork
`kalvkod`	veal
`lamkod`	lamb
`fjerkod`	poultry
`kodpaal`	sliced meats
`eggret`	egg dishes
`hamburg`	hamburgerryg
`filet`	fillet
`rgtmbr`	smoked tenderloin

2.6 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.

2.6.1 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?

2.6.2 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`.

Table 2.2: *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 $\geq$ 5mm
`nvtot`	—	total number of naevi
`burn15`	—	number of sunburns before age 15

2.6.3 Simple tabulation analysis

9. 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.

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

11. 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.

12. Same question for age.

Table 2.3: *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**

2.6.4 Simple analysis controlling for age

13. 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.

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

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

Table 2.4: *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	$p < 0.001$	

2.6.5 Introductory analyses.

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

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

18. 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.

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

20. 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`.

2.6.6 Trend tests and interactions.

21. 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`.

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

23. 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?

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

25. 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!

2.7 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].

2.7.1 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/EpiPhD/F2014/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.

2.7.2 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.