Statistical Analysis in the Lexis Diagram: Age-Period-Cohort models

Max Plack Insitute of Demographic Research, Rostock
2-6 May 2016
http://BendixCarstensen.com/APC/MPIDR-2016
1.0

Compiled Friday 29th April, 2016, 21:28
from: /home/bendix/teach/APC/MPIDR.2016/pracs/pracs.tex

Bendix Carstensen Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
www.BendixCarstensen.com

Contents

Chapter 1

Program and introduction

Course program

As the general rule, the daily program will have one lecture and one practical each morning and each afternoon.

Lectures will be between 45 and 90 minutes; normally with one or two breaks. Occasionally you will be asked to do small practical in the middle of the lectures.

The practicals will follow the lecture to fill the 3-hour slot. Sometimes we may need to push over some of the practical computing to take a bit of the beginning of the next slot.

The general rule is that there will be a walk-through of practicals after you have had a change to have a go at it yourself.

Monday 2nd

09:00 – 09:15	Welcome and introduction.
09:15 – 12:15	Morning slot:
– L1:	Follow-up time and rates from register data `surv-rate`
–	Lexis machinery in `Epi lifetable`
–	Follow-up time and rates from population data `tab-mod`
– P1:	Regression, linear algebra and reparametrization
–	Danish prime ministers `pm`
13:15 – 16:15	Afternoon slot:
– L2:	Likelihood for rates: Cox and Poisson
–	Cox as limit of the Poisson `WntCma`
–	Poisson model for rates: Factor models
–	Practical handling of linear contrasts in `R` using `ci.lin() tab-mod`
– P2:	Rates and survival, RR and RD
–	Linear and curved effects

Tuesday 3rd

09:00 – 09:30	**Recap of Monday**
09:30 – 12:15	Morning slot:
– L3:	The age-period and the age-cohort model. `AP-AC`
–	The Age-drift model
– P3:	Age-period model `age-per`
–	Age-cohort model `age-coh`
–	Age-drift model `age-drift`
13:15 – 16:15	Afternoon slot:
– L4:	The Age-period-cohort model
–	Parametrizations
–	Lexis triangles
– P4:	Age-period-cohort model
–	Using `apc.fit`

Wednesday 4th

09:00 – 09:30	**Recap of Tuesday**
09:30 – 12:15	Morning slot:
– L5:	Parametrization revisited: The general case.
–	The Lee-Carter model
– P5:	Age-period-cohort model for triangles
– L5:	The implementation of `apc.fit`.
–	Parametrizations.
–	The residual parametrization.
– P5:	Lee-Carter: Lung cancer in Danish women
13:15 – 16:15	Afternoon slot:
– L6:	Several rates compared with APC-models:
–	Estimation and reporting of effects.
–	Parametrization options for several rates.
– P6:	Lung cancer differences by sex `lung-sex`

Thursday 5th

09:00 – 16:00	Study free: Working with the assignments.

Friday 6th

09:00 – 09:30		**Recap of Wednesday**
09:30 – 12:15		Morning slot:
–	L7:	Predictions based on APC models
–		Managing splines for prediction
–	P7:	Predicting lung cancer `lung-pred`
–		Predicting breast cancer `breast-pred`
13:15 – 16:00		Afternoon slot:
–	L8:	APC-models for continuous outcome
–	P8:	BMI in Australia
16:00 – 16:15		Wrapping up, closure, evaluation and farewell

1.1 Reading

It would be helpful if you had read the papers which cover the essentials of the models that we will cover: [1, 2, 3, 4]

1.2 Introduction to exercises

Most of the following exercises all require basic skills in computing, in R, in particular the use of the graphical facilities.

1.2.1 Datasets and how to access them.

All the datasets for the exercises in this section are in the folder `APC\data`. This can be accessed through the homepage of the course, as:
`http://BendixCarstensen.com/APC/data`.

The datasets with `.txt` extension are plain text files where variable names are found in the first line. Such datasets can be read into R with the command `read.table`

1.2.2 R-functions and packages

Most functions for this course (and several more) are supplied in the R-package `Epi`, which can be downloaded from CRAN (on the R-website). It is also recommended that you get the packages `demography` and `ilc`.

```
> library( Epi )
> sessionInfo()
```

The latter command will list the attached packages and their version numbers. Yur version of Epi should be at least 2.3.

1.2.3 Solutions

This document also contains some suggestions for solutions of the assignments. They should *not* be taken as the *only* possible solutions to the practicals.

It is a good idea to give it a shot to do the practicals before you look in the solutions. However, the odd solution proposal may contain a twist to the analyses that you may find useful. Any suggestions for improving the solutions would be most welcome.

The R-code used in the solutions is available in the folder `http://bendixcarstensen.com/APC/MPIDR-2016/R/`, the filenames are shown at the top of each of the solution sections.

Chapter 2

Basic concepts in analysis of rates and survival

The following is a summary of relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

2.1 Probability

Survival function:

$$\begin{aligned} S(t) &= \mathrm{P}\left\{\text{survival at least till } t\right\} \\ &= \mathrm{P}\left\{T > t\right\} = 1 - \mathrm{P}\left\{T \leq t\right\} = 1 - F(t) \end{aligned}$$

Conditional survival function:

$$\begin{aligned} S(t|t_{\text{entry}}) &= \mathrm{P}\left\{\text{survival at least till } t | \text{ alive at } t_{\text{entry}}\right\} \\ &= S(t)/S(t_{\text{entry}}) \end{aligned}$$

Cumulative distribution function of death times (cumulative risk):

$$\begin{aligned} F(t) &= \mathrm{P}\left\{\text{death before } t\right\} \\ &= \mathrm{P}\left\{T \leq t\right\} = 1 - S(t) \end{aligned}$$

Density function of death times:

$$f(t) = \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{death in } (t, t+h)\right\}/h = \lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{h} = F'(t)$$

Intensity:

$$\begin{aligned} \lambda(t) &= \lim_{h \rightarrow 0} \mathrm{P}\left\{\text{event in } (t, t+h] \mid \text{alive at } t\right\}/h \\ &= \lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\ &= \lim_{h \rightarrow 0} -\frac{S(t+h) - S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} \end{aligned}$$

The intensity is also known as the hazard function, hazard rate, mortality/morbidity rate or simply “rate”.

Note that f and λ are *scaled* quantities, they have dimension time^{-1}.

Relationships between terms:

$$\begin{aligned} -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\ &\Updownarrow \\ S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\big(-\Lambda(t)\big) \end{aligned}$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity (rate), it is dimensionless, despite its name.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

The cumulative *risk* of an event (to time t) is:

$$F(t) = \mathrm{P}\left\{\text{Event before time } t\right\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

2.2 Statistics

Likelihood contribution from follow up of one person:

The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$\begin{aligned} \mathrm{P}\left\{\text{event at } t_4|\text{entry at } t_0\right\} &= \mathrm{P}\left\{\text{survive } (t_0,t_1)|\text{ alive at } t_0\right\} \times \\ &\quad\ \mathrm{P}\left\{\text{survive } (t_1,t_2)|\text{ alive at } t_1\right\} \times \\ &\quad\ \mathrm{P}\left\{\text{survive } (t_2,t_3)|\text{ alive at } t_2\right\} \times \\ &\quad\ \mathrm{P}\left\{\text{event at } t_4|\text{ alive at } t_3\right\} \end{aligned}$$

Each term in this expression corresponds to one *empirical rate*[1] $(d, y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length y. Each person can contribute many empirical rates, most with $d = 0$; d can only be 1 for the *last* empirical rate for a person.

Log-likelihood for one empirical rate (d, y):

$$\ell(\lambda) = d\log(\lambda) - \lambda y$$

This is under the assumption that the rate (λ) is constant over the interval that the empirical rate refers to.

[1] This is a concept coined by BxC, and so is not necessarily generally recognized.

Log-likelihood for several persons. Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D\log(\lambda) - \lambda Y,$$

where Y is the total follow-up time, and D is the total number of failures.

Note: The Poisson log-likelihood for an observation D with mean λY is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

The term $D\log(Y)$ does not involve the parameter λ, so the likelihood for an observed rate can be maximized by pretending that the no. of cases D is Poisson with mean λY. But this does *not* imply that D follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

A linear model for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\bigl(\log(\lambda) + \log(Y)\bigr) = \exp\bigl(X\beta + \log(Y)\bigr)$$

Therefore, in order to get a linear model for $\log(\lambda)$ we must require that $\log(Y)$ appear as a variable in the model for $D \ \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called *offset*-term in the linear predictor.

2.3 Competing risks

Competing risks: If there is more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

$$\lambda_c(a) = \lim_{h\rightarrow 0} \mathrm{P}\left\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\right\}/h, \quad c = 1,2,3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

because you have to escape all 3 causes of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

$$\mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the "cause-specific survival", but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed and that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1, 2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

Subdistribution hazard Fine and Gray defined models for the so-called subdistribution hazard. Recall the relationship between between the hazard (λ) and the cumulative risk (F):

$$\lambda(a) = -\frac{\mathrm{d}\log\bigl(S(a)\bigr)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\bigl(1-F(a)\bigr)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\bigl(1-F_1(a)\bigr)}{\mathrm{d}a}$$

This is what is called the subdistribution hazard, it depends on the survival function S, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risk. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard.

2.4 Demography

Expected residual lifetime: The expected lifetime (at birth) is simply the variable age (a) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty af(a)\,\mathrm{d}a$$

where f is the density of the distribution of lifetime (age at death).

The relation between the density f and the survival function S is $f(a) = -S'(a)$, so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\bigl(-S'(a)\bigr)\,\mathrm{d}a = -\Bigl[aS(a)\Bigr]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and a by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected *residual* lifetime at age a is calculated as the integral of the *conditional* survival function for a person aged a:

$$\mathrm{EL}(a) = \int_a^\infty S(u)/S(a)\,\mathrm{d}u$$

Lifetime lost due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(u)/S_{\mathrm{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a decision required in order to devise an estimate of S_{Well}.

Lifetime lost by cause of death is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned} S(a) = 1 &- \mathrm{P}\left\{\text{dead from cause 1 at } a\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a\right\} \end{aligned}$$

and hence:

$$\begin{aligned} S_{\text{Well}}(a) - S_{\text{Diseased}}(a) = &\ \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Diseased}\right\} \\ &+ \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Diseased}\right\} \\ &+ \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Diseased}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Well}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Well}\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Well}\right\} \end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned} \mathrm{LL}_2(a) = \int_a^\infty &\mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\right\} \\ &- \mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\right\} \,\mathrm{d}u \end{aligned}$$

These quantities have the property that their sum is the total years of life lost due to the disease:

$$\mathrm{LL}(a) = \mathrm{LL}_1(a) + \mathrm{LL}_2(a) + \mathrm{LL}_3(a)$$

The terms in the integral are computed as (see the section on competing risks):

$$\begin{aligned} \mathrm{P}\left\{\text{dead from cause 2 at } x|\text{Diseased \& alive at } a\right\} &= \int_a^x \lambda_{2,\text{Dis}}(u) S_{\text{Dis}}(u)/S_{\text{Dis}}(a)\,\mathrm{d}u \\ \mathrm{P}\left\{\text{dead from cause 2 at } x|\text{Well \& alive at } a\right\} &= \int_a^x \lambda_{2,\text{Well}}(u) S_{\text{Well}}(u)/S_{\text{Well}}(a)\,\mathrm{d}u \end{aligned}$$

Chapter 3

Practical exercises

3.1 Regression, linear algebra and projection

This exercise is aimed at reminding you about the linear algebra behind linear models. Therefor we use artificial data

1. First generate a continuous variable `x`, and a factor `f` on 3 levels, each with 100 units, say:

   ```
   x <- runif(100,20,50)
   f <- factor( sample(letters[1:3],100,replace=T) )
   x
   table( f )
   ```

   Then generate a response variable `y` by some function (the exact shape is immaterial):

   ```
   y <- 0.2*x + 0.02*(x-25)^2 + 3*as.integer(f) + rnorm(100,0,1)
   plot( x, y, col=f, pch=16 )
   ```

2. Now fit the same model using `lm`, so this should get your parameter estimates back (almost):

   ```
   mm <- lm( y ~ x + I(x^2) + f )
   summary( mm )
   ```

3. Now verify that you get the same results using the matrix formulae. You will first have to generate the design matrix:

   ```
   X <- cbind( 1, x, x^2, f=="b", f=="c" )
   ```

   Recall that the matrix formula for the estimates is:

   $$\hat{\beta} = (X'X)^{-1}X'y$$

   To make this calculation explicitly in R you will need the transpose `t()` and the matrix inversion `solve()` functions, as well as the matrix multiplication operator `%*%`.

   An explicit calculation then gives:

   ```
   bb <- solve( t(X) %*% X ) %*% t(X) %*% y
   cbind( bb, coef(mm) )
   ```

3.2 Reparametrization of models

This exercise is aimed at showing you how to reparametrize a model: Suppose you have a model parametrized by the linear predictor $X\beta$, but that you really wanted the parametrization $A\gamma$, where the columns of X and A span the same linear space.

So $X\beta = A\gamma$, and we assume that both X and A are of full rank, $\dim(X) = \dim(A) = n \times p$, say.

We want to find γ given that we know $X\beta$ and that $X\beta = A\gamma$. Since we have that $p < n$, we have that $A^-A = I$, by the properties of G-inverses, and hence:

$$\gamma = A^-A\gamma = A^-X\beta$$

1. try to generate a dataset with a response hat is normally distributed in three groups, and then fit the model using the "usual" parametrization:

```
f <- factor( sample(letters[1:3],20,replace=T) )
y <- 5+2*as.integer(f) + rnorm(20,0,1)
mm <- lm( y ~ f )
library( Epi )
ci.lin( mm )
```

2. Set up the model matrix `X` for this regression, and versify that you get the same results by entering `X` as regression in `lm`

```
( X <- cbind( 1, f=="b", f=="c" ) )
ci.lin( lm( y ~ X-1 ) )
```

3. Now suppose you want a parametrization with the last level as reference instead. You could then easily convert the parameters, but use the formulae from above to do it, by first setting up `A` corresponding to the desired parametrization, and then using `ginv` from the `MASS` library:

```
library( MASS )
( A <- cbind( 1, f=="a", f=="b" ) )
ginv(A) %*% X
ginv(A) %*% X %*% ci.lin( mm )[,1]
```

4. Verify that you get the results you expect:

```
( X <- cbind( 1, f=="b", f=="c" ) )
( A <- cbind( 1, f=="a", f=="b" ) )
ginv(A) %*% X
```

5. Try to obtain the conversion from the parametrization with an intercept and two contrasts to the parametrization with a separate level in each group by constructing the matrices using the `model.matrix` function.

```
( X <- model.matrix( ~f   ) )
( A <- model.matrix( ~f-1 ) )
ginv(A) %*% X
```

The essences of these calculations are:

- Given that you have a set of fitted values in a model (*in casu* $\hat{y} = X\beta$) and you want the parameter estimates you would get if you had used the model matrix A. Then they are $\gamma = A^{-}\hat{y} = A^{-}X\beta$.

- Given that you have a set of parameters β, from fitting a model with design matrix X, and you would like the parameters γ, you would have got had you used the model matrix A. Then they are $\gamma = A^{-}X\beta$.

3.3 Danish prime ministers

The following table shows all Danish prime ministers in office since the war. They are ordered by the period in office, hence some appear twice. Entry end exit refer to the office of prime minister. A missing date of death means that the person was alive at 31 March 2016.

Name	Birth	Death	Entry	Exit
Vilhelm Buhl	16/10/1881	18/12/1954	05/05/1945	07/11/1945
Knud Kristensen	26/10/1880	29/09/1962	07/11/1945	13/11/1947
Hans Hedtoft	21/04/1903	29/01/1955	13/11/1947	30/10/1950
Erik Eriksen	20/11/1902	07/10/1972	30/10/1950	30/09/1953
Hans Hedtoft	21/04/1903	29/01/1955	30/09/1953	29/01/1955
H C Hansen	08/11/1906	19/02/1960	01/02/1955	19/02/1960
Viggo Kampmann	21/07/1910	03/06/1976	21/02/1960	03/09/1962
Jens Otto Kragh	15/09/1914	22/06/1978	03/09/1962	02/02/1968
Hilmar Baunsgaard	26/02/1920	30/06/1989	02/02/1968	11/10/1971
Jens Otto Kragh	15/09/1914	22/06/1978	11/10/1971	05/10/1972
Anker Jorgensen	13/07/1922	20/03/2016	05/10/1972	19/12/1973
Poul Hartling	14/08/1914	30/04/2000	19/12/1973	13/02/1975
Anker Jorgensen	13/07/1922	20/03/2016	13/02/1975	10/09/1982
Poul Schlüter	03/04/1929	.	10/09/1982	25/01/1993
Poul Nyrup Rasmussen	15/06/1943	.	25/01/1993	27/11/2001
Anders Fogh Rasmussen	26/01/1953	.	27/11/2001	05/04/2007
Lars Løkke Rasmussen	15/05/1964	.	21/01/2009	03/10/2011
Helle Thorning-Schmidt	14/12/1966	.	03/10/2011	28/06/2015
Lars Løkke Rasmussen	15/05/1964	.	28/06/2015	.

The data in the table can be found in the file `pm-dk.txt`.

```
st <- read.table( "../data/pm-dk.txt", header=T, as.is=T, na.strings="." )
st
str( st )
```

1. Draw a Lexis diagram with life-lines of the persons, for example by using the `Lexis` machinery from the `Epi` package:

```
library( Epi )
# Change the character variables with dates to fractional calendar
# years
for( i in 2:5 ) st[,i] <- cal.yr( st[,i], format="%d/%m/%Y" )
# Attach the data for those still alive
st$fail <- !is.na(st$death)
st[is.na(st$exit),"exit"] <- cal.yr( Sys.Date() )
st[     !st$fail,"death"] <- cal.yr( Sys.Date() )
st
# Lexis object
L <- Lexis( entry = list(per=birth),
             exit = list(per=death, age=death-birth),
             exit.status=fail,
             data=st )
# Plot Lexis diagram
par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, xaxt="n" ) # Omit x-labels
plot( L, xlim=c(1945,2020), ylim=c(20,95),
         xaxs="i", yaxs="i", lwd=3, las=1,
         grid=0:20*5, col="black", xlab = "Calendar time", ylab="Age" )
points( L, pch=c(NA,16)[L$lex.Xst+1] )
# Put names of the prime ministers on the plot
with( st, text( death, death-birth, Name, adj=c(1.05,-0.05), cex=0.7 ) )
par( xaxt="s" )
axis( side=1, at=seq(1950,2010,10) ) # x-labels at nice places
```

2. Mark with a different color the periods where they have been in office. You could try something like:

```
# New Lexis object describing periods in an office
# and lines added to a picture
st <- transform( st,
                 in_office = c( rep(FALSE,nrow(st)-1),TRUE ),
                      exit = ifelse( is.na(exit), 2011, exit ) )
Lo <- Lexis( entry = list(per=entry),
              exit = list(per=exit, age=exit-birth),
              exit.status=in_office,
              data = st )
lines( Lo, lwd=3, las=1, col="red" )
# the same may be plotted using command segments
box()
with( st, segments( birth, 0, death, death-birth, lwd=2 ) )
with( st, segments( entry, entry-birth, exit, exit-birth, lwd=4, col="red" ) )
```

3. Draw the line representing age 50 years.

4. How many 50th birthdays have been celebrated in office since the war?

5. Draw the line representing 2 October 1972. (Why just that?)

```
abline( v=cal.yr( "2/10/1972", format="%d/%m/%Y" ) )
```

6. How many present and former prime ministes were alive at 31st December 2008?

7. Which period(s) since the war has seen the maximal number of former post-war prime ministers alive?

```
# New lexis object - since entry to the office to the death
Ln <- Lexis( entry = list(per = entry),
              exit = list(per = death,
                           age = death-entry ),
       exit.status = death,
              data = st )
ny <- 2008-1945
n_alive <- vector( "numeric", ny )
for (i in 1:ny)
{
alive <- ( (Ln$death >=(1944+i))&(Ln$entry<=(1944+i)) )
n_alive[i] <- nlevels( as.factor( subset( Ln$Name, alive==T ) ) )
}
plot( n_alive~seq(1945,(1945+ny-1),1), type="l", xlab="Calendar year",
      ylab = "Maximal numbers of former prime ministers alive" )
```

8. Mark the area in the diagram with person years lived by persons aged 50 to 70 in the period 1 January 1970 through 1 January 1990.

9. Mark the area for the lifetime experience of those who were between 10 and 20 years old in 1945.

```
polygon( c(1955,2010,2010,1965,1955), c(30,85,75,30,30), lwd=2,
     border="blue", col="lightblue" )
```

10. How many prime-minister-years have been spent time in each of these sets? And in the intersection of them?

```
# Prime-minister years lived by persons
#  aged 50 to 70 in the period 1 January 1970 through 1 January 1990.
x1 <- splitLexis( Lo ,breaks = c(0,50,70,100), time.scale="age" )
x2 <- splitLexis( x1, breaks = c(1900,1970,1990,2010), time.scale="per" )
summary( x2 )
tapply( status(x2,"exit")==1, list( timeBand(x2,"age","left"),
                                    timeBand(x2,"per","left") ), sum )
tapply( dur(x2),  list( timeBand(x2,"age","left"),
                        timeBand(x2,"per","left") ), sum )
# Computing the person-years in the 1925-35 cohort
x3 <- subset( Lo, birth>1925 & birth<=1935 )
summary( x3 )
dur( x3 )
# Computing person years in the intersection
x4 <- subset( x2 , birth>1925 & birth<=1935 )
summary( x4 )
dur( x4 )
```

3.4 Rates and survival

1. Consider the following data:

Year of birth	Year of death		Age at death
	1994	1995	
1994	2,900	500	0
1993	120	130	1
1992	50	60	2
1991	45	55	3
1990	40	40	4

2. Represent these data in a Lexis diagram. You could use the `Lexis.diagram` function from the `Epi` package and the print the no. of cases on the digram with `text`.
3. On the basis of these data, can you calculate the age-specific death rate for two-year-olds (${}_1m_2$) in 1994? If you can, do it. If you cannot, explain what additional information you would need.
4. On the basis of these data, can you calculate the probability of surviving from age 2 to age 3 (${}_1q_2$) in for the cohort born in 1992? If you can, do it. If you cannot, explain what additional information you would need.
5. Now consider the following data:
   - Live births during 1991: 142,000
   - Number of infants born in 1991 who did not survive until the end of 1991: 2,900
   - Number of infants born in 1991 who survived to the end of 1991, but did not reach their first birthday: 500
   - Live births during 1992: 138,000
   - Number of infants born in 1992 who did not survive until the end of 1992: 2,600
   - Number of infants born in 1992 who survived to the end of 1992, but did not reach their first birthday: 450
6. Represent the data on a Lexis diagram.
7. Calculate the infant mortality rate (IMR) for 1992 under the assumption that you were only able to observe events occurring in 1992, and that you did not know the birth dates of infants dying during that year.
8. Same as above, except that now you do know the birth dates of infants dying during 1992.
9. Assume all data are known: Calculate the IMR.
10. What is the IMR for the 1992 birth cohort?

3.5 Calculation of rates, RR and RD

This exercise is *very* prescriptive, so you should make an effort to really understand everything you type into R. Consult the relevant slides of the lecture on “Poisson regression for rates ...”

3.5.1 Hand calculations for a single rate

Let λ be the true hazard rate or theoretical incidence rate, its estimator being the empirical incidence rate $\widehat{\lambda} = D/Y$ = 'no. cases/person-years'. Recall that the standard error of the empirical rate is $\mathrm{SE}(\widehat{\lambda}) = \widehat{\lambda}/\sqrt{D}$.

The simplest approximate 95% confidence interval (CI) for λ is given by

$$\widehat{\lambda} \pm 1.96 \times \mathrm{SE}(\widehat{\lambda})$$

An alternative approach is based on logarithmic transformation of the empirical rate. The standard error of the log-rate $\widehat{\theta} = \log(\widehat{\lambda})$ is $\mathrm{SE}(\widehat{\theta}) = 1/\sqrt{D}$. Thus, a simple approximate 95% confidence interval for the log-hazard $\theta = \log(\lambda)$ is obtained from

$$\widehat{\theta} \pm 1.96/\sqrt{D} = \log(\widehat{\lambda}) \pm 1.96/\sqrt{D}$$

When taking the exponential from the above limits, we get another approximate confidence interval for the hazard λ itself:

$$\exp\{\log(\widehat{\lambda}) \pm 1.96/\sqrt{D}\} = \widehat{\lambda} \overset{\times}{\div} \mathrm{EF},$$

where $\mathrm{EF} = \exp\{1.96 \times \mathrm{SE}[\log(\widehat{\lambda})]\}$ is the *error factor* associated with the 95% interval. This approach provides a more accurate approximation with small numbers of cases. (However, both these methods fail when $D = 0$, in which case an *exact* method or one based on *profile-likelihood* is needed.)

1. Suppose you have 15 events during 5532 person-years. Let's use R as a simple desk calculator to derive the rate (in 1000 person-years) and the first version of an approximate confidence interval:

```
> library( Epi )
> options(digits=4)  #  to cut down decimal points in the output
```

```
> D <- 15
> Y <- 5.532    # thousands of years
> rate <- D / Y
> SE.rate <- rate/sqrt(D)
> c(rate, SE.rate, rate + c(-1.96, 1.96)*SE.rate )
```

2. Compute now the approximate confidence interval using the method based on log-transformation and compare the result with that in item (a)

```
> SE.logr <- 1/sqrt(D)
> EF <- exp( 1.96 * SE.logr )
> c(log(rate), SE.logr)
> c( rate, EF, rate/EF, rate*EF )
```

3.5.2 Poisson model for a single rate with logarithmic link

You are able to estimate λ and compute its CI with a Poisson model, as described in the relevant slides in the lecture handout.

3. Use the number of events as the response and the log-person-years as an *offset* term, and fit the Poisson model with log-link

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

What is the interpretation of the parameter in this model?

4. The summary method produces too much output. You can extract CIs for the model parameters directly from the fitted model on the scale determined by the *link* function with the `ci.lin()`-function. Thus, the estimate, SE, and confidence limits for the log-rate $\theta = \log(\lambda)$ are obtained by:

```
> ci.lin( m )
```

However, to get the confidence limits for the rate $\lambda = \exp(\theta)$ on the original scale, the results must be exp-transformed:

```
> ci.lin( m, Exp=T)
```

To get just the point estimate and CI for λ from log-transformed quantities you are recommended to use function `ci.exp()`, which is actually a wrapper of `ci.lin()`:

```
> ci.exp( m)
> ci.lin( m, Exp=T)[, 5:7]
```

Both functions are found from `Epi` package. – Note that the test statistic and P-value are rarely interesting quantities for a single rate.

5. There is an alternative way of fitting a Poisson model: Use the empirical rate $\widehat{\lambda} = D/Y$ as a *scaled* Poisson response, and the person-years as *weight* instead of offset (albeit it will give you a warning about non-integer response in a Poisson model, but you can ignore this warning):

```
> mw <- glm( D/Y ~ 1, family=poisson, weight=Y )
> ci.exp( mw)
```

Verify that this gave the same results as above.

3.5.3 Poisson model for a single rate with identity link

The advantage of the approach based on weighting is that it allows sensible use of the *identity* link. The response is the same but the parameter estimated is now the rate itself, not the log-rate.

6. Fit the Poisson model with identity link

```
> mi <- glm( D/Y ~ 1, family=poisson(link=identity), weight=Y )
> coef(mi)
```

What is the meaning of the intercept in this model?

Verify that you actually get the same rate estimate as before.

7. Now use ci.lin() to produce the estimate and the confidence intervals from this model:

```
> ci.lin( mi )
> ci.lin( mi )[, c(1,5,6)]
```

3.5.4 Poisson model assuming same rate for several periods

Now, suppose the events and person years are collected over three periods.

8. Read in the data and compute period-specific rates

```
> Dx <- c(3,7,5)
> Yx <- c(1.412,2.783,1.337)
> Px <- 1:3
> rates <- Dx/Yx
> rates
```

9. Fit the same model as before, assuming a single rate to the data for the separate periods. Compare the result from previous ones

```
> m3 <- glm( Dx ~ 1, family=poisson, offset=log(Yx)  )
> ci.exp(m3)
```

10. Now test whether the rates are the same in the three periods: Try to fit a model with the period as a factor in the model:

```
> mp <- glm( Dx ~ factor(Px), offset=log(Yx), family=poisson )
```

and compare the two models using anova() with the argument test="Chisq":

```
> anova( m3, mp, test="Chisq" )
```

Compare the test statistic to the deviance of the model mp.

What is the deviance good for?

3.5.5 Analysis of rate ratio

We now switch to comparison of two rates λ_1 and λ_0, i.e. the hazard in an exposed group vs. that in an unexposed one. Consider first estimation of the true rate ratio $\rho = \lambda_1/\lambda_0$ between the groups. Suppose we have pertinent empirical data (cases and person-times) from both groups, (D_1, Y_1) and (D_0, Y_0). The point estimate of ρ is the empirical rate ratio

$$\text{RR} = \frac{D_1/Y_1}{D_0/Y_0}.$$

It is known that the variance of log(RR), that is, the difference of the log of the empirical rates $\log(\widehat{\lambda}_1) - \log(\widehat{\lambda}_0)$ is estimated as

$$\begin{aligned}
\text{var}(\log(\text{RR})) &= \text{var}\{\log(\widehat{\lambda}_1/\widehat{\lambda}_0)\} \\
&= \text{var}\{\log(\widehat{\lambda}_1)\} + \text{var}\{\log(\widehat{\lambda}_0)\} \\
&= 1/D_1 + 1/D_0
\end{aligned}$$

Based on a similar argument as before, an approximate 95% CI for the true rate ratio λ_1/λ_0 is then:

$$\text{RR} \overset{\times}{\div} \exp\left(1.96\sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)$$

Suppose you have 15 events during 5532 person-years in an unexposed group and 28 events during 4783 person-years in an exposed group:

11. Calculate the the rate-ratio and CI by direct application of the above formulae:

```
> D0 <- 15   ; D1 <- 28
> Y0 <- 5.532 ; Y1 <- 4.783
> RR <- (D1/Y1)/(D0/Y0)
> SE.lrr <- sqrt(1/D0+1/D1)
> EF <- exp( 1.96 * SE.lrr)
> c( RR, RR/EF, RR*EF )
```

12. Now achieve this using a Poisson model:

```
> D <- c(D0,D1) ; Y <- c(Y0,Y1); expos <- 0:1
> mm <- glm( D ~ factor(expos), family=poisson, offset=log(Y) )
```

What do the parameters mean in this model?

13. You can extract the exponentiated parameters in two ways:

```
> ci.exp( mm)
> ci.lin( mm, E=T)[,5:7]
```

3.5.6 Analysis of rate difference

When estimating the true rate difference $\delta = \lambda_1 - \lambda_0$, the variance of the natural estimator $\mathrm{RD} = D_1/Y_1 - D_0/Y_0$ is (since the empirical rates are based on independent samples) just the sum of the variances:

$$\begin{aligned}\mathrm{var}(\mathrm{RD}) &= \mathrm{var}(\widehat{\lambda}_1) + \mathrm{var}(\widehat{\lambda}_0) \\ &= D_1/Y_1^2 + D_0/Y_0^2\end{aligned}$$

14. Use this formula to compute the rate difference and a 95% confidence interval for it:

```
> rd <- diff( D/Y )
> sed <- sqrt( sum( D/Y^2 ) )
> c( rd, rd+c(-1,1)*1.96*sed )
```

15.

16. Verify that this is the confidence interval you get when you fit an additive model with exposure as factor:

```
> ma <- glm( D/Y ~ factor(expos),
+            family=poisson(link=identity), weight=Y )
> ci.lin( ma )[, c(1,5,6)]
```

3.5.7 Calculations using matrix tools

NB. This subsection requires some familiarity with matrix algebra.

17. You can explore the function `ci.mat()`, which lets you use matrix multiplication (operator `'%*%'` in R) to produce confidence interval from an estimate and its standard error (or CIs from whole columns of estimates and SEs):

```
> ci.mat
> ci.mat()
```

Apply this to the single rate calculations in 1.6.1:

```
> c( rate, SE.rate ) %*% ci.mat()
> exp( c( log(rate), SE.logr ) %*% ci.mat() )
```

18. For computing the rate ratio and its CI as in 1.6.5, matrix multiplication with `ci.mat()` should give the same result as there:

```
> exp( c( log(RR), SE.lrr ) %*% ci.mat() )
```

19. Look again the model used to analyse the rate ratio in 1.6.5(b). Often one would like to get simultaneously both the rates and the ratio between them. This can be achieved in one go using the *contrast matrix* argument `ctr.mat` to `ci.lin()` or `ci.exp()`. Try:

```
> CM <- rbind( c(1,0), c(1,1), c(0,1) )
> rownames( CM ) <- c("rate 0","rate 1","RR 1 vs. 0")
> CM
> mm <- glm( D ~ factor(expos),
+               family=poisson(link=log),  offset=log(Y) )
> ci.exp( mm, ctr.mat=CM)
```

20. Use the same machinery to the additive model to get the rates and the rate-difference in one go. Note that the annotation of the resulting estimates are via the column-names of the contrast matrix.

```
> rownames( CM ) <- c("rate 0","rate 1","RD 1 vs. 0")
> ma <- glm( D/Y ~ factor(expos),
+                  family=poisson(link=identity), weight=Y )
> ci.lin( ma, ctr.mat=CM )[, c(1,5,6)]
```

3.6 Estimation and reporting of linear and curved effects

In this exercise we will use the `testisDK` data from the `Epi` package, which contains the number of cases of testis cancer in Denmark 1943–96:

1. First load the Danish testis cancer data, and inspect the dataset:

```
 library( Epi )
 sessionInfo()
 data( testisDK )
 str( testisDK )
 head( testisDK )
```

   Tabulate both events and person-years using `stat.table`, in say 10-year age-groups and 10-year periods of follow-up. In which ages are the age-specific testis cancer rates highest?

2. Now fit a Poisson-model for the mortality rates with a linear term for age at follow-up (current age, attained age):

```
 ml <- glm( D ~ A, offset=log(Y), family=poisson, data=testisDK )
 ci.exp( ml )
```

   What do the parameters mean?

3. Work out the the predicted log-mortality rates for ages 25 to 45, say, by doing a hand-calculation based on the coefficients:

```
 ( cf <- coef( ml ) )
```

4. However, we do not have the standard errors of these mortality rates, and hence neither the confidence intervals. This is implemented in `ci.exp`; if we provide the argument `ctr.mat=` as a matrix where each row corresponds to a prediction point and each column to a parameter from the model. Look at the help page for `ci.exp` and then try:

```
( CM <- cbind( 1, 25:45 ) )
round( ci.exp( ml, ctr.mat=CM )*10^5, 3 )
```

5. Use this machinery to derive and plot the mortality rates over the range from 15 to 65 years, say:

   ```
   C1 <- cbind( 1, 15:65 )
   matplot( 15:65, ci.exp( ml, ctr.mat=C1 )*10^5,
            log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY",
            type="l", lty=1, lwd=c(3,1,1), col="black" )
   ```

6. Now check if the mortality rates really are eksponentially increasing by age (that is linearly on the log-scale), by adding a quadratic term to the model. Note that you must use the expression `I(A^2)` in the modelleing in order to avoid that the "^" is interpreted as part of the model formula:

   ```
   mq <- glm( D ~ A + I(A^2), offset=log(Y), family=poisson, data=testisDK )
   ci.exp( mq, Exp=F )
   ```

   Then plot the estimated rates under the quadratic model.

   ```
   aa <- 15:65
   C2 <- cbind( 1, aa, aa^2 )
   matplot( aa, ci.exp( mq, ctr.mat=C2 )*10^5,
            log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY",
            type="l", lty=1, lwd=c(3,1,1), col="black" )
   ```

   Try to overlay the estimated rates from the model with linear efect of age — you will need the function `matlines`.

7. Repeat the same using a 3rd degree polynomial.

8. Instead of continuing with higher powers of age we could use fractions of powers, or we could use splines, piecevise polynomial curves, that fit nicely together at join points (knots). This is implemented in the `splines` package, in the function `ns`, which returns a matrix. There is a wrapper `Ns` in the `Epi`-package that automatically designate the smallest and largest knots a *boundary knots*, beyond which the resulting curve is linear:

   ```
   library( splines )
   ms <- glm( D ~ Ns(A,knots=seq(15,65,10)), offset=log(Y),
              family=poisson, data=testisDK )
   ```

   In order to extract the estimated effects, construct a contrast matrix that correspond to the parameters of the model:

   ```
   As <- Ns( aa, knots=seq(15,65,10) )
   matplot( aa, ci.exp( ms, ctr.mat=cbind(1,As) )*10^5,
            log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY",
            type="l", lty=1, lwd=c(3,1,1), col="black" )
   ```

9. Now add a linear term in calendar time `P` to the model, and make a prediction of the incidence rates in 1970, say:

```
msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P, offset=log(Y), family=poisson, data=testisDK )
matplot( aa, ci.exp( msp, ctr.mat=cbind(1,As,1970) )*10^5,
         log="y", xlab="Age", ylab="Testis cancer incidence rate per 100,000 PY",
         type="l", lty=1, lwd=c(3,1,1), col="black" )
```

Note that cbind automatically will expand the 1 and the 1970 to match the number of rows of As.

10. Extract the RR relative to 1970, by using the subset argument to ci.exp:

    ```
    ci.exp( msp, subset="P" )
    ```

    What is the annual relative increase in the testis cancer incidence rates? Show the RR of testis cancer by year relative to 1970 by multipling the log-RR for period with the distance form 1970, such as:

    ```
    yy  <- 1943:1996
    Cp1 <- cbind( yy - 1970 )
    matplot( yy, ci.exp( msp, ctr.mat=Cp1, subset="P" ),
             log="y", xlab="Date", ylab="RR of Testis cancer",
             type="l", lty=1, lwd=c(3,1,1), col="black" )
    abline( h=1 )
    ```

11. Try to add a quadratic term to the period effect, and plot the resulting RR relative to 1970.
    Hint: In order to extract the quadratic effects relative to 1970, you must form the matrix of linear and quadratic period, and a corresponding matrix where all rows are identical to the 1970 row:

    ```
    msp <- glm( D ~ Ns(A,knots=seq(15,65,10)) + P + I(P^2),
                    offset=log(Y), family=poisson, data=testisDK )
    Cq <- cbind( yy, yy^2 ) - cbind( rep(1970,length(yy)), 1970^2 )
    ```

    Use this matrix as arguent to ci.exp

12. Now investigate if there is any non-linearity in period beyond the quadratic, by fitting fit a spline for (P) as well, and comparing the models. Plot the resulting RR by year, relative to 1970 too. You must define a contrast matrix corresponding to the years where the prediction is made, as well as a matrix with the same number of rows, but with all rows identical to the one corresponding to the reference year. You must use the differenec of these two as the arument to ctr.mat in ci.exp.

13. Plot the estimated age-specific rates in 1970 from this model. Note that you need a reference matrix for the period with all rows identical to the 1970 row, but this time with the same number of rows as the *age*-prediciton points.

14. Collect these steps in a general outline, where you first define the knots, and the points of age and period prediction, and then fit the model and do the two plots.

15. Form a new variable in the data frame, B=P-A, the data of birth, and repeat the last analysis with this variable instead of P.

3.7 Age-period model

The following exercise is aimed at familiarizing you with the parametrization of the age-period model. It will give you the opportunity explore how to extract and and plot parameter estimates from models. It is based on Danish male lung cancer incidence data in 5-year classes.

1. Read the data in the file `lung5-M.txt` as in the tabulation exercise:

   ```
   lung <- read.table( "../data/lung5-M.txt", header=T )
   lung
   with( lung , table( A ) )
   with( lung , table( P ) )
   with( lung , tapply( Y, list(A,P), sum ) )
   ```

   What do these tables show?

2. Fit a Poisson model with effects of age (A) and period (P) as class variables:

   ```
   ap.1 <- glm( D ~ factor(A) + factor(P) + offset(log(Y)),
                family=poisson, data=lung )
   summary( ap.1 )
   ```

   What do the parameters refer to, i.e. which ones are log-rates and which ones are rate-ratios?

3. Fit the same model without intercept (use `-1` in the model formula); call it `ap.0` — we shall refer to this subsequently. What do the parameters now refer to?

4. Fit the same model, using the period 1968–72 as the reference period, by using the `relevel` command for factors to make 1968 the first level:

   ```
   ap.3 <- glm( D ~ factor(A) - 1 + relevel(factor(P),"1968") + offset(log(Y)),
                  family=poisson, data=lung )
   ```

5. Extract the prameters from the model, by doing:

   ```
   ap.cf <- summary( ap.3 )$coef
   ```

6. Now plot the estimated age-specific incidence rates, remembering to annoatte them with the correct scale. We need the first 10 parameters, with their standard errors:

   ```
   age.cf <- ap.cf[1:10,1:2]
   ```

   This means that we take rows 1–10 and columns 1–2. The corresponding age classes are $40, \ldots, 85$. The midpoints of these age-classes are 2.5 years higher. The ages can be generated in R by saying `seq(40,85,5)+2.5`.

   Now put confidence limits on the curves by taking ±1.96 × s.e.. The line of the estimates can be over-drawn once more in a thicker style:

   ```
   lines( seq(40,85,5)+2.5, exp(age.cf[,1]), lwd=3 )
   ```

7. Now for the rate-ratio-parameters, take the rest of the coefficients:

   ```
   RR.cf <- ap.cf[11:20,1:2]
   ```

   But the reference group is missing, so we must stick two 0s in the correct place. We use the command `rbind` (row-bind):

   ```
   RR.cf <- rbind( RR.cf[1:5,], c(0,0), RR.cf[6:10,] )
   ```

   Now we have the same situation as for the age-specific rates, and can plot the relative risks (relative to 1968) in precisely the same way as for the agespecific rates.

   Make a line-plot of the relative risks with confidence intervals.

8. However, the relevant parameters may also be extracted directly from the model without intercept, using the function `ci.lin` (remember to read the documentation for this!)

   The point is to define a *contrast matrix*, which multiplied to (a subset of) the parameters gives the rates in the reference period. The log-rates in the reference period (the first level of `factor(P)` are the age-parameters. The log-rates in the period labelled `1968` are these *plus* the period estimate from `1968`.

   Now construct the following matrix and look at it:

   ```
   cm.A <- cbind( diag( nlevels( factor(lung$A) ) ), 1 )
   ```

   Now look at the parameters extracted by `ci.lin`, using the `subset=` argument:

   ```
   ci.lin( ap.0, subset=c("A","1968") )
   ```

   Now use the argument `ctr.mat=` in `ci.lin` to produce the rates in period 1968 and plot them on a log-scale.

9. Save the estimates of age aned period effects along with the age-points and period-points, using `save` (look up the help page if you are not familiar with it. You will need these in the next exercise on the age-cohort model.

10. We can also use the same machinery to extract the rate-ratios relative to 1968. The contrast matrix to use is the difference between two: The first one is the one that extracts the rate-ratios with a prefixed 0:

    ```
    cm.P <- rbind(0,diag( nlevels(factor(lung$P))-1 ) )
    cm.P
    ci.lin( ap.0, subset="P", ctr.mat=cm.P )
    ```

    In order to subtract the value corresponding to `1968`, we must subtract a 11 × 10 matrix, that just selects the `1968` column:

    ```
    cm.Pref <- cm.P * 0
    cm.Pref[,5] <- 1
    cm.Pref
    ```

    The contrast matrix to use is the difference between these two:

    ```
    cm.P - cm.Pref
    ci.lin( ap.0, subset="P", ctr.mat=cm.P-cm.Pref )
    ```

Use the Exp=TRUE argument to get the rate-ratios and plot these with confidence intervals on a log-scale.

11. *For the* **real** *nerds:* Plot the rates and the rate ratios beside each other, and make sure that the physical extent of the units on both the x-axis and the y-axis are the same.

    Hint: You may want to use par(mar=c(0,0,0,0), oma=), the function layout as well as the xaxs="i" argument to plot.

3.8 Age-cohort model

This exercise is aimed at familiarizing you with the parametrization of the age-cohort model. It will give you the opportunity explore how to extract and and plot parameter estimates from models. It is parallel to the exercise on the age-period model and is therefor less detailed.

1. Read the data in the file lung5-M.txt as in the tabulation exercise:

    ```
    library(Epi)
    lung <- read.table( "../data/lung5-M.txt", header=T )
    lung
    attach( lung )
    table( A )
    table( P )
    table( P-A )
    ```

    What do these tables show?

2. Fit a Poisson model with effects of age (A) and cohort (C) as class variables. You will need to form the variable C (cohort) as $P - A$ first.

    What do the parameters refer to ?

3. Fit the same model without intercept. What do the parameters now refer to ?

    Hint: Use -1 in the model formula.

4. Fit the same model, using the cohort 1908 as the reference cohort. What do the parameters represent now?

    Hint: Use the Relevel command for factors to make 1968 the first level.

5. What is the range of birth dates represented in the cohort 1908?

6. Extract the age-specific incidence parameters from the model and plot then against age. Remember to annotate them with the correct units. Add 95% confidence intervals.

    Hint: Use the function ci.lin from the Epi package.

7. Extract the cohort-specific rate-ratio parameters and plot then against the date of birth (cohort). Add 95% confidence intervals.

8. Now load the estimates from the age-period model, and plot the estimated age-specific rates from the two models on top of each other.

    Why are they different? In particular, why do they have different slopes?

3.9 Age-drift model

This exercise is aimed at introducing the age-drift model and make you familiar with the two different ways of parametrizing this model. Like the two previous exercises it is based on the male lung cancer data.

1. First read the data in the file `lung5-M.txt` and create the cohort variable:

   ```
   lung <- read.table( "../data/lung5-M.txt", header=T )
   lung$C <- lung$P - lung$A
   ```

   Alternatively you can do:

   ```
   lung <- transform( lung, C = P - A )
   ```

2. Fit a Poisson model with effects of age as class variable and period P as continuous variable.

   What do the parameters refer to ?

3. Fit the same model without intercept. What do the parameters now refer to?

4. Fit the same model, using the period 1968–72 as the reference period.

   Hint: When you center a variable on a reference value `ref`, say, by entering `P-ref` directly in the model formula will cause a crash, because the "`-`" is interpreted as a model operator. You must "hide" the minus from the model formula interpretation by using the identity function, i.e. use: `I(P-ref)`.

   Now what do the parameters represent?

5. Fit a model with cohort as a continuous variable, using 1908 as the reference, and without intercept. What do the resulting parameters represent?

6. Compare the deviances and the slope estimates from the models with cohort drift and period drift.

7. What is the relationship between the estimated age-effects in the two models?

   Verify this empirically by converting one set of age-parameters to the other.

8. Plot the age-specific incidence rates from the two different models in the same panel.

9. The rates from the model are:

   $$\log(\lambda_{ap}) = \alpha_p + \delta(p - 1970.5)$$

   Therefore, with an x-variable: (1943,...,1993) + 2.5, the log rate ratio relative to 1970.5 will be:

   $$\log \mathrm{RR} = \hat{\delta} \times x$$

   and the upper and lower confidence bands:

   $$\log \mathrm{RR} = (\hat{\delta} \pm 1.96 \times \mathrm{s.e.}(\delta)) \times x$$

   Now extract the slope parameter, and plot the rate-ratio functions as a function of period.

3.10 Age-period-cohort model

The following exercise is aimed at familiarizing you with the parametrization of the age-period-cohort model and with the realtionship of the APC-model to the other model that you have been working with, so we will refer back to those, and assume that you have the results from them at hand.

1. Read the data in the file `lung5-M.txt` as in the tabulation exercise:

   ```
   lung <- read.table( "../data/lung5-M.txt", header=T )
   lung
   attach( lung )
   ```

2. Fit a Poisson model with effects of age (A), period (P) and cohort (C) as class variables. Also fit a model with age alone as a class variable. Write down a scheme showing the deviances and degrees of freedom for the 5 models you have models fitted to this dataset.

3. Compare the models that can be compared, with likelihood-ratio tetsts. You will want to use `anova` (or specifically `anova.glm`) with the argument `test="Chisq"`.

4. Next, fit the same model without intercept, and with the first and last period parameters and the 1908 cohort parameter set to 0. Before you do so a few practical things must be fixed:

   You can merge the first and the last period level using the `Relevel` function (look at the documentation for it).

   ```
   lung$Pr <- Relevel( factor(lung$P), list("first-last"=c("1943","1993") ) )
   ```

   You can also use this function to make the 1908 cohort the first level of the cohort factor:

   ```
   lung$Cr <- Relevel( factor(lung$P-lung$A), "1908" )
   ```

   It is a good idea to tabulate the new factor against the old one (i.e. that variable from which it was created) in order to meake sure that the relevelling actually is as you intended it to be.

5. Now you can fit the model, using the factors you just defined. What do the parameters now refer to?

6. Make a graph of the parameters. Remember to take the exponential to convert the age-parameters to rates (and find out what the units are) and the period and cohort parameters to rate ratios. Also use a log-scale for the y-axis. You may want to use `ci.lin` to facilitate this.

7. Fit the same model, using the period 1968–72 as the reference period and two cohorts of your choice as references. To decide which of the cohorts to alias it may be useful to see how many observations there are in each:

```
with( lung, table(P-A) )
with( lung, tapply(D,list(P-A),sum) )
```

Having fitted the model, now what do the parameters in it represent?

8. Make a plot of these parameters.

   Add the parameters from the previous parametrization to the same graph.

3.11 Age-period-cohort model for Lexis triangles

The following exercise is aimed at showing the problems associated with age-period-cohort modelling for triangular data.

Also you will learn how to overcome these problems by parametric modelling of the three effects.

1. Read the Danish male lung cancer data tabulated by age period *and* birth cohort, `lung5-Mc.txt`. List the first few lines of the dataset and make sure you understand what the variables refer to. Also define nthe synthetic cohorts as `P5-A5`:

```
library( Epi )
ltri <- read.table( "../data/lung5-Mc.txt", header=T )
ltri$S5 <- ltri$P5 - ltri$A5
attach( ltri )
```

2. Make a Lexis diagram showing the subdivision of the follow-data. You will explore the function `Lexis.diagram`.

```
Lexis.diagram( age=c(40,90), date=c(1943,1998), coh.grid=TRUE )
```

3. Use the variables `A5` and `P5` to fit a traditional age-period-cohort model with synthetic cohort defined above as `S5=P5-A5`:

```
ms <- glm( D ~ -1 + factor(A5) + factor(P5) + factor(S5) + offset(log(Y)),
           family=poisson, data=ltri )
```

   How many parameters does this model have? (Use the `summary()` function)

4. Now try to fit the model with the “real” cohort variable `C5`:

```
mc <- glm( D ~ -1 + factor(A5) + factor(P5) + factor(C5) + offset(log(Y)),
           family=poisson, data=ltri )
summary( mc )$df
```

   How many parameters does this model have?

5. Plot the parameter estimates from the two models on top of each other, with confidence intervals. Remember to put the correct scales on the plot.

```
par( mfrow=c(1,3) )
a.pt <- as.numeric( levels(factor(A5)) )
p.pt <- as.numeric( levels(factor(P5)) )
s.pt <- as.numeric( levels(factor(S5)) )
c.pt <- as.numeric( levels(factor(C5)) )
matplot( a.pt, ci.lin( ms, subset="A5", Exp=TRUE )[,5:7]/10^5,
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Age", ylab="Rates", log="y" )
matlines( a.pt, ci.lin( mc, subset="A5", Exp=TRUE )[,5:7]/10^5,
          type="l", lty=1, lwd=c(3,1,1), col="blue" )
matplot( p.pt, rbind( c(1,1,1), ci.lin( ms, subset="P5",Exp=TRUE )[,5:7] ),
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Period", ylab="RR", log="y" )
matlines( p.pt, rbind( c(1,1,1), ci.lin( mc, subset="P5",Exp=TRUE )[,5:7] ),
          type="l", lty=1, lwd=c(3,1,1), col="blue" )
matplot( s.pt, rbind(c(1,1,1),ci.lin( ms, subset="S5", Exp=TRUE )[,5:7]),
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Cohort", ylab="RR", log="y" )
matlines( c.pt, rbind(c(1,1,1),ci.lin( mc, subset="C5", Exp=TRUE )[,5:7]),
          type="l", lty=1, lwd=c(3,1,1), col="blue" )
```

How do the confidence limits compare between the three effects?

6. Now fit the model using the proper midpoints of the triangles as factor levels. How many parameters does this model have?

```
mt <- glm( D ~ -1 + factor(Ax) + factor(Px) + factor(Cx) + offset(log(Y)),
           family=poisson, data=ltri )
summary( mt )$df
```

7. Plot the parameters from this model in three panels as for the previous two models.

```
par( mfrow=c(1,3) )
a.pt <- as.numeric( levels(factor(Ax)) )
p.pt <- as.numeric( levels(factor(Px)) )
c.pt <- as.numeric( levels(factor(Cx)) )
matplot( a.pt, ci.lin( mt, subset="Ax", Exp=TRUE )[,5:7]/10^5,
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Age", ylab="Rates", log="y" )
matplot( p.pt, rbind( c(1,1,1), ci.lin( mt, subset="Px",Exp=TRUE )[,5:7] ),
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Period", ylab="RR", log="y" )
matplot( c.pt, rbind(c(1,1,1),ci.lin( mt, subset="Cx", Exp=TRUE )[,5:7]),
         type="l", lty=1, lwd=c(3,1,1), col="black",
         xlab="Cohort", ylab="RR", log="y" )
```

We see that the parameters clearly do not convey a reasonable picture of the effects; som severe indeterminacy has crept in.

8. What is the residual deviance of this model?

```
summary( mt )$deviance
```

9. The dataset also has a variable `up`, which indicates whether the observation comes from an upper or lower triangle. Try to tabulate this variable against `P5-A5-C5`.

```
table( up, P5-A5-C5 )
```

10. Fit an age-period cohort model separately for the subset of the dataset from the upper triangles and from the lowere triangles. What is the residual deviance from each of these models and what is the sum of these. Compare to the model using the proper midpoints as factor levels.

```
m.up <- glm( D ~ -1 + factor(A5) + factor(P5) + factor(S5) + offset(log(Y)),
             family=poisson, data=subset(ltri,up==1) )
summary( m.up )$deviance
m.lo <- glm( D ~ -1 + factor(A5) + factor(P5) + factor(S5) + offset(log(Y)),
             family=poisson, data=subset(ltri,up==0) )
summary( m.lo )$deviance
summary( m.lo )$deviance + summary( m.up )$deviance
summary( mt )$deviance
```

11. Next, repeat the plots of the parameters from the model using the proper midpoints as factor levels, but now super-posing the estimates (in different color) from each of the two models just fitted. What goes on?

```
par( mfrow=c(1,3) )
a.pt <- as.numeric( levels(factor(Ax)) )
p.pt <- as.numeric( levels(factor(Px)) )
c.pt <- as.numeric( levels(factor(Cx)) )
a5.pt <- as.numeric( levels(factor(A5)) )
p5.pt <- as.numeric( levels(factor(P5)) )
s5.pt <- as.numeric( levels(factor(S5)) )
matplot( a.pt, ci.lin( mt, subset="Ax", Exp=TRUE )[,5:7]/10^5,
         type="l", lty=1, lwd=c(2,1,1), col=gray(0.7),
         xlab="Age", ylab="Rates", log="y" )
matpoints( a5.pt, ci.lin( m.up, subset="A5", Exp=TRUE )[,5:7]/10^5,
           pch=c(16,3,3), col="blue" )
matpoints( a5.pt, ci.lin( m.lo, subset="A5", Exp=TRUE )[,5:7]/10^5,
           pch=c(16,3,3), col="red" )
matplot( p.pt, rbind( c(1,1,1), ci.lin( mt, subset="Px",Exp=TRUE )[,5:7] ),
         type="l", lty=1, lwd=c(2,1,1), col=gray(0.7),
         xlab="Period", ylab="RR", log="y" )
matpoints( p5.pt[-1], ci.lin( m.up, subset="P5", Exp=TRUE )[,5:7],
           pch=c(16,3,3), col="blue" )
matpoints( p5.pt[-1], ci.lin( m.lo, subset="P5", Exp=TRUE )[,5:7],
           pch=c(16,3,3), col="red" )
matplot( c.pt, rbind(c(1,1,1),ci.lin( mt, subset="Cx", Exp=TRUE )[,5:7]),
         type="l", lty=1, lwd=c(2,1,1), col=gray(0.7),
         xlab="Cohort", ylab="RR", log="y" )
matpoints( s5.pt[-1], ci.lin( m.up, subset="S5", Exp=TRUE )[,5:7],
           pch=c(16,3,3), col="blue" )
matpoints( s5.pt[-1], ci.lin( m.lo, subset="S5", Exp=TRUE )[,5:7],
           pch=c(16,3,3), col="red" )
```

12. Now, load the splines package and fit a model using the correct midpoints of the triangles as quantitative variables in restricted cubic splines, using the function `ns`:

```
library( splines )
mspl <- glm( D ~ -1 + ns(Ax,df=7,intercept=T)
                    + ns(Px,df=6,intercept=F)
                    + ns(Cx,df=6,intercept=F) + offset(log(Y)),
             family=poisson, data=ltri )
```

13. Compute the residual degrees of freedom for the two models and compare the deviance of the models with these

```
summary( mspl )
summary( mt )$deviance - summary( mspl )$deviance
summary( mt )$df       - summary( mspl )$df
```

How do the deviances compare?

14. Make a prediction of the terms, using `predict.glm` using the argument `type="terms"`, and plot these estimated terms.

15. Repeat the last three questions based on a moedl where you have interchanged the sequence of the period and cohort term.

3.12 Using apc.fit etc.

This exercise is aimed at introducing the functions for fitting and plotting the results from age-period-cohort models: `apc.fit` `apc.plot` `apc.lines` and `apc.frame`.

You should read the help page for the `apc.fit` function, in particular you should be aware of the meaning of the argument

1. Read the testis cancer data and collapse the cases over the histological subtypes:

   ```
   th <- read.table( "../data/testis-hist.txt", header=T )
   str( th )
   ```

   Knowing the names of the variables in the dataset, you can collapse the dataset over the histological subtypes. You may want to use the function `aggregate`; note that there is no need to tabulate by cohort, because even for the triangular data the relationship $c = p - a$ holds.

   Note that the original data had three subtypes of testis cancer, so while it is OK to sum the number of cases (`D`), risk time should not be aggregated across histological subtypes — the aggregation is basically as for competing risks only events are added up, the risk time is the same. (Take a look at the help page for `aggregate`):

2. Present the rates in 5-year age and period classes from age 15 to age 59 using `rateplot`. Consider the function `subset`. To this end you must make a table, for example using something like:

   ```
   with( tc, tapply( D, list(floor(A/5)*5+2.5,
                             floor((P-1943)/5)*5+1945.5), sum ) )
   ```

   — assuming your aggregated data is in the data frame `tc`. and a similar construction for the risk time.

3. Fit an age-period-cohort model to the data using the machinery implemented in `apc.fit`. The function returns a fitted model *and* a parametrization, hence you must choose how to parametrize it, in this case `"ACP"` with all the drift included in the cohort effect and the reference cohort being 1918.

   ```
   tapc <- apc.fit( subset( tc, A>15 & A<60 ), npar=c(10,10,10), parm="ACP", ref.c=1918 )
   ```

   Can any of the effects be omitted from the model?

4. Plot the estimates using the `apc.plot` function:

   ```
   apc.plot( tapc, ci=TRUE )
   ```

5. Now explore in more depth the cohort effect by increasing the number of parameters used for it:

   ```
   tapc <- apc.fit( subset( tc, A>15 & A<60 ), npar=c(10,10,20),
                    parm="ACP", ref.c=1918, scale=10^5 )
   fp <- apc.plot( tapc, ci=TRUE )
   ```

   Do the extra parameters for the cohort effect have any influence on the model fit?

6. Explore the effect of using the residual method instead, and over-plot the estimates from this method on the existing plot:

7. The standard display is not very pretty — it gives an overview, but certainly not anything worth publishing, hence a bit of handwork is needed. Use the `apc.frame` for this, and create a nicer plot of the estimates from the residual model. You may not agree with all the parameters suggested here:

   ```
   par( mar=c(3,4,1,4), mgp=c(3,1,0)/1.7, las=1 )
   fp <- apc.frame( a.lab=seq(20,60,10),
                    a.tic=seq(10,60,5),
                   cp.lab=seq(1900,2000,20),
                   cp.tic=seq(1885,2000,5),
                    r.lab=c(c(1,2,5)/10,1,2,5,10),
                    r.tic=c(1:9/10,1:10),
                      gap=8,
                   rr.ref=1)
   apc.lines( tapc, ci=TRUE, col="blue", frame.par=fp )
   apc.lines( tac.p, ci=TRUE, col="red", frame.par=fp )
   ```

8. Try to repeat the exercise using period as the primary timescale, and add this to the plot as well.

   What is revealed by looking at the data this way?

3.13 Statin use in the Netherlands

Bijlsma *et al.* published an analysis of the prevalence of statin use in the Netherlands [5], available as http://bendixcarstensen.com/APC/MPIDR-2016/Bijlsma.2012.pdf. The authors have kindly put the data at our disposal, so this exercise is partly replicating the analysis in the paper, partly assessing how variants of the model behave.

1. Start by reading the data from the paper — slightly modified so that A and P now are coded as quantitative variables corresponding to the mean in each subset of the Lexis diagram:

   ```
   statin <- read.csv( "../data/statin.csv" )
   str( statin )
   head( statin )
   ```

2. Now fit AP and APC models as described in the paper. In order to fix cohort effects to be 0 for specific cohorts you will need to explore the levels of `factor(P-A)` and subsequently use the function `Relevel` to merge the two levels to the first.

   Make sure you know where the overall prevalence rates goes (with the age-effect perhaps?)

3. Then try to fit a model with suitable smooth terms in age, period and cohort, using for example `apc.fit`. Are the conclusions substantially different with respect to the period and cohort effects?

4. The outcome variable (`D`) is the number of persons that in a given period (calendar year) take out at least one prescription of statins, and the exposure `Y` is the average number of persons in the period. One might then argue that the outcome were better modeled as a fraction and not a rate; that is with a binomial distribution of `D` out of `Y` persons.

   Try the same sequence of models as before and check if similar conclusion emerge when using logit link, log link and complementary log−log link (available as argument to the `binomial` family argument).

5. Finally check if any of the Lee-Carter models provide viable alternatives to the APC-models.

3.14 Lung cancer in Danish women

This exercise is parallel to the example on male lung cancer from the lectures. The point is to fit age-period-cohort models as well as Lee-Carter models and inspect their relative merits and different fits to data on female lung cancer in Denmark.

1. Read the lung cancer data from the file `lung-md.txt` from the data repository, and subset to women only (`sex==2`), and inspect no. of cases per 5-year age-class:

   ```
   library( Epi )
   lC <- read.table( "../data/lung-mf.txt", header=TRUE )
   lF <- subset( lC, sex==2 )
   ```

2. Use `xtabs` to get an overview of cases and incidence rates (per 1000 PY, say), and derive the rates for use with the function `rateplot`.

3. When fitting APC-models and Lee-Carter models we shall use natural splines for fitting, so we must devise knots on the age and time-scales for the splines. Since the informtion in the data on event rates is in the number of *cases*, we would like to place the n knots such that there is $1/n$ between each pair of successive knots and $1/2n$ below the first and above the last knot. Now use the `quantile` function for this, using for example (we do not necessarily want 8 knots):

   ```
   quantile( rep(  A,D), probs=(1:8-0.5)/8 )
   ```

4. Use `apc.fit` to fit an APC-model to data using the chosen knots. You must contemplate the type of parametrization and possible reference points on the perido and cohort scales — read the help page for `apc.fit`.

5. Plot the estimated effects uisng `plot.apc` and possible `apc.frame` for increased control of the plot.

6. For comparison with the APC-model, fit the two Lee-Carter models, one with age-period and one with age-cohort interaction, and compare the fit of these models with the fit of the APC-model. You should use the `LCa.fit` function from the `Epi` package. In order that models be comparable, you must use the same knots for age, period and cohort effects. Alternatively the `lca.rh` function from the `ilc` package.

7. Plot the estimated components of the Lee-Carter models.

8. (This exercise is quite long-winded). In order to get a better view of the behaviour of the different models, plot the predicted rates from the two Lee-Carter models over the time-span of the data frame at select ages (say 50, 60, 70 and 80), using both period and cohort as time-axis. Compare with the fits from the AP, AC and APC-models. Make similar plots of the predicted age-specific rates for select period and cohorts, and again compare the 5 different model fits.

3.15 Histological subtypes of testis cancer

The purpose of this exercise is to handle two different rates that both obey (possibly different) age-period-cohort models. The analysis shall compare rates of seminoma and non-seminoma testis cancer.

1. Read the testis cancer data:

   ```
   th <- read.table( "../data/testis-hist.txt", header=T )
   str( th )
   ```

2. Restrict the dataset to seminomas (`hist`=1) and non-seminomas (`hist`=2), and define `hist` as factor with two levels, suitably named. Also restrict to the age-range relevant for testis cancer analysis, 15–65 years.

3. Make the four classical rate-plots:

   (a) for data grouped in 5 × 5year classes of age and period.

   (b) for data grouped in 3 × 3year classes of age and period.

4. Fit separate APC-models for the two histological types of testis cancer, and plot them together in a single plot.

5. Check whether age, period or cohort effects are similar between the two types:

   (a) by testing formally the interactions

   (b) by plotting the relevant interactions and visually inspecting whether they are alike.

What restrictions are imposed on the parameters for the two models? What restrictions are imposed on the parameters for the rate-ratio?

6. Define a sensible model for description of the two histological types, and report:

   (a) The rates for one type

   (b) The rate-ratio between the types

7. Conlude on the data and graphs.

3.16 Lung cancer: the sex difference

The purpose of this exercise to analyse lung cancer incidence rates in Danish men and women and make comparisons of the effects between the two.

1. Read the lung cancer dataset from the

   ```
   lung <- read.table("../data/apc-Lung.txt", header=T )
   str( lung )
   summary( lung )
   ```

   These data are tabulated by sex, age, period and cohort in 1-year classes, i.e. each observation corresponds to a triangle in the Lexis diagram.

2. The variables `A`, `P` and `C` are the left endpoints of the tabulation intervals. In order to be able to properly analyse data, compute the correct midpoints for each of the triangles.

3. Produce a suitable overview of the rates using the `rateplot` on suitably grouped rates. Make the plots separately for men and women.

4. Fit an age-period-cohort model for male and female rates separately. Plot them in separate displays using `apc.plot`. Use `apc.frame` to set up a display that will accomodate plotting of both sets of estimates.

5. Can you find a way of estimating the ratios of rates and the ratios of RRs between the two sexes (including confidence intervals for them) using only the `apc` objects for males and females separately?

6. Use the function `ns` (from the splines package) to create model matrices describing age, period and cohort effects respectively. Then use the function `detrend` to remove intercept and trend from the cohort and period terms.

   Fit the age-period-cohort model with these terms separately for each sex, for example by introducing an interaction between sex and all the variables (remember that sex must be a factor for this to be meaningful).

7. Are there any of the effects that possibly could be assumed to be similar between males and females?

8. Fit a model where the period effect is assumed to be identical between males and females and plot the resulting fit for the male/female rate-ratios, and comment on this.

3.17 Prediction of breast cancer rates

1. Read the breast cancer data from the text file:

   ```
   library(Epi)
   breast <- read.table("../data/breast.txt", header=T )
   ```

   These data are tabulated be age, period and cohort, i.e. each observation correspond to a triangle in the Lexis diagram.

2. The variables A, P and C are the left endpoints of the tabulation intervals. In order to be able to proper analyse data, compute the correct midpoints for each of the triangles.

3. Produce a suitable overview of the rates using the `rateplot` on suitably grouped rates.

4. Fit the age-period-cohort model with natural splines and plot the parameters (the estimated splines) in a age-period-cohort display.

5. As a starting point for predictions, add the prediction of the period and cohort effects to the plot of the effects, and in particular evaluate the trend in the period respectively cohort trends. You will need to look into the single components of the `apc` object from `apc.fit`. Are these trends invariant under reparametrization ? Which function(s) of them are ?

6. Based on the model fitted, make a prediction of future rates of breast cancer:

   - at the years 2020, 2025, 2030.
   - in the 1960, 65 and 70 generations.

   Use extensions of the estimated period and cohort effects from the natural spline model — note that you will have to refit the model with `glm` in order to make predictions with `ci.pred` sinc the `Model` art of the `apc` object is useless for this.

7. Now fit a model where the knots for period and cohort effecst are moved a bit downward, so that the last piece from which the prediction is done is a bit longer. A simple approach would be to omit the last knot in the natural splines for period and cohort. Compute the identifiable slope at the and of the period resp. cohort effcts.

8. Now fit `glm` versions of these models and compare the predictions for the same dates and cohorts as before between the three models.

3.18 BMI in Australia

The APC-problems are not necessarily tied to analysis of rates and proportions; the identifiabilty problem is on the linear predictor scale. Here is an example of an APC-problem from analysis of a continuous measure, namely BMI.

There are regular health surveys in Australia, and amongst other things information on the body mass index (BMI) of the surveyed persons are collected. In 2014, Peeters *et al.*

published an analysis of the timetrends in BMI in the Australia [6]. For this course the paper is available as
http://bendixcarstensen.com/APC/MPIDR-2016/Peeters.2014.pdf.

A distorted version of the underlying data is available, all dates (birth and survey) have been changes by a small random quantity, so no person's data is traceable.

There is one measurement of BMI per person, and for each measurement we have the sex, date of birth and date of survey (date of measurement). The persons may be regarded as a random sample of the Australian population, so in principle we have measurements of BMI by age and calendar time for each sex.

1. Read the data from the file `bmi.txt` using for example `read.table` and plot the measurement points by age and calendar time.

2. Fit separate linear regression models for the two sexes to the BMI-measurements with non-linear effects of age and calendar time (splines, for example). Show the resulting effects, and check the validity of the model assumptions, in particular the symmetry of the residuals.

3. Check if adding a non-linear cohort effect improves the fit. Consider how to parametrize the resulting model when showing the effects. You would have a look at the function `detrend` for use in modeling and showing the relevant parametrization.

4. Check if a log-transform of the BMI-values improves the fit.

5. (Somewhat log-winded) Get the `quantreg` package and perform separate analyses of BMI for the percentiles (say) 10, 25, 50, 75 and 90. Figure out how to show the results from the different perventiles *jointly*. What is the conclusion?

References

[1] TR Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39:311–324, 1983.

[2] D. Clayton and E. Schifflers. Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine*, 6:449–467, 1987.

[3] D. Clayton and E. Schifflers. Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine*, 6:469–481, 1987.

[4] B Carstensen. Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045, July 2007.

[5] M. J. Bijlsma, E. Hak, J. H. Bos, L. T. de Jong-van den Berg, and F. Janssen. Inclusion of the birth cohort dimension improved description and explanation of trends in statin use. *J Clin Epidemiol*, 65(10):1052–1060, Oct 2012.

[6] A. Peeters, E. Gearon, K. Backholer, and B. Carstensen. Trends in the skewness of the body mass index distribution among urban Australian adults, 1980 to 2007. *Ann Epidemiol*, 25(1):26–33, Jan 2015.