Age-Period-Cohort models:

Statistical inference in the Lexis diagram

www.biostat.ku.dk/~bxc/APC

Department of Biostatistics
Institute of Public Health
University of Copenhagen
March 2005

Bendix Carstensen
Steno Diabetes Center &
Department of Biostatistics
Institute of Public Health
University of Copenhagen
bxc@steno.dk www.biostat.ku.dk/~bxc

Niels Keiding
Department of Biostatistics
Institute of Public Health
University of Copenhagen
nk@biostat.ku.dk

Contents

Chapter 0

Introduction

0.1 Introduction

These notes were written for a course in age-period-cohort modelling or more correctly, a course in "Statistical inference in the Lexis diagram". The course was first given in the spring semester of 2004 at the Department of Biostatistics, Institute of Public Health, University of Copenhagen.

0.1.1 Notation

We use a to denote age, p to denote calendar time (period) and c to denote date of birth (cohort). We use d as event indicator, D as event counter, y and Y for risk time and ℓ as denoting population size at a given date.

This is in contrast to the classical demographic notation where x is used to denote age and t to denote calendar time.

Chapter 1

History of the Lexis diagram

1.1 Introduction

In a very short time span around 1870 a number of pathbreaking expositions of the use of graphical representations as a help in conceptualizing mortality measurement were published, all in German: G.F. Knapp (Leipzig), G. Zeuner (Zürich), K. Becker (Berlin) and W. Lexis (Strassburg (now Strasbourg))/Dorpat (now Tartu)). The dissertations by Brasche (Würzburg) and Verweij (Utrecht) also belong to this effort, and Perozzo (1880a) delivered an impressive *finale*. The graphical representations were accompanied by analytical descriptions, mathematical theory, in particular Zeuner (1869) gave a representation analogous to the fundamental partial differential equation later associated with McKendrick and Von Foerster.

This preliminary presentation does not aim to explain why this golden decade of German demography happened, nor why so little notice was taken of it in cultures with different language. Perhaps Knapp (1874, p. 4) gave a clue:

> So hat sich insbesondere an den Aufgaben über Sterblichkeit in den letzten Jahren eine noch sehr jugendliche Disciplin herangebildet, die man hier und da als mathematische Statistik bezeichnet. Darin wird bekanntlich die Forderung aufgestellt, dass in der Theorie der Statistik die Erscheinungen in ihrer wirklichen Gestalt zu erfassen seien, während man sich früher mit einem willkürlich zugerichteten Abbild begnügte, und ferner wird jene unsicher tastende Rechenkunst, die früher gebräuchlich war, verdrängt durch streng zu begründende Messungsmethoden.

It does indeed seem that these men had a lucky combination of mathematical training, systematic spirit, and realistic motivation. A further cultural historical explanation of this might be interesting.

1.2 The density of the living: the Zeuner sheet

Zeuner (1869) (in a book titled *Abhandlungen aus der Mathematischen Statistik*, cf. the use of this term by Knapp, cited above), studied a function $z = f(t, x)$ of time of birth t and age x defined by the property that

$$V(x) = \int_{t_1}^{t_2} f(t, x)dt$$

is the number of individuals born in $[t_1, t_2]$ who survived age x. Zeuner called the function $f(t, 0)$ (i.e. $P_1PP'P_2$) the *birth curve* and any function $x \to f(t, x)$ (e.g. PMQ) a *mortality curve*. The

current time (or *census time*) $\tau = t + x$ is constant on lines (such as that through B and D) in the (t, x) plane with slope -1. Zeuner's book was inspired by the earlier work of Knapp (1868), but is much more clearly written and contains a total of 27 graphs, most of them representing views of the manifold (we could call it the *Zeuner surface* or the *Zeuner sheet*) $z = f(t, x)$ in three dimensions.

In Zeuner's formalism the *first primary set (erste Hauptgesammtheit) of the living* is $V(x) = \int_{t_1}^{t_2} f(t, x)dt$, the number in generation $[t_1, t_2]$ who survive age x, represented by the area of the figure $B_1M_1M_2B_2$ in Fig. 1

Fig. 1 (Zeuner, 1869, Fig. 1)

and the *second primary set of the living* is $V(\tau) = \int_{t_1}^{t_2} f(t, \tau - t)dt$, the number in generation $[t_1, t_2]$ alive at census time τ, is represented in Fig. 2 by the projection $A_1S_1S_2A_2$ on the YZ plane of the figure $C_1N_1N_2C_2$.

Fig. 2 (Zeuner, 1869, Fig. 2)

Zeuner also defined *secondary sets (Nebengesammtheiten) of the living* such as

$$\int_{t_1}^{\tau_2 - x} f(t, x)dt$$

the number of individuals born after t_1 who survive age x before census at τ_2.

From the survival density $f(t, x)$ Zeuner derived the fundamental general result that the number of deaths in some region A of the (t, x) plane is given as

$$-\int\int_A \frac{\partial}{\partial x} f(t, x)dxdt \quad . \tag{1.1}$$

Three classical *primary sets of deaths* are now delineated by A being a rectangle $[t_1, t_2]\times[x_1, x_2]$, a parallelogram $\{(t, x) : \ t \in [t_1, t_2], \ t+x \in [\tau_1, \tau_2]\}$ or a parallelogram $\{(t, x) : \ t+x \in [\tau_1, \tau_2], \ x \in [x_1, x_2]\}$, see Figs. 3-5. There are also *secondary sets of deaths*.

Fig. 3 (Zeuner, 1869, Fig. 4)

Fig. 4 (Zeuner, 1869, Fig. 5)

Fig. 5 (Zeuner, 1869, Fig. 6)

Zeuner's treatment of practical calculation of life-tables was held in general terms, although this allowed him to isolate a main feature concerning exact calculation of a (generation) life table, according to himself an original observation (... sind *aber bis jetzt nirgends ermittelt worden*). For a fixed generation (born in the interval) $[t_1, t_2]$ a census is taken at τ. If $x_1 = \tau - t_2, x_2 = \tau - t_1$, it is then possible to calculate the exact survival probability $V(x_2)/V(x_1)$ of x_1-years old to age x_2. Indeed, cf. Fig. 6

$$V(x_1) = V(\tau) + M'(x_1, x_2)$$
$$V(x_2) = V(\tau) - M''(x_1, x_2)$$

where $M'(x_1, x_2)$ is the number of deaths of individuals aged between x_1 and x_2 *before* census ($\Delta B_1C_1C_2$) and $M''(x_1, x_2)$ those died between ages x_1 and x_2 *after* census ($\Delta B_2C_1C_2$). The above relations are then obvious from the hatched projections on the YZ plane.

Fig. 6 (Zeuner, 1869, Fig. 13)

The practical implication is that death registers should contain not only age and year at death but also year at birth, and Zeuner strongly argued that this additional information (which "leicht aus Kirchenbuchregistern nachwiesen liesse") be routinely collected.

Zeuner proceeded to develop several approximation formulae, again leaning consistently on his stereometric representations. He took the opportunity to criticize Knapp (1868) for the simplifying proportionality assumption $f(t,x) = \varphi(t)f(x)$. His first article concluded with general treatment of expected life length and mean age of sets of living and sets of dead, always in a consistently simple and elegant, continuous-time notation.

1.3 Knapp's first life lines

Knapp (1869, 1874) took the necessarily discrete nature of actual observations more seriously than Zeuner. To obtain a concrete graphical representation of individual lives, he simply plotted line segments from birth to death, stacked above a time axis, see Fig. 7. The left-hand curve represents the succession of births, and by shifting this curve a certain amount to the right one may read off how many individuals were still alive at that age. The figure $MNOP$ contains a third primary set of deaths ("verhältnismässig die schwierigste").

Fig. 7 (Knapp, 1874, Erste Abhandlung, Fig. 1)

Knapp (1874, erste Abhandlung, dated 1870) went on to rederive all the simpler relations of mortality analysis in discrete age and time (we would say: for the estimators) by extensive use of this graphical representation, and he developed a complete calculus of finite differences for further analysis of the empirical quantities.

1.4 Life lines: the Becker and Lexis age-period-cohort diagrams

Apparently neither Knapp nor Zeuner came across the later so obvious idea of combining their ideas of the individual life line and a (cohort, age) diagram. This seems to have been achieved first by Becker (1874), who modestly characterized his idea as a modification of Knapp's representation, so that the life line of a person born at time t and dead at age x, i.e. time $\tau = t + x$, is the horizontal line segment (t,t) to $(t,t+x)$. Thus, Becker had arrived at a $(t,t+x) = (t,\tau)$-diagram, cf. Fig. 8, where we keep the notation from Zeuner's exposition: birth time t, age x and current time (or census time) $\tau = t + x$, or in modern terms cohort t, age x and period $\tau = t+x$. Becker took care to allow migrations, as is obvious from Fig. 8. In this representation the three primary sets of dead (Knapp's terminology) are represented as the parallelogram $eqng$, the rectangle $ikmo$, and the parallelogram $dlpf$.

Fig. 8 (Becker, 1874, Fig. 1)

In an elaborate critical discussion of historical sources of mortality for use in life insurance calculations, Roghé (1891), a student of Knapp, used this representation (although with the ordinate axis pointing downwards) to pinpoint what he considered crude approximations, see Fig. 9. We recognize the horizontal life lines (here called *Aufenthalte*) and the various parallelogram-shaped observation areas.

Fig. 9 (Roghé, 1891, Anhang 2)

In simultaneous independent work, still carefully acknowledging the pioneering efforts of Knapp (1868), Lexis (1875) studied individual lives in Zeuner's (t, x) plane, see Fig. 10. Lexis (1875) did not explicitly talk about life lines, but *birth points* and *death points* and the lines separating set of these. The vertical lines ($P\Pi$ etc.) separate persons born before or after a given time, while the *isochronical lines*, $t + x = \tau$ constant, have slope -1 ($\Pi Z'$). The primary sets of living and dead, as introduced by Knapp and discussed by Zeuner (cf. Section 2), are all directly illustrated in this figure. For example, the first primary set of the dead is the rectangle *bcgh*, the second (called the third by Knapp and Zeuner) the parallelogram *ekmo* and the third (called the second by Knapp and Zeuner) the parallelogram *peig*. Lexis (1903) later emphasized that his representation corresponds to vertical lines, cf. the entertaining Fig. 11, in which Becker's horizontal life lines are virtually turned 90°! In this latter reference Lexis also took pains in pointing out that his emphasis (on mortality points and their density) was different from Zeuner's.

Fig. 10 (Lexis Fig. 1)

Fig. 11 (Lexis, 1903, I. Abhandlung, Fig. 3)

Lexis (1875) was concerned that the three relevant time dimensions are not represented symmetrically in Fig. 10 (In der Figur ist es vielleicht etwas störend für diese Auffassung, dass die Elementardreiecke gleichschenkelig und nicht gleichseitig sind), and he therefore proposed an alternative *equilateral* diagram, see Fig. 12. In this diagram age, period and cohort are all on the same scale. The equilateral diagram was further developed by Lewin (1876), but I have not yet been able to consult that reference.

Fig. 12 (Lexis, 1875, Fig. 2)

1.5 Perozzo's survey and coloured graphs

A comprehensive survey on the graphical approaches of the above authors was given by Perozzo (1880a), engineer and cartographer at the "Direzione della statistica generale" in Rome. Perozzo's work was immediately translated into German by Lexis (Perozzo, 1880b), who also added a lengthy discussion (Lexis, 1880).

Perozzo's paper is also noteworthy for its *coloured* graphs and for drawing attention to an impressive stereometric representation by Berg (1860) concerning life and death of the Swedish population since 1750. Perozzo preferred the Lexis-Lewin equilateral representation in the (cohort, age)-plane and combined that with Zeuner's sheet to a "Zeuner-Lewin"-representation, cf. Fig. 1.1.

1.6 The dissertations of Brasche and Verweij

Two doctoral dissertations around this time were directly concerned with similar graphical representations:

Fig. 14 (Brasche, 1870)

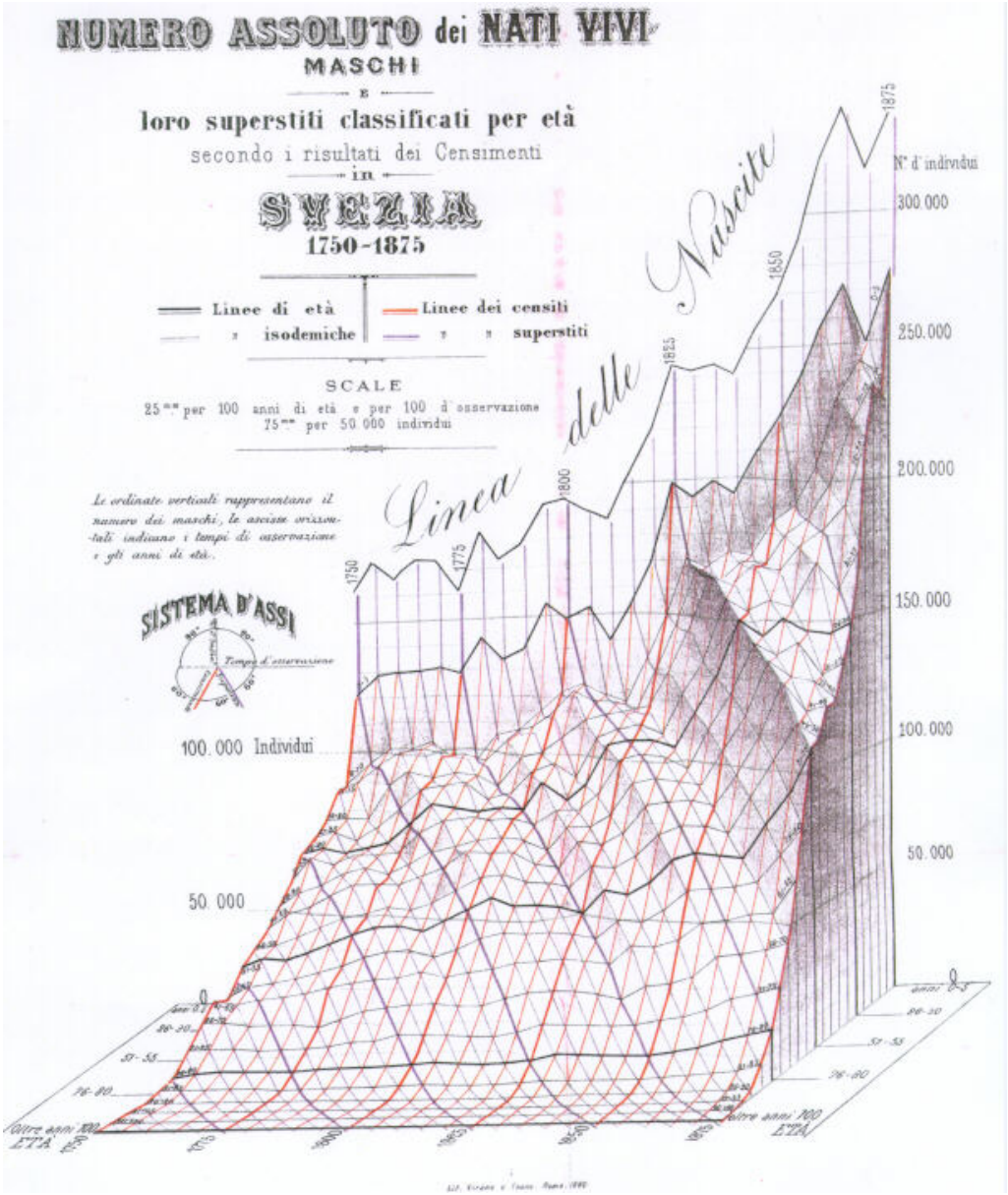


Figure 1.1: *Perozzo, Annali di Statistica, ser. 2, vol. 12, 1880, Tavola V.*

First, the (calendar time, age)-representation by Brasche (1870), Würzburg, cf. Fig. 14. Brasche gave general reference to Knapp (1868), whose graphical representations were however quite different from Brasche's. Indeed, I have identified no other (calendar time, age)-diagrams in the 19th century. Brasche used his graph to give careful and detailed exposition of basic concepts of mortality analysis "*ohne* das Hülfsmittel der Mathematik ...". Vandeschrick (2000) chronicled in detail how relatively little attention was paid to this early work by the other actors in the area.

Secondly, Verweij (1874), Utrecht, translated as Verwey (1875), studied a (cohort, age) diagram with explicit life lines, see Fig. 15. Verweij's work was independent of Lexis's, as acknowledged by the latter (Lexis, 1880).

Fig. 15 (Verwey, 1875, Fig. 1)

1.7 Multistate models: Zeuner on disability, Lexis on marriage and death

Zeuner (1869) devoted the second of his book's three sections to disability models. His most interesting contribution in the present context is the treatment of mortality in open populations, allowing for migration. Fig. 16 shows an age-specific mortality curve M_1M_2 (as usual for Zeuner for a given "generation"), as well as curves B_1B_2 representing immigrations, C_1C_2 emigrations, both in the same density interpretation as Zeuner's mortality density, cf. Section 2. Zeuner provided a detailed mathematical exposition, focusing on

Fig. 16 (Zeuner, 1869, Fig. 22)

convenient approximations, based on three different hypotheses concerning the migrations. Wittstein's hypothesis was that both curves were constant in age, Heym's that deaths and emigrations were constant in age (Zeuner added the assumption of constant immigrations) leading to standard simple differential equations with exponential formula solutions. Finally Zeuner proposed a "new hypothesis", that the immigrations and emigrations were both *proportional* to the mortality curve.

Lexis (1875) devoted his Chapter IV to sets with several changes of states, taking marriage and death as example. For this three-state event history model (we would say), Lexis developed a *stereometric* construction, for which he chose the coordinates time of birth t, age unmarried x, and marriage duration d, cf. Fig. 17. Lexis carefully described all possible primary and secondary sets as well as the interpretation of several sectional planes at many possible angles.

Fig. 17 (Lexis, 1875, Fig. 5)

1.8 Lexis' elaborations

Lexis continued to propose graphical tools for illustrating demographic phenomena. Thus in a paper ambitiously titled "Gesammtübersicht der demographischen Elemente" to the Rome session of the International Statistical Institute (of which he was then vice president) Lexis (1892) proposed all kinds of different symbols for life events (Fig. 18) and he gave some early proposals

Fig. 18 (Lexis, 1903, Fig. 6)

on how to produce summary plots for large populations for which there would be too may life lines. He used (Fig. 19) various centers of gravity and symbols representing the underlying numbers of individuals. Interestingly, Lexis (1892) emphasized that these figures, if useful in practice, would need to use colour. I here quote the figures from the revised version of this article (Lexis, 1903).

Fig. 19 (Lexis, 1903, Fig. 7, 8, 9)

1.9 The handing down of the tradition to the 20th century

Czuber (1903) gave a concise but complete, well-referenced text-book exposition of the Knapp-Zeuner-Becker-Lexis diagrams for explaining basic themes as surveyed in Sections 2-4 above. In another influential textbook Blaschke (1906) took Becker's representation as starting point but went on to explain the "Knapp-Zeuner" theory, including a Zeuner stereogram. Westergaard (1915), in a Danish textbook, made similar choices, while Wicksell (1920) (in Swedish) gave a (cohort, time)-diagram with vertical life lines (Fig. 20), in Becker's style but with the axes interchanged.

Fig. 20 (Wicksell, 1920, Fig. 11)

Wicksell went on to add Zeuner's third dimension, creating, possibly for the first time, a "Zeuner-Becker" stereogram, Fig. 21.

Fig. 21 (Wicksell, 1920, Fig. 14)

Today the "planimetric" representation of life lines, now called the Lexis diagram, is by far the most important of the graphical representations discussed above. However we use neither Becker's (time, cohort) = (τ, t)-diagram with horizontal life lines nor Lexis' (cohort, age) = (t, x)-diagram with vertical life lines, but rather the third possibility: the (time, age) = (τ, x)-diagram with life lines at slope 1.

It is not easy to pinpoint the origin of this representation. I mentioned in Section 6 above that Brasche (1870) used it in his dissertation; another sporadic occurrence is by Elderton (1906) in a discussion not of human lives, but durations of insurance policies (Fig. 22).

Fig. 22 (Elderton, 1906, p. 227)

In the French tradition this version is ascribed (Vandeschrick, 1992) to R. Pressat, who used it in official INSEE publications since the 1950s and in his well-known textbook (Pressat, 1961).

However, the version in Fig. 23 was given in an elementary textbook published in Danish and German by Westergaard and Nybølle (1927, 1928).

Fig. 23 (Westergaard and Nybølle, 1927, p. 364, 1928, Fig. 15)

1.10 The Zeuner sheet and the McKendrick-Von Foerster equation

Following Czuber (1903), define the *mortality density* $\varphi(t, x)$ by

$$\int\int_A \varphi(t, x) dx dt$$

being the number of deaths in the region A of the (t, x) plane. Then Zeuner's fundamental formula (1) is equivalent to the differential equation

$$\frac{\partial}{\partial x} f(t, x) = -\varphi(t, x) \quad .$$

Changing to the coordinates $(\tau, x) = (t + x, x)$ of the present-day "Lexis diagram", let

$$n(\tau, x) = f(\tau - x, x)$$

and

$$\gamma(\tau, x) = \varphi(\tau - x, x) = n(\tau, x)\mu(\tau, x)$$

with $\mu(\tau, x)$ the usual death intensity per individual alive at (τ, x). Since $f(t, x) = n(t+x, x)$ and $\varphi(t, x) = \gamma(t + x, x)$ we then get

$$\frac{\partial}{\partial x}\ f(t, x) = \frac{\partial}{\partial \tau}\ n(t + x, x) + \frac{\partial}{\partial x}\ n(t + x, x)$$

$$= -\varphi(t, x) = -\gamma(t + x, x) = -n(t + x, x)\mu(t + x, x)$$

or

$$\frac{\partial}{\partial \tau}\ n(\tau, x) + \frac{\partial}{\partial x}\ n(\tau, x) = -n(\tau, x)\mu(\tau, x) \quad . \tag{1.2}$$

This is the celebrated *McKendrick-Von Foerster equation*, for which the standard references are McKendrick (1926) and Von Foerster (1959), see e.g. the survey by Keyfitz & Keyfitz (1997). The equation was however mentioned as a routine matter by Westergaard (1925) and given a comprehensive discussion in an appendix to the elementary textbook by Westergaard and Nybølle (1927, pp. 515-521; 1928, pp. 627-634).

With (2) as their basic tool Arthur and Vaupel (1984) revisited the Zeuner sheet, in very much the same spirit as Knapp and Zeuner, but in the modern (time, age)-notation and without explicitly connecting back to Zeuner.

Brillinger (1986) gave a full point-process based statistical theory, in some sense bringing Zeuner's (1869) excellent exposition up to date by incorporating sampling distributions. Keiding (1991) generalized this to three-state illness-death models with the principal aim of developing a basic continuous-time statistical theory of incidence and prevalence. Lund (2000) gave a full point-process based discussion of this work.

Brunet and Struchiner (1996, 1999) returned to Zeuner's mathematical framework, studying differential equations for incidence and prevalence in rather more general situations than Keiding's.

My own experience using Lexis diagram techniques, including level plots, weighted events, and smoothed surfaces, was reported by Keiding et al. (1989), Keiding (1990) (also containing an attempt at a survey of statistical techniques for the Lexis diagram), and Ogata et al. (2000). See Vaupel et al. (1998) for a large set of Lexis diagram contour plots (many in colour).

The Lexis diagram is today a ubiquitous tool in demography, but keeps being reinvented (in orthogonal and equilateral forms) in epidemiology and biostatistics, as surveyed by Keiding (1998).

Acknowledgements. This is a revised version of the manuscript of my talk "Graphical representations in mortality measurement: Knapp, Zeuner, Becker, Lexis" given at the workshop "Lexis in context", Max Planck Institut für demografische Forschung, Rostock, 28 August 2000.

I am grateful for comments and help with references from G. Feichtinger, J. Fleischhacker, A. Hald, G. Scalia-Tomba, C. Vandeschrick and J.W. Vaupel.

Chapter 2

Likelihood for rates

2.1 Statistical models for follow-up data

The ideal version of a register would be a continuous monitoring of the entire population, where each person at regular intervals were classified as having experienced an event or not.

This can be mimicked from truly continuous surveillance data like for example cancer registries by subdividing the available follow-up time for each subject into small intervals. If we assume that intervals are chosen so small that the risk time of each subject is subdivided in a number of intervals each of (a small fixed) length y, say.

Each small interval for a person contributes an observation of an *empirical* rate, (d, y), where d is the number of events in the interval (0 or 1), and y is the length of the interval, i.e. the risk time. This definition is slightly different from the traditional as d/y; it is designed to keep the entire information content in the demographic observation, even if the number of events is 0.

The *theoretical* rate of event occurrence is defined as a function, usually depending on some time scale, t:

$$\lambda(t) = \lim_{h \searrow 0} \frac{\mathrm{P}\left\{\text{event in } (t, t+h) | \text{ at risk at time } t\right\}}{h}$$

However the rate can depend on any number of covariates; incidentally on none at all. Note that in this formulation t has the status of a covariate and h is the risk time.

This definition can immediately be inverted to give the likelihood contribution from an observed empirical rate (d, y), namely the Bernoulli likelihood[1] with probability λy:

$$\begin{aligned} L(\lambda|(d,y)) &= (\lambda y)^d \times (1-\lambda y)^{1-d} = \left(\frac{\lambda y}{1-\lambda y}\right)^d (1-\lambda y) \\ \ell(\lambda|(d,y)) &= d \ln\left(\frac{\lambda y}{1-\lambda y}\right) + \ln(1-\lambda y) \approx d\ln(\lambda) + d\ln(y) - \lambda y \end{aligned}$$

where the term $d\ln(y)$ can be dispensed with because it does not depend on the parameter.

Observation of several independent empirical rates with the same theoretical rate parameter will give rise to a log likelihood that depends on the empirical rates only through $D = \sum d$ and $Y = \sum y$ of the form:

$$\ell(\lambda|(D,Y)) = D\ln(\lambda) - \lambda Y \tag{2.1}$$

[1] The random variables event (0/1) and follow-up time for each person have in this formulation been transformed into a random number of 0/1 variables (of which at most the last can be 1). Hence the validity of the binomial argument, y is not a random quantity, but a fixed quantity.

The contributions to the likelihood from one person will not be independent; but they will be conditionally independent; the total likelihood from one person will be the product of conditional probabilities of the form:

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

Hence the likelihood for a set of empirical rates *looks like* a likelihood for independent observations, but it is not, it is a product (and the log-likelihood a sum) of conditional probabilities.

Thus follow-up studies can be analysed this way in any desired detail; it depends on how large parts of the time scale one is prepared to model with constant rates. Of course the amount and spacing of events limits how detailed the rates can be modelled.

2.2 Regression models for rates

The likelihood for a set of rates is proportional to a likelihood for a Poisson observation with mean λY, and hence the model can be fitted with software that fits Poisson-models.

Programs maximising a likelihood for Poisson observations with mean λY will maximize the log-likelihood function:

$$\ell(\lambda|(D,Y)) = D\ln(\lambda Y) - \lambda Y = D\ln(\lambda) + D\ln(Y) - \lambda Y$$

which has the same maximum as the desired likelihood (2.1), since the extra term $D\ln(Y)$ does not depend on the parameters.

If covariate effects are modelled multiplicatively, the log-rate is modelled additively, and the likelihood for observations with common covariate vector x will be models for $\lambda Y = \exp(x'\beta + \ln(Y))$, so the log-likelihood is:

$$\ell(\beta|(D,Y)) = D(x'\beta + \ln(Y)) - \exp(x'\beta + \ln(Y))$$

This is a likelihood for a Poisson variate with mean $\mu = \exp(x'\beta + \ln(Y))$, so the natural log of the mean is linear in the parameters (the βs). The term $\eta = x'\beta + \ln(Y)$ is called the *linear predictor*, and the natural log is called the *link*-function, because it links the mean and the linear predictor: $\log(\mu) = \eta$.

The total likelihood for an entire study is the sum of such terms over all covariate patterns. It is not necessary to tabulate data by covariate pattern prior to fitting the model; Poisson modelling will work even with single empirical rates for very small intervals. This is because the first term is additive in D, the $\ln(Y)$ irrelevant in the first term and the second term is additive in Y.

The linear predictor contains a term $\ln(Y)$, which can be seen as a variable in the model with a regression coefficient fixed to be 1. This can be accommodated in most software packages declaring $\ln(Y)$ as an *offset*-variable.

If theoretical rates are modelled multiplicatively (a log-linear model), then the parameters $\exp(\beta)$ are estimates of rate ratios corresponding to a change of one in the corresponding covariate.

2.3 Model fit statistics for Poisson regression

The response data used for Poisson regression of rates are the event counter D and the risk time variable Y, and a set of covariates describing the variation of the rates.

In the technical specification of the model it is only the event counter D that appear as response variable; the other part of the response, Y, is used as offset variable after log-transformation. But this is a technical trick to make the program maximize the likelihood for the rates — the model for the *counts* (D) is not a Poisson model, the model for the rates just happens to have the same likelihood as a Poisson model for the counts with the person-years treated as fixed: two different models can have identical likelihoods (but not vice versa).

The Poisson-likelihood was derived under the assumption that the rate is constant; i.e. the assumption is that during all risk time contributing to a single observation in the dataset, the rate has been constant. Thus, the maximal degree of detail that one can obtain in modelling the rates is determined by the size of the cells of the tabulation of D and Y.

When fitting a Poisson model a common so called goodness of fit statistic is output by most programs: the deviance (or as termed by R: the *residual* deviance). This statistic is the log-likelihood ratio statistic comparing the fitted model with the model with a perfect fit, i.e. with one parameter per observation in the dataset. Thus, the deviance is is a function of the model fitted *and* the dataset (how finely the data are tabulated). If a model assuming rates to be constant in 5-year classes were fitted using a dataset where cases and person-years were tabulated in 1-year classes, one would get exactly the same estimates as if data were tabulated in 5-year classes, but the deviance would be different.

As an example we analysed lung cancer incidence among Danish men by an age-period model assuming that rates to be constant in 5-year classes. Using a dataset with cases and person-years tabulated in one-year classes (50 age-classes 40–89, and 54 periods 1943–97) we get a deviance of 6703.90 on 2680 df., whereas analysis based on data tabulated in 5-year classes gave a deviance of 2723.47 on 90 df. But both analyses give exactly the same set of parameter estimates, and so predict the same rates.

Programs that illustrate this can be found in the file `deviance-ex.R` and the output in figure 2.1.

However *differences* between deviances between models are log-likelihood ratio statistics for testing reduction of one model to another. The requirement for this to be valid is that one model is a submodel the other. The distribution of the difference between two deviances is χ^2-distributed with a number of degrees of freedom equal to the difference in numbers of parameters in the two models.

In the lung cancer example mentioned above we added synthetic 5-year cohorts to the model and the resulting age-period-cohort model had a deviance of 4188.99 on 2662 d.f. for the one-year tabulated data, and 208.55 on 72 d.f. for the 5-year tabulated data. The difference to the deviance for the age-period model is:

$$2723.46 - 208.55 = 6703.90 - 4188.99 = 2514.91$$

independent of the data-tabulation.

The deviance cannot in general be taken as a goodness of fit statistic for a model. In order for this to be the case, the saturated model, i.e. the one with one parameter per observation in the dataset must be a relevant comparison model. The mere fact that the saturated model is based on a more or less arbitrarily chosen fineness of tabulation does not make the comparison sensible. And therefore a formal test based on the deviance may not be relevant.

The deviance is primarily a quantity to be used for *comparing* models.

```
 R 1.9.0
 ------------------------------------------------
 Program:  ..\r\deviance-ex.R
  Folder:  C:\Bendix\Undervis\APC\notes
 Started:  mandag 05. april 2004, 16:31:47
 ------------------------------------------------
> # Read the 1 by 1 by 1 year tabulated data on lung cancer in DK and restrict to males
> L1 <- read.table( "../data/lung-apc.txt", header=T )
> L1 <- L1[L1$sex==1,]
> L1$A1 <- floor( L1$A )
> L1$P1 <- floor( L1$P )
>
> # Tabulate in 5-year intervals and make a data-frame
> d5 <- tapply( L1$D, list( L1$A5, L1$P5 ), sum )
> y5 <- tapply( L1$Y, list( L1$A5, L1$P5 ), sum )
> L5 <- data.frame( expand.grid( dimnames( d5 ) ),
+                   D=as.vector(d5), Y=as.vector(y5) )
> names( L5 )[1:2] <- c("A5","P5")
> L5$A5 <- as.numeric( as.character( L5$A5 ) )
> L5$P5 <- as.numeric( as.character( L5$P5 ) )
> L5$C5 <- L5$P5 - L5$A5
>
> # Tabulate in 1-year intervals and make an data-frame
> d1 <- tapply( L1$D, list( floor(L1$A), floor(L1$P) ), sum )
> y1 <- tapply( L1$Y, list( floor(L1$A), floor(L1$P) ), sum )
> L1 <- data.frame( expand.grid( dimnames( d1 ) ),
+                   D=as.vector(d1), Y=as.vector(y1) )
> names( L1 )[1:2] <- c("A1","P1")
> L1$A5 <- floor(  as.numeric(as.character(L1$A1))/5 ) * 5
> L1$P5 <- floor( (as.numeric(as.character(L1$P1))-1943)/5 ) * 5 + 1943
> L1$C5 <- L1$P5 - L1$A5
>
> # Fit the models:
> ap5  <- glm( D ~ factor(A5) - 1 + factor(P5) + offset(log(Y)), family=poisson, data=L5 )
> apc5 <- update( ap5, . ~ . + factor(C5) )
> ap1  <- update( ap5 , data=L1 )
> apc1 <- update( apc5, data=L1 )
>
> # Compare parameter estimates from the two tabulations
> cbind( summary( ap5 )$coef[,1:2], summary( ap1 )$coef[,1:2] )
                  Estimate Std. Error    Estimate Std. Error
factor(A5)40   -10.3423479 0.04192098 -10.3423479 0.04192098
factor(A5)45    -9.3897676 0.03453519  -9.3897676 0.03453519
factor(A5)50    -8.5599766 0.03145070  -8.5599766 0.03145070
factor(A5)55    -7.9282241 0.03020492  -7.9282241 0.03020491
factor(A5)60    -7.4797607 0.02970184  -7.4797607 0.02970184
factor(A5)65    -7.1907539 0.02956000  -7.1907539 0.02955999
factor(A5)70    -7.0245116 0.02969777  -7.0245116 0.02969777
factor(A5)75    -7.0325494 0.03030666  -7.0325494 0.03030666
factor(A5)80    -7.1659457 0.03208700  -7.1659457 0.03208700
factor(A5)85    -7.4325185 0.03846618  -7.4325185 0.03846618
factor(P5)1948   0.3920570 0.03629482   0.3920570 0.03629482
factor(P5)1953   0.6759239 0.03403607   0.6759239 0.03403607
factor(P5)1958   1.0143362 0.03226353   1.0143362 0.03226353
factor(P5)1963   1.2666610 0.03130075   1.2666610 0.03130075
factor(P5)1968   1.4871744 0.03066768   1.4871744 0.03066768
factor(P5)1973   1.5923909 0.03038760   1.5923909 0.03038760
factor(P5)1978   1.6799356 0.03020190   1.6799356 0.03020189
factor(P5)1983   1.6990178 0.03015189   1.6990178 0.03015189
factor(P5)1988   1.5995837 0.03028010   1.5995837 0.03028010
factor(P5)1993   1.5255770 0.03077608   1.5255770 0.03077607
>
> # Print the deviances
> rbind( unlist( summary( ap5  )[c("deviance","df")] )[c(1,3)],
+        unlist( summary( apc5 )[c("deviance","df")] )[c(1,3)],
+        unlist( summary( ap1  )[c("deviance","df")] )[c(1,3)],
+        unlist( summary( apc1 )[c("deviance","df")] )[c(1,3)] )
      deviance  df2
[1,] 2723.4660   90
[2,]  208.5476   72
[3,] 6703.9036 2680
[4,] 4188.9852 2662
>
>
 ------------------------------------------------
 Program: ..\r\deviance-ex.R
  Folder: C:\Bendix\Undervis\APC\notes
   Ended: mandag 05. april 2004, 16:31:50
 Elapsed: 00:00:03
 ------------------------------------------------
```

Figure 2.1: *Sample* R*-program illustrating the deviance dependence on data for the Danish lung cancer data.*

Chapter 3

Classical approach to age-period-cohort modelling.

This chaper is largely based on the two papers by Clayton & Schifflers [3, 4].

The classical approach to descriptive analysis of register data has been to tabulate events by age and date of event, in rectangular (mostly quadratic and mostly 5×5 year) subsets of the Lexis diagram. Corresponding population risk time figures are then computed to a good approximation by averaging the population figures at the ends of each period, or by using the population size at the middle of the period and multiplying it with the period length. (As we shall see later, a better approximation can be obtained).

3.1 Age-period tabulation in the Lexis diagram

Suppose for example cases of cancer are tabulated by age and date of diagnosis (period) in 5-year intervals, and that corresponding risk time figures are available. As an example consider the number of testis cancer cases in Denmark 1948–92 given in table 3.1 and cooresponding male person-years in 3.2.

Each cell in the table represents an observation in the analysis dataset, which in this example has $9 \times 9 = 81$ observations. The rates are descibed by the count variable D (table 3.1) and the person-years variable Y (table 3.2).

Description of these rates by age at diagnosis, date of diagnosis (period) and date of birth (cohort), requires a coding of these three variables, as seen in the tables 3.3, 3.4 and 3.5, respectively. The cohort variable (date of birth) is simply defined as period−age.

Table 3.1: *Number of cases (D) of testis cancer in Denmark.*

	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92
15–19	7	13	13	15	33	35	37	49	51
20–24	31	46	49	55	85	110	140	151	150
25–29	62	63	82	87	103	153	201	214	268
30–34	66	82	88	103	124	164	207	209	258
35–39	56	56	67	99	124	142	152	188	209
40–44	47	65	64	67	85	103	119	121	155
45–49	30	37	54	45	64	63	66	92	86
50–54	28	22	27	46	36	50	49	61	64
55–59	14	16	25	26	29	28	43	42	34

Table 3.2: *Number of person-years (Y) for men in Denmark.*

	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92
15–19	744.2	794.1	972.9	1051.5	961.0	952.5	1011.1	1005.0	929.8
20–24	744.7	721.8	770.9	960.3	1053.8	967.5	953.0	1019.7	1017.3
25–29	781.8	723.0	698.6	764.8	962.7	1056.1	960.9	956.2	1031.6
30–34	774.5	769.3	711.6	700.1	769.9	960.4	1045.3	955.0	957.1
35–39	782.9	760.2	760.5	711.6	702.3	767.5	951.9	1035.7	948.6
40–44	754.3	768.5	749.9	756.5	709.8	696.5	757.8	940.3	1023.7
45–49	676.7	737.9	753.5	738.1	746.4	698.2	682.4	743.1	923.4
50–54	600.3	653.9	715.4	732.7	718.3	724.2	675.5	660.8	721.1
55–59	512.8	571.1	622.5	680.8	698.2	683.8	686.4	640.9	627.7

Table 3.3: *Mean age in the cells of the tables.*

	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92
15–19	17.5	17.5	17.5	17.5	17.5	17.5	17.5	17.5	17.5
20–24	22.5	22.5	22.5	22.5	22.5	22.5	22.5	22.5	22.5
25–29	27.5	27.5	27.5	27.5	27.5	27.5	27.5	27.5	27.5
30–34	32.5	32.5	32.5	32.5	32.5	32.5	32.5	32.5	32.5
35–39	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5	37.5
40–44	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5	42.5
45–49	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5
50–54	52.5	52.5	52.5	52.5	52.5	52.5	52.5	52.5	52.5
55–59	57.5	57.5	57.5	57.5	57.5	57.5	57.5	57.5	57.5

Thus the dataset required has 5 variables:

1. D - number of cases.
2. Y - the amount of risk time (person-years)
3. A - age class at diagnosis
4. P - period of diagnosis
5. C - period of birth (cohort)

The cohort variable $C = P - A$ represents what is normally termed *synthetic* cohorts. Any person will contribute risk time in two adjacent synthetic cohorts. Each synthetic cohort will

Table 3.4: *Mean date of diagnosis (period) in the cells of the tables. The convention is that 1950.0 refers to 1 January 1950.*

	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92
15–19	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
20–24	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
25–29	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
30–34	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
35–39	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
40–44	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
45–49	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
50–54	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5
55–59	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5

Table 3.5: *Mean date of birth in the cells of the tables. The convention is that 1950.0 refers to 1 January 1950.*

	1948–52	1953–57	1958–62	1963–67	1968–72	1973–77	1978–82	1983–87	1988–92
15–19	1933.0	1938.0	1943.0	1948.0	1953.0	1958.0	1963.0	1968.0	1973.0
20–24	1928.0	1933.0	1938.0	1943.0	1948.0	1953.0	1958.0	1963.0	1968.0
25–29	1923.0	1928.0	1933.0	1938.0	1943.0	1948.0	1953.0	1958.0	1963.0
30–34	1918.0	1923.0	1928.0	1933.0	1938.0	1943.0	1948.0	1953.0	1958.0
35–39	1913.0	1918.0	1923.0	1928.0	1933.0	1938.0	1943.0	1948.0	1953.0
40–44	1908.0	1913.0	1918.0	1923.0	1928.0	1933.0	1938.0	1943.0	1948.0
45–49	1903.0	1908.0	1913.0	1918.0	1923.0	1928.0	1933.0	1938.0	1943.0
50–54	1898.0	1903.0	1908.0	1913.0	1918.0	1923.0	1928.0	1933.0	1938.0
55–59	1893.0	1898.0	1903.0	1908.0	1913.0	1918.0	1923.0	1928.0	1933.0

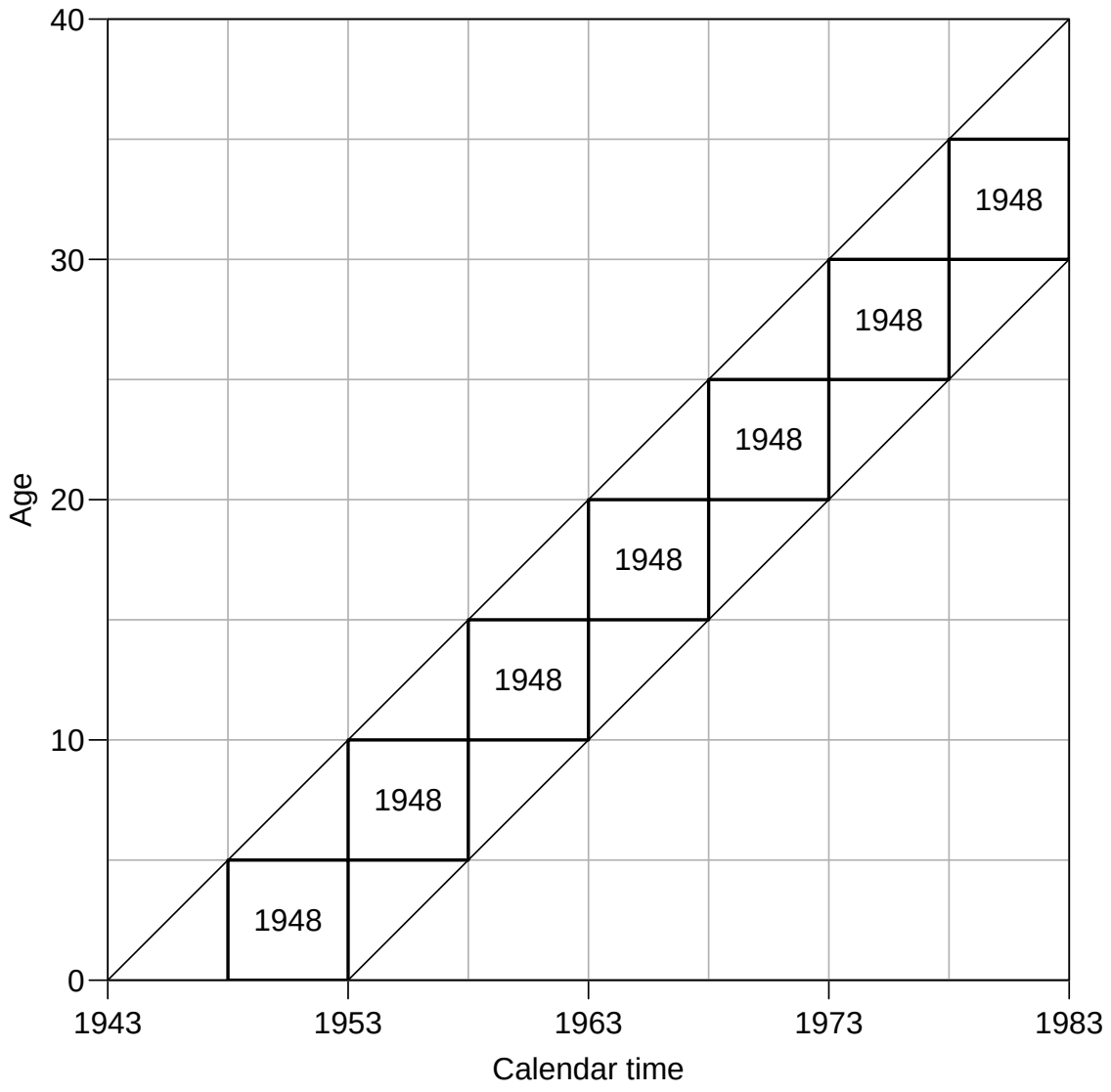


Figure 3.1: *The synthetic cohort 1948 in the Lexis diagram with age and period classification.*

include risk time (and cases) from persons with birth dates in a 10-year interval. For example, for the age class 20–24 years with mean age 22.5 in period 1968–72 with mean date of diagnosis 1970.5 (= 1 July 1970) has mean date of birth $1970.5 - 22.5 = 1948$. But it comprises persons born between $1968 - 25 = 1943$(=1 January 1943) and $1973 - 20 = 1953$(=31 December 1952), i.e. a ten-year period.

This is illustrated in figure 3.1.

The rates can be modelled as functions of age, period and cohort by letting D be the response, $\log(Y)$ the offset and A, P and C categorical explanatory variables in a Poisson model.

The latter means that indicators for each of the levels of the three are generated and entered into the model. The details of such models are elaborated below.

However it is recommendable to start out by graphing the data.

3.1.1 Classical graphing of rates

If rates are observed in a regular grid, i.e. tabulated by two factors with the same class width, as for example the lung cancer data that are tabulated by age and period in 5-year classes, it is relevant to plot the empirical rates for these subgruops.

There are four basic plots of interest:

Age specific rates for each period Here we plot the age-specific rates for each period of observation in a coordinate system with age along the x-axis and rates along the y-axis, the latter usually devised as a logaritmic axis in order to be able to see the structure in the smaller rates too. Thus we get one curve for each period, as shown in figure **??**. If these curves are

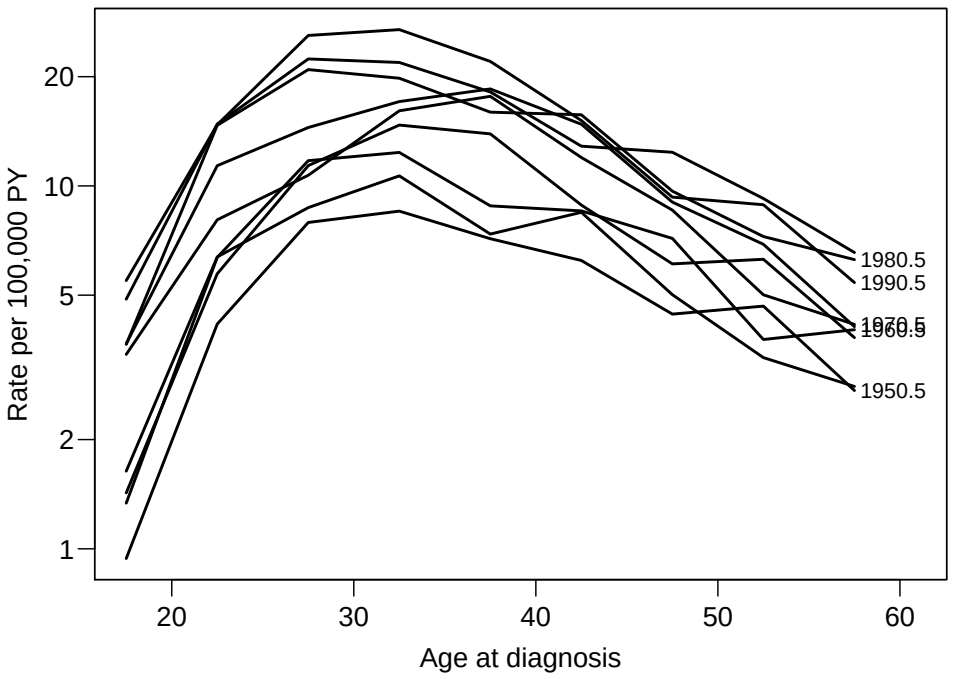


Figure 3.2: *Age-specific rates of testis cancer in Denmark for the periods 1948–52,. . . ,1988–92. Rates connected within periods of diagnosis.*

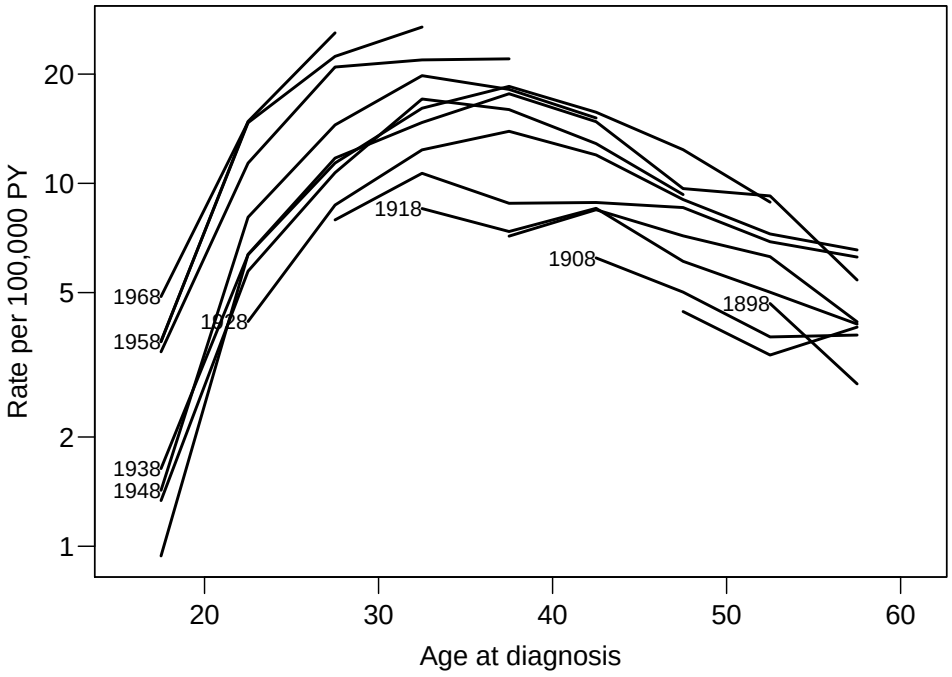


Figure 3.3: *Age-specific rates of testis cancer in Denmark for the periods 1948–52,. . . ,1988–92. Rates connected within birth-cohorts.*

approximately parallel, it is an indication that that the all age-specific rates vary in the same way by period.

Age specific rates for each cohort Here we plot the age-specific rates for each cohort in a coordinate system with age along the x-axis and rates along the y-axis. Thus we get one curve for each cohort, as shown in figure 3.3. If the observation plan is for a fixed period and for all ages in that period, the curves will have different lengths. Note that the points connected are the same as the points in the plot in figure 3.2, just connected differently; connecting them both ways can produce informative or totally fuzzy plots.

If these curves are approximately parallel, it is an indication that that the age-specific rates vary in the same way by cohort.

Rates for each age versus period Here the rates for a given age are plotted against period of diagnosis. This another way of seeing whether the major variation in the rates are by period; if this is the case, these curves should be parallel. The plot is illustrated for testis cancer in Denmark in figure 3.4.

Rates for each age versus cohort Here the rates for a given age are plotted against period of birth (cohort). This another way of seeing whether the major variation in the rates are by cohort; if this is the case, these curves should be parallel. The plot is illustrated for testis cancer in Denmerk in figure 3.5. Note that this plot is showing the same curves as those in figure 3.4, only offset by the length of one period against each other.

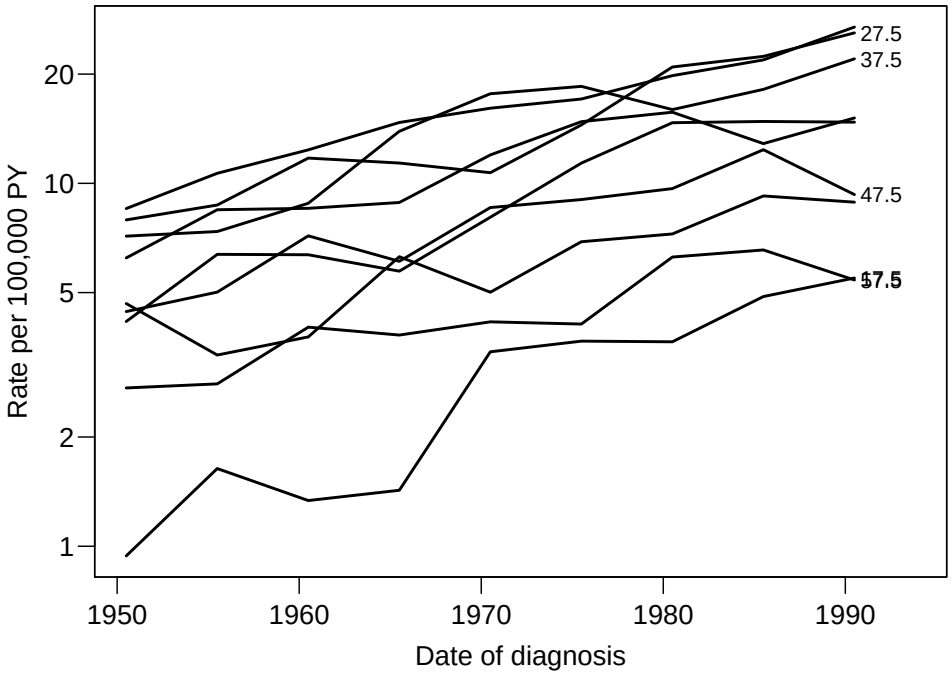


Figure 3.4: *Age-specific rates of testis cancer in Denmark for the periods 1948–52,. . . ,1988–92. Rates in each age-class plotted against date of diagnosis.*

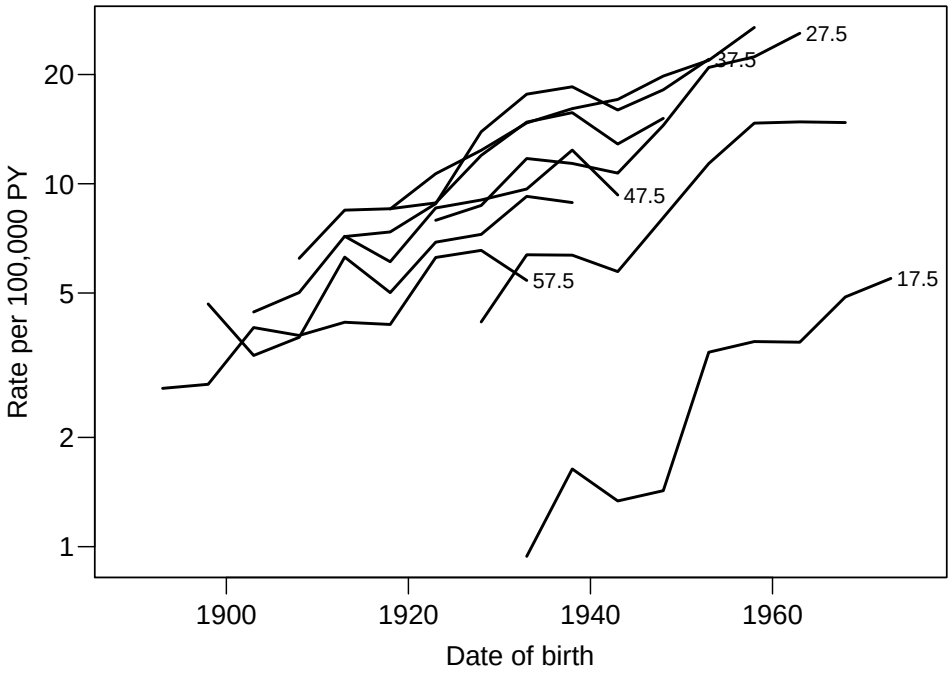


Figure 3.5: *Age-specific rates of testis cancer in Denmark for the periods 1948–52,. . . ,1988–92. Rates in each age-class plotted against date of birth.*

3.2 The age-period model

The age-period model states that the age-specific rates have the same shape in all periods, but possibly with a varying level:

$$\lambda(a,p) = a_a \times b_p \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \beta_p$$

This model has one parameter per age class and one per period, but as always there is a one-parameter unidentifiability in the formulation — a constant may be added to the αs provided the same constant is subtracted from the βs.

The natural constraint from an epidemiological point of view is to fix one period parameter to be 0, $\beta_{p_0} = 0$. For period p_0 we will then have:

$$\log[\lambda(a,p_0)] = \alpha_a + \beta_{p_0} = \alpha_a$$

Thus the αs are logs of age-specific rates for period p_0, and hence the age-specific rates in that period are $\exp(\alpha_a)$.

Comparing rates in any age class between period p and period p_0 gives:

$$\log[\text{RR}] = \log[\lambda(a,p)/\lambda(a,p_0)] = \alpha_a + \beta_p - (\alpha_a + \beta_{p_0}) = \beta_p$$

so the βs are log rate-ratios relative to period p_0. This rate ratio is the same for all age-classes.

The results from this model should be reported as two curves: The age-specific rates in reference period p_0 and the rate ratios relative to this period. Note that the units in which the population risk time ("person-years") are supplied to the computer program will be reflected in the scale of the estimated rates. Thus if the risk time is entered as person-millenia (1000 person-years), the $\exp(\alpha_a)$s will be rates per 1000 person years.

Computer programs will provide estimates and standard errors of the parameters, that are used to obtain confidence intervals as estimate $\pm 1.96 \times$ standard error, and these can then be plotted against age and date of diagnosis, respectively.

The estimates emerging from this parametrization are shown in figure 3.7. Note that we have chosen to display the rate-estimates and the rate-ratio-estimates on a log-scale. This is the natural thing to do when we use a multiplicative model.

The age-specific rates are *cross-sectional* rates referring to the period p_0. Thus they do not have a biological interpretation, but rater a demographic — it is what we would expect to see in a population during a short period of time. Similarly the period parameters describe how these change as time goes by. The age-period model is set up to have this interpretation.

Once the model is fitted it is desirable to compare the fit of the model to the empirical rates, i.e. to plot the observed and predicted rates in the same coordinate system. This is best done using the two period-displays, i.e. plotting rates against age for each period and rates against period for each age. Both these displays will produce a set of parallel curves on a log-scale.

They are illustrated in figure **??**

3.2.1 Practicalities in fitting the age-period model

The parametrization suggested above is however not standard in neither R nor ***SAS***, so a little care is needed to obtain the desired parameters.

In order to obtain the estimates as desired one must use both age and period as factors (R) or class variables (***SAS***). In order to obtain log-rates as parameters, the default intercept must be excluded from the model and age must be the first term mentioned in the model. Finally the

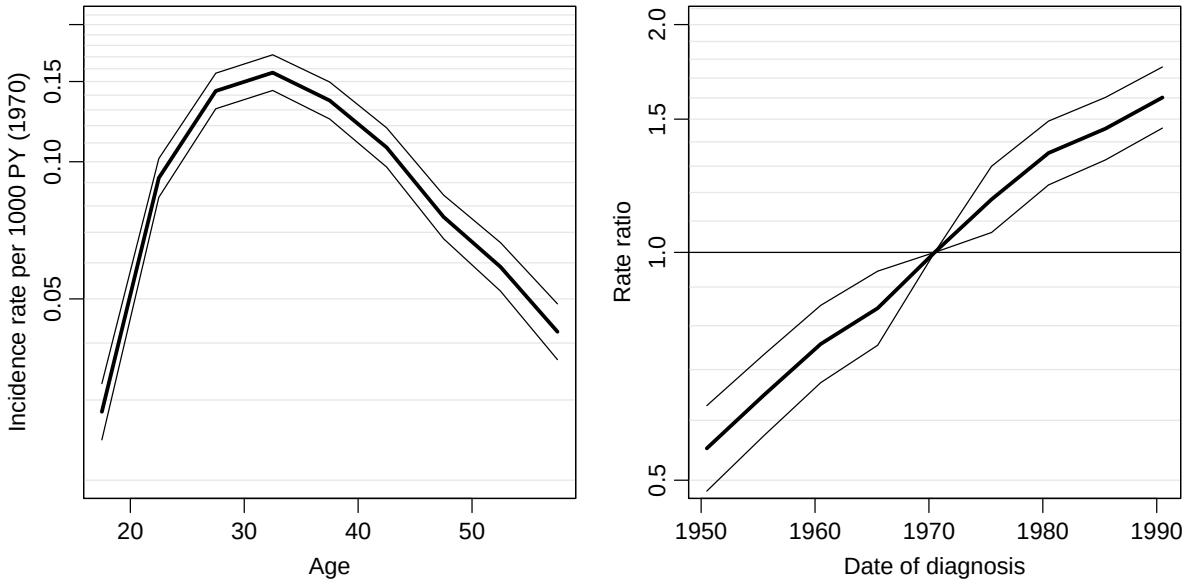


Figure 3.6: *Estimates from the age-period model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits. Note how the confidence limits on rate-ratio plot reveals the reference period.*

reference period must be chosen in some way. Otherwise R will use the first period and ***SAS*** the last as reference.

A sample R-code for this would look like (assuming we want the 5th period as the reference):

```
ap <- glm( D ~ factor( A ) - 1 + relevel( factor( P ), 5 ) + offset( log( Y ) ),
           family = poisson, data = testis )
```

The "`-1`" is the R-notation for omitting the intercept.

From the output of this program one could then derive the rates and the rate ratios with confidence intervals, and make appropriate plots of them, as shown in figure 3.7.

3.3 The age-cohort model

The age-cohort model is similar to the age-period model; it states that the age-specific rates have the same shape for any cohort, but possibly with a varying level:

$$\lambda(a,c) = a_a \times c_c \qquad \text{or} \qquad \log[\lambda(a,p)] = \alpha_a + \gamma_c$$

Again, the natural constraint from an epidemiological point of view is to fix one cohort parameter to be 0, $\gamma_{c_0} = 0$. For period c_0 we will then have:

$$\log[\lambda(a,c_0)] = \alpha_a + \gamma_{c_0} = \alpha_a$$

The αs has the interpretation as age-specific log-rates for cohort c_0, and hence the age-specific rates are $\exp(\alpha_a)$. The γs are log rate-ratios relative to cohort c_0. This rate ratio is the same for all age-classes. But as opposed to the age-period model the estimates relating to the youngest and oldest cohort are less precise because they will be based on only a few cells that are likely to have rather little information.

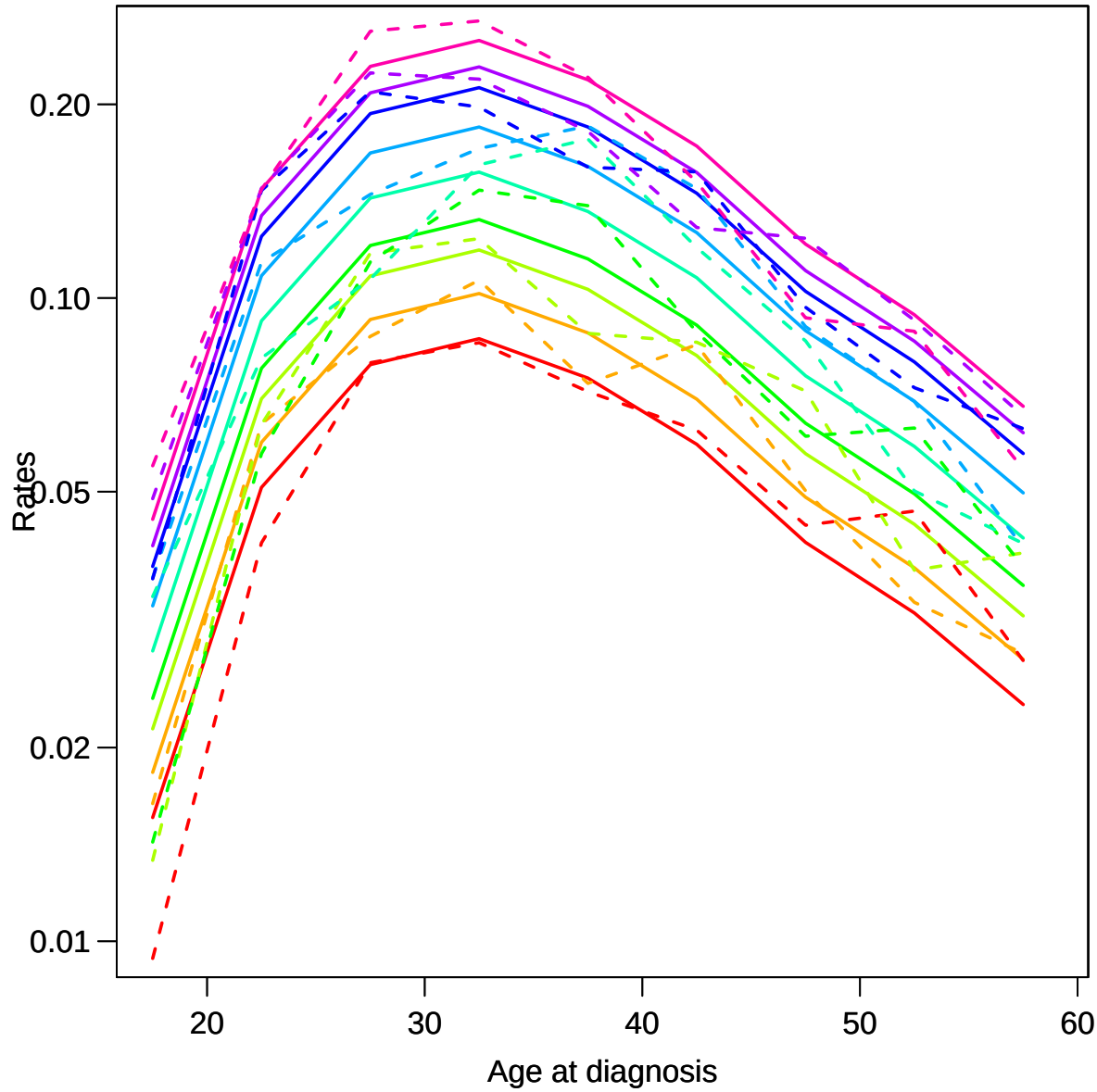


Figure 3.7: *Estimated (predicted) rates from the age-period model fitted to the Danish testis cancer data, together with the observed rates (dotted lines).*

The results from this model should be reported as two curves: The age-specific rates in reference cohort c_0 and the rate ratios relative to this cohort. The estimates emerging from this parametrization are shown in figure 3.8. Note how the confidence limits for the cohort effects illustrate the phenomenon with less information in the early and late cohorts, the are wider towars the ends, as opposed to what was the case for the period parameters in the age-period model.

These age-specific rates are thus rates that one would expect to see in a group of persons born at c_0 if they were followed over the ages in the analysis. So these kind of rates more readily lend themselves to a biological interpretation. The same is the case with the cohort parameters; they describe the changes in incidence rates from one cohort to the next, so they

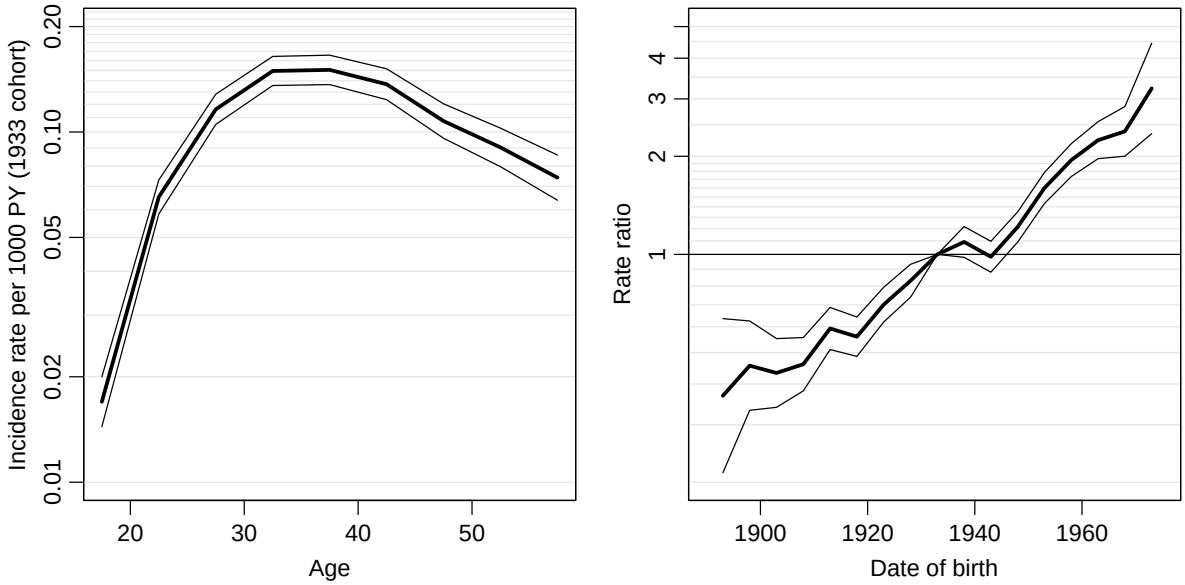


Figure 3.8: *Estimates from the age-cohort model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits. Note how the confidence limits on rate-ratio plot reveals the reference cohort.*

are in a sense comparing different groups of persons. The period parameters from the age-period model compare different time period based on data from roughly the same set of persons (at least if periods are close).

The practical aspects of fitting this model are exactly the same as for the age-period model.

3.4 The age-drift model

Inspection of the rate-ratio plots in figure 3.7 and could suggest to replace the period parameters by a linear trend in log-rates:

$$\log[\lambda(a,p)] = \alpha_a + \beta_p = \alpha_a + \beta(p - p_0)$$

that is $\beta_p = \beta(p - p_0)$.

This would imply that the rate-ratio display (on the log-scale) would show a straight line. The result of fitting this model is given in figure 3.9

Likewise, possibly requiring a slightly more brutal approach, the same approach may be taken to the age-cohort model, substituting it by:

$$\log[\lambda(a,p)] = \tilde{\alpha}_a + \gamma_c = \tilde{\alpha}_a + \gamma(c - c_0)$$

that is $\gamma_c = \gamma(c - c_0)$ The result of fitting this model is given in figure 3.10.

If we look at the computer-output from the two models we will note that the (residual) deviance and other goodness of fit statistics are identical. Furthermore the fitted rates will also be the same under the two models, and finally the estimates of β and γ are identical with identical standard errors as well.

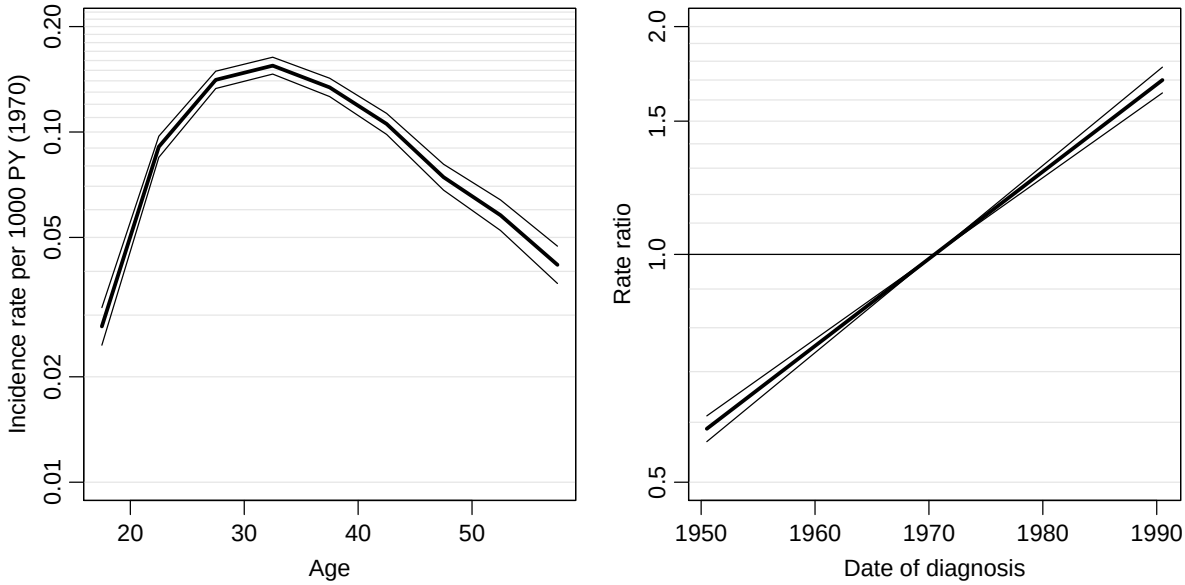


Figure 3.9: *Estimates from the age-period-drift model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits.*

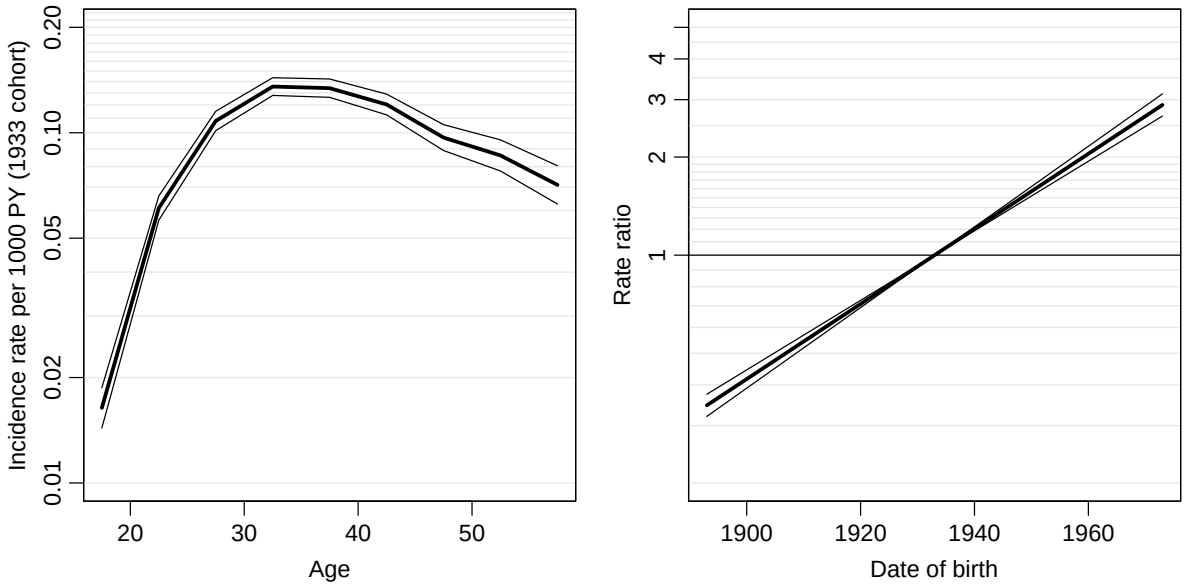


Figure 3.10: *Estimates from the age-cohort-drift model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits.*

Analytically, it can be seen that the two models are the same, by using that $p = a + c$ (and defining $a_0 = p_0 - c_0$):

$$\alpha_a + \beta(p - p_0) = \alpha_a + \beta\big(a + c - (a_0 + c_0)\big) = \alpha_a + \beta(a - a_0) + \beta(c - c_0)$$

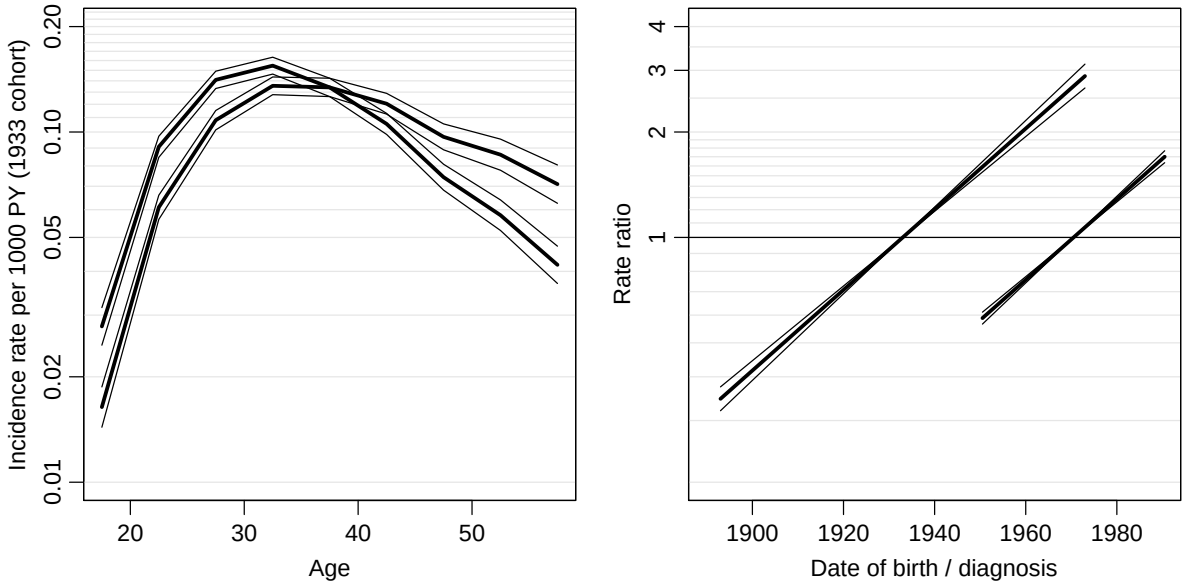


Figure 3.11: *Estimates from the age-drift model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. Parametrizations by period drift or cohort drift. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits.*

Thus, there is only one age-drift model. The interpretation of this is that rates increase exponentially by time (calendar time or cohort) at the same pace, $\exp(\beta) = \exp(\gamma)$ per year for all age classes.

So going from the age-period-drift model to the age-cohort-drift model is just to replace the age effect α_a by $\tilde{\alpha}_a = \alpha_a + \beta(a - a_0)$. The age-specific rates from the cohort formulation increase steeper by age if $\beta > 0$, i.e. if rates are increasing by time.

Thus, the age-drift model is a submodel of both the general age-period model and of the general age-cohort model. In particular we note that when we have a constant annual change in rates it makes no sense to attribute this to either period or cohort. Whatever the "true" mechanism behind such a regular temporal variation of rates were, the observed rates would be the same.

The two sets of age-specific rates and the corresponding slopes by age-time are given in figure 3.11. We note that as we have put the reference period to 1968–72 (with midpoint 1970.5) and the reference cohort is 1933 (1928–1937), the two age-curves cross in $1970.5 - 1933 = 37.5$ years of age, that is, the two models have the same estimate for the rates in age-class 35–39.

3.4.1 Interpretation

The two sets of age-specific rates have different interpretations:

The estimates from the age-period-drift formulation are estimated cross-sctional rates for the period 1968–72, and the model predicts that rates for each year increase by a factor $\exp(0.0265)$, that is from each 5-year period to the next by a factor $\exp(0.0265 \times 5) = 1.14$ — 14% each 5 years.

The estimates from the age-cohort-drift formulation are estimated cohort or longitudinal rates for the (synthetic) birth cohort 1933. They are the predicted rates for the 1933 generation. The drift parameter predicts an increase of rates from each (one-year) generation to the next of

$\exp(0.0265)$, that is from each 5-year cohort to the next by a factor $\exp(0.0265 \times 5) = 1.14$ — 14% each 5 years.

Both interpretations are equally valid; the choice of parametrization must be based on context or additional data — separation is not possible on the basis of the data recorded in the Lexis diagram.

3.4.2 Praticalities in fitting the age-drift model

The age-drift model is fitted straightforward by defining the variable $p - p_0$ or $c - c_0$ and including this in the model as a continuous covariate.

In R this would look like:

```
C0 <- 1933
m.dr <- glm( D ~ factor(A) - 1 + I(C-C0) + offset( log( Y ) ), family=poisson )
summary( m.dr )
```

The age-specific log-rates (in the examples above, the longitudinal rates) are parameters in the model, so they can be used directly with their estimated standard errors to form confidence intervals for the log-rates. Confidence intervals for rates are then formed by transformation of these by the exponential function.

The log rate-ratio in a specific cohort c (relative to the reference c_0) is $\hat{\gamma} \times (c - c_0)$, and the (95%) confidence interval is $\big(\hat{\gamma} \pm 1.96 \times \text{s.e.}(\gamma)\big) \times (c - c_0)$. Hence, to form rate-ratios with confidence interval for rate-ratios over a whole sequence of cs we compute:

$$\exp\big(\hat{\gamma} \times (c - c_0)\big), \quad \exp\big(\hat{\gamma} \pm 1.96 \times \text{s.e.}(\gamma)) \times (c - c_0)\big)$$

These curves can then be plotted against c; preferably using a log-cale on the y-axis to show the log-linear structure of the model. This was done in figure 3.11

3.5 The age-period-cohort model

If we take the age-period model, which has $1 + (A - 1) + (P - 1)$ parameters, and add cohort as a factor (class variable), we will find that we get only $C - 2$ and not $C - 1$ new parameters as we would expect. Similarly if we start with the age-cohort model which has $1 + (A - 1) + (C - 1)$ parameters and add period as a factor we get only $P - 2$ and not $P - 1$ new parameters.

So when we put all three time scales age, period and cohort in the model as factors we get one parameter less than we would expect from a model with three factors. This comes from the common "drift" term that we found as the intersection between the age-period and the age-cohort model. Depending on how we fix the three levels of period and cohort that will be set to 0 we get dramatically differently looking estimates.

Four different examples are shown in figure 3.12. Each column represent one parametrization. But for estimates in one column the product of the age-effect, the period effect and the cohort effect gives exactly the same fitted rates. There is an infinity of ways we can produce sets of estimates that multiply to the set of fitted rates.

A mathematical illustration of this is as follows: Suppose we fitted a model with a certain set of parameters:

$$\log[\lambda(a, p)] = \alpha_a + \beta_p + \gamma_c \tag{3.1}$$

The reason that a new factor with F levels only produces $F - 1$ new parameters is that we have a constant, in the model, that is we can replace the formula in (3.2) by:

$$\log[\lambda(a, p)] = (\alpha_a + \mu_a) + (\beta_p + \mu_p) + (\gamma_c + \mu_c) \tag{3.2}$$

as long as $\mu_a + \mu_p + \mu_c = 0 \Leftrightarrow \mu_a = -\mu_p - \mu_c$. The special thing about the age-period cohort model is that for all observations we have that $a = p - c \Leftrightarrow a - p + c = 0$, so we can also replace (3.2) by:

$$\begin{aligned} \log[\lambda(a,p)] &= (\alpha_a + \delta a - \mu_p - \mu_c) + \\ &\quad (\beta_p - \delta p + \mu_p) + \\ &\quad (\gamma_c + \delta c + \mu_c) \end{aligned} \tag{3.3}$$

for any μ_p, μ_c and δ, and still get the same ste of predicted rates, i.e. the *model*[1].

The parametrizations produced by setting certain period and cohort effects to 0 correspond to choosing values of the three arbitrary parameters μ_p, μ_c and δ.

Mathematically we can only identify the 2nd order differences, i.e.

$$(\alpha_1 - \alpha_2) - (\alpha_2 - \alpha_3) = \alpha_1 - 2\alpha_2 + \alpha_3, \quad \alpha_2 - 2\alpha_3 + \alpha_4, \quad \alpha_3 - 2\alpha_4 + \alpha_5, \ldots$$

(assuming that classes are numbered 1,2,...). This is because the constant μ_p, μ_c and any linear trend will cancel when we form second order differences. This argument is only true if age-classes are equidistant, otherwise the linear term will not vanish. The second order differences of the parameters are the same no matter which one of the possible parametrizations one chooses. This goes for the period and cohort effects as well.

But such second order differences are very difficult to comprehend, in particular because they do not enter as additive components in expressions for log-rates, so they are not attractive as quantities to report.

In most studies published one will see the log rate ratio for a central period and the outer cohorts constrained to 0. But as can be seen from the displays of the results for the testis cancer data this is not necessarily giving a description of the rates that is unique in any clearly defined sense.

Holford [10] suggested to first fit a model with any parametrization of the effects, and then regress the age-estimates on age, the period estimates on period and the cohort estimates on cohort. The resulting residuals would then be a reasonable representation of the deviation of the effects from linearity and they would be the same regardless of the initial parametrization. However the estimated slopes by age, period and cohort cannot be uniquely assigned to any of the three time scales.

Thus there is no way to get a reasonably unique way of displaying three effects that sum to the fitted rates, unless one is willing to make some assumptions of the relative importance of the effects.

Suppose that we follow the line of Holford and regress the age-parameter on age etc., so that we have:

$$\lambda(a,p) = \tilde{\alpha}_a + \hat{\mu}_a + \hat{\delta}_a a + \tilde{\beta}_p + \hat{\mu}_p + \hat{\delta}_p p + \tilde{\gamma}_c + \hat{\mu}_c + \hat{\delta}_c c$$

where the $\tilde{}$-parameters are the residuals and $(\hat{\mu}, \hat{\delta})$ are the regression parameters. The residuals are on average 0 across age, period and cohort respectively. Holford suggest to report the three sets of residuals as a convenient way of showing the curvature, but the time trends remains difficult to report.

[1] We use the term model for a description of the data by a set of fitted values, whereas Clayton & Schiffers [3, 4] use the term for a certain parametrization of a model.

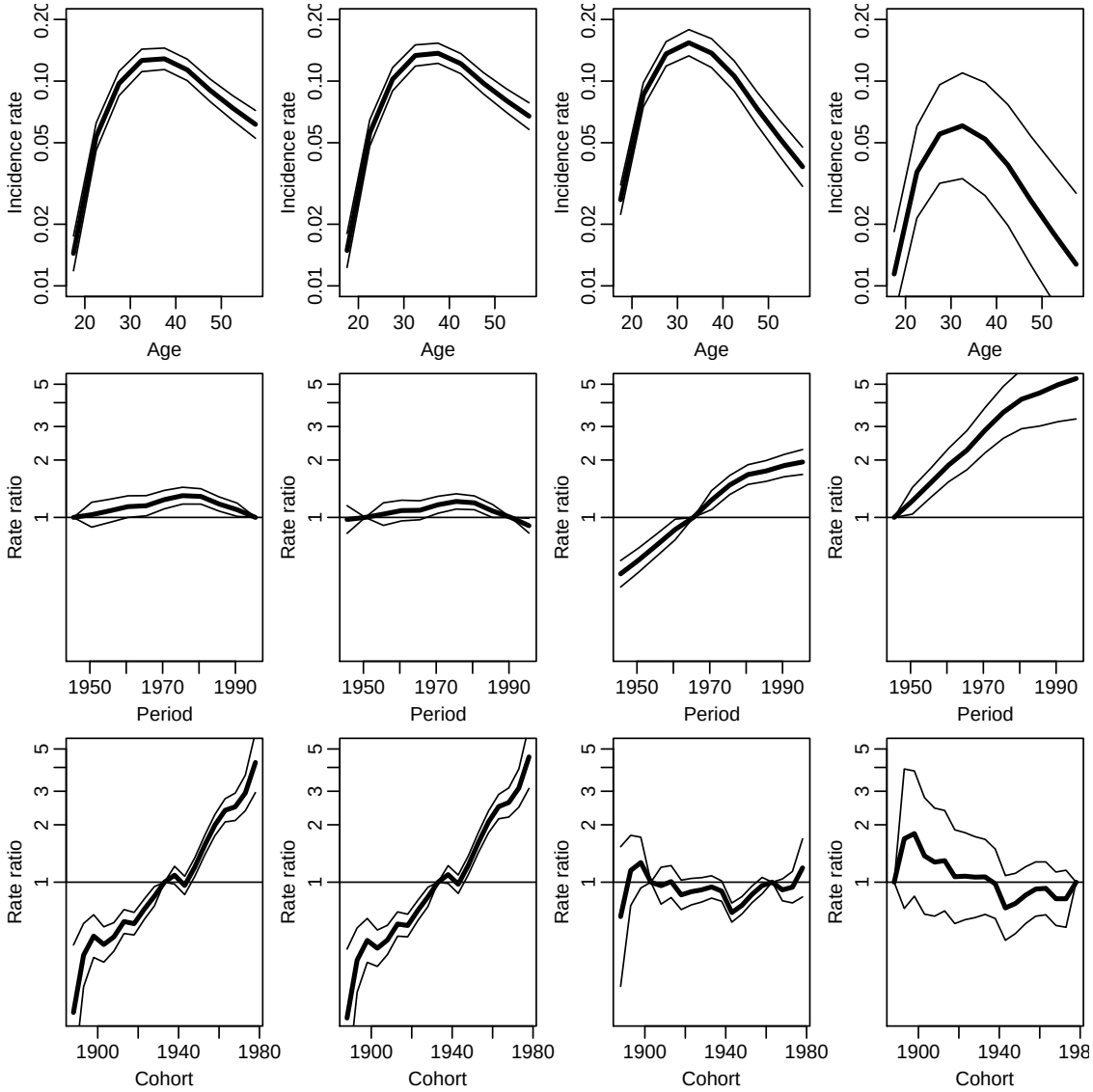


Figure 3.12: *Estimates from the age-period-cohort model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. Panels in the same column represent one parametrization. The thick lines connect the estimates, and the thin lines the 95% pointwise confidence limits.*

3.5.1 A simple suggestion for a parametrization

If we decide that age is the most important scale, and period the least important, we could choose a parametrization based on the following assumptions:

- Periods effect should be 0 on average.
- Cohort effect should be relative risk relative to some central cohort.
- Age effect should represent age-specific rates in a reference cohort after correction for period effects that are 0 on average. That is some average incidence rate in the cohort.

Starting from the top we could use

$$g(p) = \tilde{\beta}_p = \beta_p - \hat{\mu}_p - \hat{\delta}_p p$$

as period effects. Next, the cohort effect should then absorb all the time trend. But since there are two such terms, one for cohort ($\delta_c c$) and one for period ($\delta_p p = \delta_p(c+a)$), the linear cohort effect should be $\delta_c + \delta_p$. In order to make c_0 the reference cohort we must subtract the level in that cohort and place that with the age-effect, so the cohort effect will be:

$$h(c) = \gamma_c - \gamma_{c_0} + \hat{\delta}_p(c - c_0)$$

this is indeed 0 for c_0 and has the right slope. The remaining will be the age-effect:

$$f(a) = \alpha_a + \hat{\mu}_p + \hat{\delta}_p a + \hat{\delta}_p c_0 + \gamma_{c_0}$$

It is easily verified that $\alpha_a + \beta_p + \gamma_c = f(a) + g(p) + h(c)$, so we have just made a reparametrization of the model by partitioning the effects in a unique and well defined way between the three factors, but it relies on the arbitrary decisions that:

- Age is the major time scale.
- Cohort is the secondary time scale (the major secular trend).
- Period is the residual time scale.
- c_0 is the relevant reference cohort.

Such assumptions may be reasonable from a biological point of view or irrelevant from another point of view. The role of period and cohort could equally well have been interchanged. The important thing to realize is that there is no way that data can tell whether it is reasonable or not.

There is however one problem with this way of parametrizing the model. The regression on period is a complicated functions of the original data so the suggested parameters have virtually intractable standard errors, so there is no practical way to get standard errors of these. This is probably the main reason that this kind of approach has gained limited popularity (to put it mildly).

3.5.2 A practical suggestion

If however one is willing to prioritise as indicated above among the time scales and rank period lowest, one approach could be to fit an age-cohort model with c_0 as reference, and subsequently fit a period model to the residuals.

The procedure would then be first to fit the age-cohort model, with cohort c_0 as reference and get estimates $\hat{\alpha}_a$ and $\hat{\gamma}_c$. If the "true" model is an age-period-cohort model, we still need to accommodate the period effect. If we settle for the residual period effect conditional on the already estimated age and cohort effects we have:

$$\log[\lambda(a,p)] = \hat{\alpha}_a + \hat{\gamma}_c + \beta_p$$

The residual period effect can be estimated if we note that for the number of cases we have:

$$\log(\text{expected cases}) = \log[\lambda(a,p)Y] = \hat{\alpha}_a + \hat{\gamma}_c + \log(Y) + \beta_p$$

This is analogous to the expression for a Poisson model in general, but now is the offset not just $\log(Y)$ but $\hat{\alpha}_a + \hat{\gamma}_c + \log(Y)$, the log of the fitted values from the age-cohort model. The β_ps can the be estimated in a Poisson model with these fitted values as offset. The procedure can of course be applied with the roles of cohort and period reversed.

The procedure does not give maximum likelihood estimates, but *marginal* estimates of age and cohort effects and *conditional* estimates of the period effects.

The advantage of this procedure is that confidence intervals can be computed for all parameters very easily. But they are of course not maximum likelihood estimates. The standard errors from the software will be marginal standard errors for age and cohort and conditional for period.

The results from this simple stepwise procedure and from using Holford's suggestion and participating the maximum likelihood estimates are shown in figure 3.13 for the testis cancer data, and the results are virtually indistinguishable.

But again: The choice is essentially arbitrary, and only represents one way of describing data.

In summary: There are three different ways of arriving at a parametrization of an age-period-cohort model in terms of three sets of parameters that sum to the expected log rates:

1. Constrain 2 period and 1 cohort parameter to be 0 (or vice versa).

2. Holford's residual approach. Optionally extended by putting the overall trend into the cohort, and all the rest of the effects into age.

3. Fit an age-cohort model, and subsequently a period-alone model using the log-fitted values from the age-cohort model as offset. Report the estimates from these two models with confidence bands.

The latter two approaches are contrasted in figure 3.13, where a separate panel is used for each effect. Even if we have made all the vertical scales identical in extent (25 times from lowest to highest), the horizontal scales are different, which makes it difficult to judge the relative effects of the factors. Therefore it is advisable to plot all effects with the same scale on the horizontal axis as well. This also gives the possibility of using the same display for cohort and period effects as they have a small overlap of calendar time they refer to. This is illustrated in figure 3.14

3.5.3 Practicalities in fitting the age-period-cohort model

All the practical problems in connection with the fitting of the age-period-cohort model concern the fiddling with the parametrizations.

Constraints on periods and cohorts

If we want to constrain one period and two cohorts or two period and one cohort to be 0, the trick is to get the software to do this by proper coding of the sequence of factor levels for period and cohort.

In R (assuming the default `contr.treatment` is used for factor coding), the *first* level of each factor is used as reference. If any of the later indicators of factor levels are found to be linearly related to the previous columns of the design matrix, the coefficients of these are set to 0. Thus, if we want to alias two period parameters and one cohort parameter then:

- Make the reference level of the cohort the first, using `relevel`.

- Make one of the reference periods the first and the other the last level.

- Make sure that cohort is entered before period in the model formula.

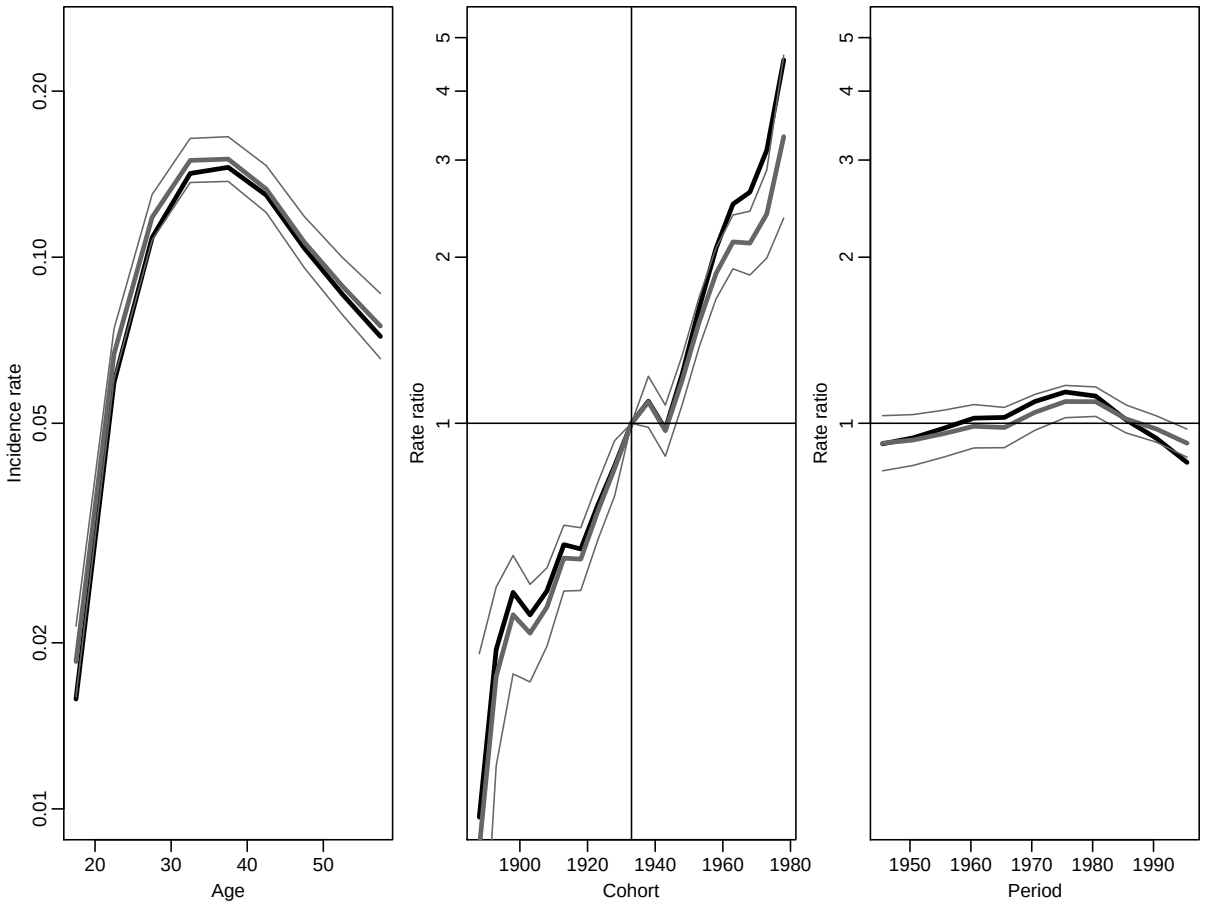


Figure 3.13: *Estimates from the age-period-cohort model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The black lines are maximum likelihood estimates, scaled so that the regression of log period effects on period is 0, and so that the cohort effect in 1933 is 1. The gray lines are those obtained by fitting first an age-cohort model and subsequently a period model to the residuals.*

Suppose we have 9 age classes, 9 periods and therefore 17 (synthetic) cohorts, and that we would have cohort 9 and periods 2 and 8 as re fences the code for this would look like:

```
m.apc <- glm( D ~ factor( A ) - 1 +
                  relevel( factor( C ), 9 ) +
                  relevel( factor( P ), c(2,1,3:7,9,8) ) +
                  offset( log( Y ) ), family=poisson )
```

The function `relevel` takes the mentioned levels of the factor as the first ones, and then leaves the rest in the original order. It is essential that the period term is after the cohort term in the model.

Note that we include the term "`-1`" in the model in order to prevent an intercept to be fitted. In this way we get the age-effects as incidence rates. In this case as cohort rate for cohort 9, after corrections for period.

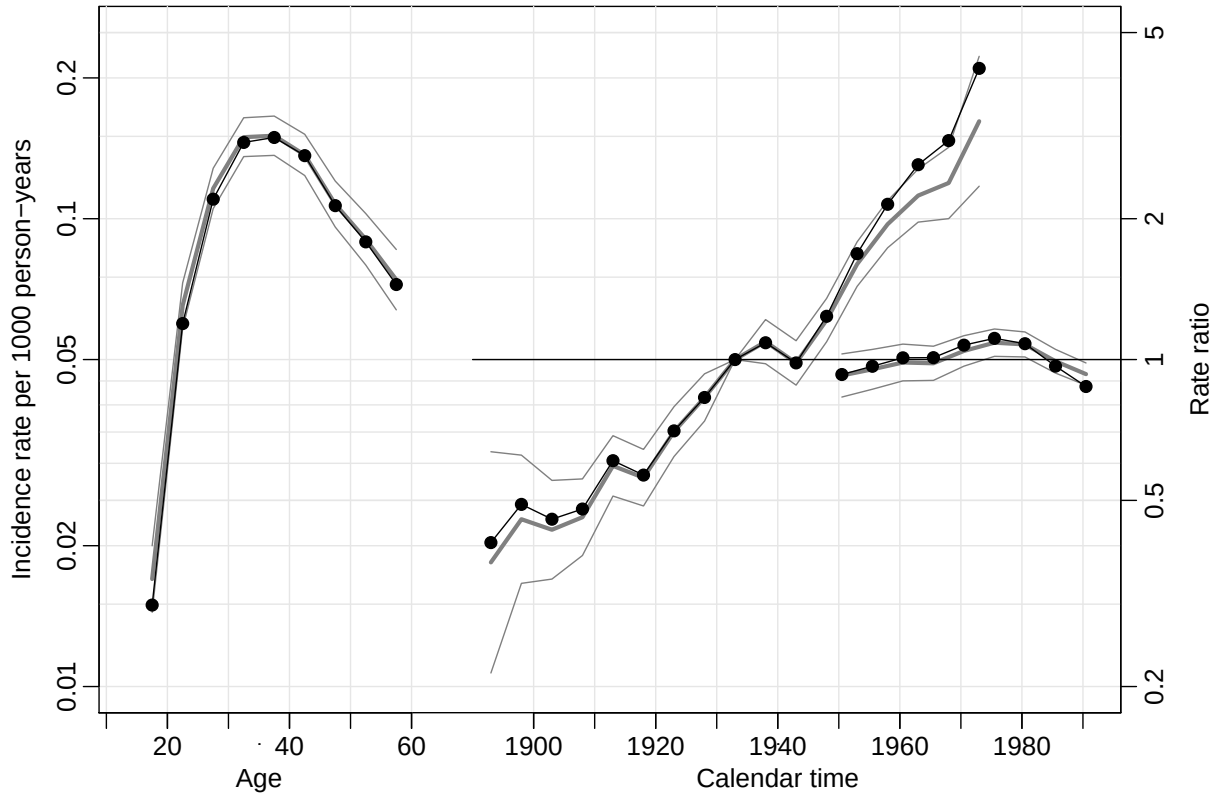


Figure 3.14: *Estimates from the age-period-cohort model fitted to the Danish testis cancer data shown in tables 3.1 and 3.2. The black lines with dots are maximum likelihood estimates, scaled so that the regression of log period effects on period is 0, and so that the cohort effect in 1933 is 1. The gray lines are those obtained by fitting first an age-cohort model and subsequently a period model to the residuals. Note the common scaling of both axes.*

Holford's suggestion

This involves extraction of the parameters belonging to each of the three factors, insertion of whatever 0s the program has left out or fixed, regression of each set of parameters on the factor levels, and finally adjustment of the parameters to the desired form.

Examples of how to do this can be found in `xtestis-apc.R` .

The sequential method

This is about fitting two models:

1. The age-cohort model to the original data.

2. A period-only model to the residuals (i.e. using fitted values as offset).

In order to make the age-parameters (log) rate-parameters and c_0 the reference cohort we need a model fit as:

```
m.ac <- glm( D ~ fcator( A ) - 1 + relevel( factor( C ), "c0" ) + offset( log( Y ) ),
             family=poisson )
```

This will allow us to extract the age-parameters with standard errors from the model by (assuming we have 9 age-classes):

```
a.par <- summary( m.ac )$coef[1:9,1:2]
```

and the log-rate-ratios by cohort as:

```
c.par <- summary( m.ac )$coef[10:25,1:2]
```

These can then be plotted against age and cohort respectively, with confidence intervals constructed on the parameter scale and then transformed to the rate, respectively rate-ratio scale by the exponential function.

The period residuals can now be estimated using the fitted or predicted values of the linear predictor as offset:

```
m.r.p <- glm( D ~ factor( P ) - 1 + offset( predict( m.ac, type="link" ) ),
              family=poisson )
```

This model is then used to produce the period residuals with standard errors:

```
p.res <- summary( m.r.p )$coef[,1:2]
```

and these can now be plotted against period.

An example of how to do this can also be found in `xtestis-apc.R`.

Chapter 4

Cohort studies, SMR and RSR

4.1 Format of cohort studies

Epidemiological cohort studies are characterised by long follow-up times, as opposed to short term survival studies. This implies that multiple time scales may be of interest; notably current age (attained age) and calendar time on top of time since entry.

Thus in cohort studies the follow-up time (and events) from each subject must be subdivided by several time-scales: age, calendar time and possibly time since entry into the study, time since first exposure, cumulative exposure etc.

The minimal data required to represent a follow-up study is:

- Date of entry into the study.
- Date of exit from the study.
- Exit status (event or censoring).

This will enable modelling of event-rates by time since entry and calendar time. If date of birth is known too, current age (and date of birth) can be used as covariates as well.

We may think of this as the same exercise is in the construction of the likelihood in chapter 2, 2.1, where each little piece of follow-up (day, say) is classified by current date, current age and time since entry into the study. This time however it is performed in practise and not only for the purpose of deriving a likelihood.

The practical approach to this is accomplished by splitting the follow-up time along each of the time-scales we want to use as covariates. The classical illustration of this is splitting the follow-up by current age and calendar time, corresponding to crossings of the grids in the Lexis diagram. If additional splitting by time since entry into the study is required we get further subdivisions of the follow-up time, as seen in figure 4.1.

The splitting of follow-up time is a data-manipulation task that replaces each record in a file of follow-up information with entry, exit and failure with a number of records with the same variables. The exit date of a new record is identical to the entry date of the next new, and failure status is censored for all records except the last from the same individual which has failure status as the original record.

In the process of splitting follow-up time this way each new record will get current values for each of the time scales that in play. In the example shown in figure 4.1, we will go from the single record:

```
id    birth    entry     exit   fail
 1   1930.2   1971.4   1978.7      1
```

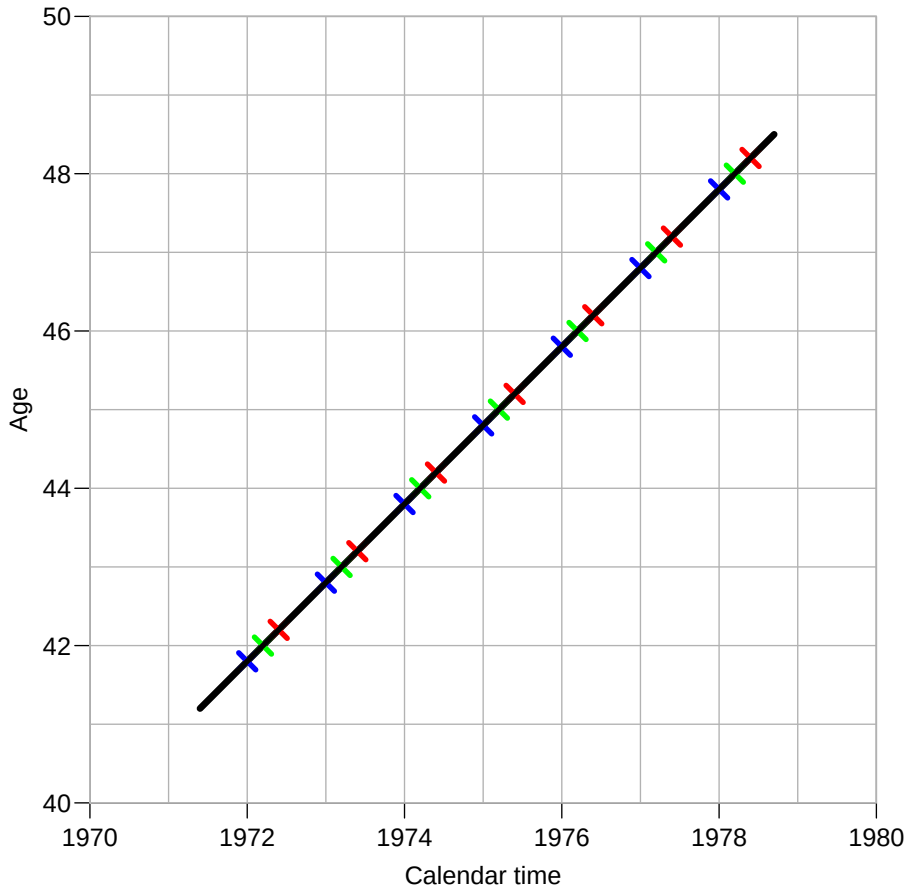


Figure 4.1: *Splitting of time by age, period and time since entry into the study.*

whereas when we split along the age-axis, the calendar time axis and along the time-since entry-axis we get the following 22 records:

	Entry	Exit	Fail	age	per	yfe	Risk
1	1971.4	1972.0	0	41	1971	0	0.6000055
2	1972.0	1972.2	0	41	1972	0	0.2000009
3	1972.2	1972.4	0	42	1972	0	0.1999936
4	1972.4	1973.0	0	42	1972	1	0.6000055
5	1973.0	1973.2	0	42	1973	1	0.2000009
6	1973.2	1973.4	0	43	1973	1	0.1999936
7	1973.4	1974.0	0	43	1973	2	0.6000055
8	1974.0	1974.2	0	43	1974	2	0.2000009
9	1974.2	1974.4	0	44	1974	2	0.1999936
10	1974.4	1975.0	0	44	1974	3	0.6000055
11	1975.0	1975.2	0	44	1975	3	0.2000009
12	1975.2	1975.4	0	45	1975	3	0.1999936
13	1975.4	1976.0	0	45	1975	4	0.6000055
14	1976.0	1976.2	0	45	1976	4	0.2000009
15	1976.2	1976.4	0	46	1976	4	0.1999936
16	1976.4	1977.0	0	46	1976	5	0.6000055
17	1977.0	1977.2	0	46	1977	5	0.2000009
18	1977.2	1977.4	0	47	1977	5	0.1999936
19	1977.4	1978.0	0	47	1977	6	0.6000055
20	1978.0	1978.2	0	47	1978	6	0.2000009
21	1978.2	1978.4	0	48	1978	6	0.1999936
22	1978.4	1978.7	1	48	1978	7	0.2999980

Three new variables have been added: `age` (current age), `per` (calendar year) and `yfe` (year

since entry). These variables are each constant over a period of 1 year; they are taken as the value at the left end-point of the current one-year interval on each of the three scales.

This highlights the somewhat paradoxical legacy of the Lexis-diagram: Each piece of follow-up is assigned a certain age-class and not the actual current age. Furthermore the splitting along several time-axes gives interval of differing lengths, in the example above of lengths 0.2, 0.2 and 0.6 years.

One might argue that we instead, given a decision to have intervals of on average 0.33 years, split the entire follow-up time in intervals of equal length, for example as time since entry, and subsequently computed the current age and calendar time at the start of each of these intervals. These values could then later be rounded when used for modelling and matching with population rates. This would give a split as:

```
      Entry       Exit   Fail         yfe        age        per        Risk
1  1971.400   1971.733      0   0.0000000   41.20000   1971.400   0.3333333
2  1971.733   1972.067      0   0.3333333   41.53333   1971.733   0.3333333
3  1972.067   1972.400      0   0.6666667   41.86667   1972.067   0.3333333
4  1972.400   1972.733      0   1.0000000   42.20000   1972.400   0.3333333
5  1972.733   1973.067      0   1.3333333   42.53333   1972.733   0.3333333
6  1973.067   1973.400      0   1.6666667   42.86667   1973.067   0.3333333
7  1973.400   1973.733      0   2.0000000   43.20000   1973.400   0.3333333
8  1973.733   1974.067      0   2.3333333   43.53333   1973.733   0.3333333
9  1974.067   1974.400      0   2.6666667   43.86667   1974.067   0.3333333
10 1974.400   1974.733      0   3.0000000   44.20000   1974.400   0.3333333
11 1974.733   1975.067      0   3.3333333   44.53333   1974.733   0.3333333
12 1975.067   1975.400      0   3.6666667   44.86667   1975.067   0.3333333
13 1975.400   1975.733      0   4.0000000   45.20000   1975.400   0.3333333
14 1975.733   1976.067      0   4.3333333   45.53333   1975.733   0.3333333
15 1976.067   1976.400      0   4.6666667   45.86667   1976.067   0.3333333
16 1976.400   1976.733      0   5.0000000   46.20000   1976.400   0.3333333
17 1976.733   1977.067      0   5.3333333   46.53333   1976.733   0.3333333
18 1977.067   1977.400      0   5.6666667   46.86667   1977.067   0.3333333
19 1977.400   1977.733      0   6.0000000   47.20000   1977.400   0.3333333
20 1977.733   1978.067      0   6.3333333   47.53333   1977.733   0.3333333
21 1978.067   1978.400      0   6.6666667   47.86667   1978.067   0.3333333
22 1978.400   1978.700      1   7.0000000   48.20000   1978.400   0.2999946
```

If the intervals are sufficiently small the two approaches are not likely to produce very different results. However if age- and period-intervals are 5 years and time since entry intervals are 1-year intervals, one might very well see differences.

The first method where follow-up time is rigidly assigned to specific classes on each time-scale is a descendant from SMR-analyses where risk time is meticulously assigned age×period regions of the Lexis diagram where population rates are known.

In contexts where parametric models for the effect of several time scales are included it would however be more appropriate to use the second approach with accurate computation of the time scales at the beginning of each follow-up interval.

4.2 Comparing cohort rates and population rates

For cohort studies with follow-up time over long periods and where persons in the study are followed over different calendar period it is desirable to control for possible confounding by the population event rates by modelling the relationship between the observed rates and the population rates.

There are two major approaches to this described in the literature:

- SMR-modelling where the ratio of rates between the cohort and the population are modelled as function of covariates. This is a generalisation of the classical SMR.

- Excess risk modelling where the difference between cohort and population rates are modelled. This is a variant of the relative survival calculations known from cancer survival studies.

Both approaches will be briefly reviewed below, but only the SMR approach will be further pursued by an example.

The data requirements for these approaches are access to population rates of the event type of interest. This will in practise limit the applicability of these types of analysis to mortality studies and cancer incidence studies.

4.2.1 SMR or the relative risk model

This approach is based on a multiplicative model for the cohort rates:

$$\lambda_c = \lambda_{\text{pop}} \times \text{RR}$$

where $\text{RR} = \exp(X\beta)$ is a term describing the effect of the covariates on the ratio of rates. The contribution to the log-likelihood from D events during Y follow-up time will be (see (2.1)):

$$\begin{aligned} \ell(\lambda_c|(D,Y)) &= D\ln(\lambda_c) - \lambda_c Y \\ &= D\ln(\lambda_{\text{pop}}\text{RR}) - \lambda_{\text{pop}}\text{RR}Y \\ &= D\ln(\text{RR}) - E\text{RR} + D\ln(\lambda_{\text{pop}}) \end{aligned}$$

where $E = \lambda_{\text{pop}}Y$ is the expected number of events under the assumption of fixed follow up time and population event rate. The last term, $D\ln(\lambda_{\text{pop}})$ does not depend on the parameters of interest in RR, so the log-likelihood for RR is exactly as the log-likelihood for a rate, just with follow-up time replaced by expected numbers.

In this derivation we assumed that population rates λ_{pop} were known and constant for each follow-up interval for cohort members. This is in most practical circumstances taken as being constant in 5 year age-classes by 5 year periods because these numbers are readily available from population statistics. Most likely because they will fit on one page in a statistics yearbook. Thus most practical analyses of SMR are based on initial splitting of follow-up time by 5-year age-classes and periods, and keeping the assignment of these classes through any further splitting of follow-up time. The latter has to some extent been carried over to analysis of cohort studies where no background rates are incorporated.

4.2.2 RSR or the excess risk model

RSR is an acronym for Relative Survival Rate which was introduced by Ederer, Axtell and Cutler [6], as a way to correct cancer patient survival for general population survival. The basic idea is to take the 5-year survival say, for a group of cancer patients and divide by the expected survival for a group of persons of the same age- and sex- composition as the cancer patients, worked out assuming that population mortality rates prevail.

If the cancer patient survival rates are $\lambda_c(t)$ and the population rates are $\lambda_{\text{pop}}(t)$ the 5-year RSR is:

$$\text{RSR}(5) = \exp\left(-\int_0^5 \lambda_c(s)\,\mathrm{d}s\right) \Big/ \exp\left(-\int_0^5 \lambda_{\text{pop}}(s)\,\mathrm{d}s\right) = \exp\left(-\int_0^5 \lambda_c(s) - \lambda_{\text{pop}}(s)\,\mathrm{d}s\right)$$

Thus, the RSR is a function of the rate difference between the cohort mortality and the population mortality. The above formula is slightly deceptive, because the technicalities arising from patients being of different ages and sex are not accounted for.

The generalisation including covariates of the RSR approach is based on an additive model for the cohort rates:

$$\lambda_c = \lambda_{\text{pop}} + \text{ER}$$

where $\text{ER} = X\beta$ is a term describing the effect of the covariates on the difference between rates[1]. The contribution to the log-likelihood from D events during Y follow-up time will be (again, see (2.1)):

$$\begin{aligned} \ell(\lambda_c|(D,Y)) &= D\ln(\lambda_c) - \lambda_c Y \\ &= D\ln(\lambda_{\text{pop}} + \text{ER}) - (\lambda_{\text{pop}} + \text{ER})Y \end{aligned}$$

The log-likelihood in this case is for a Poisson-variate with mean $(\lambda_{\text{pop}} + \text{ER})Y$, i.e. a Poisson model with identity link function, $\lambda_{\text{pop}}Y$ as offset and covariates equal to the original covariates multiplied by Y. Here it is obvious that if follow-up time is split in equally long intervals the multiplication of the covariates by Y disappears, and the effect only lies in the interpretation of regression coefficients.

Most modern computer packages have facilities for fitting these models, although they will usually require some coding. A more detailed exposition of these models is given by Dickman *et al.* [5].

4.2.3 Confounding by population rates

To the extent that the modelling of rates incorporates population rates, it may be viewed as a way to control confounding by these. The argument here is that the effect of covariate of interest may be biased if the covariate is associated with the population rates, i.e. in some way varies with the determinants of these. The classical logic of confounder control in regression analysis would then imply that the population rates be entered in the analysis as a covariate and not as an offset. Thus in the SMR-analyses we should enter the log of the population rate as covariate, and likewise for the excess risk model.

In the multiplicative case we would get a model of the form:

$$\lambda_c = \lambda_{\text{pop}}^{\theta} \times \text{RR}$$

which (if $\theta \neq 1$) would lead to an interpretation of RR as the covariate effect controlled for population mortality, but not as a SMR.

In the additive case we would have:

$$\lambda_c = \xi\lambda_{\text{pop}} + \text{ER}$$

with a similar vague interpretation.

Thus, the offset formulation is founded on interpretability and only partly on the need for confounder control.

[1] There is also a large number of models around describing ER as a multiplicative function of the covariates. We shall not go into details with that here.

Chapter 5

Classification by age, period and cohort in the Lexis diagram

Analysis of rates from a complete observation in a Lexis digram need not be restricted to the classical sets classified by two factors:

A-sets: Classification by age and period. (□)

B-sets: Classification by cohort and period. (◰)

C-sets: Classification by age and cohort. (▱)

Analyses of incidence data from registries is commonly based on tabulation of cases by age and period. The corresponding person-years are normally derived from census data either by averaging presence data at ends of each period or by using mid-period estimates of population size.

Additional classification of cases by date of birth (cohort) is a subdivision of the A sets (□) into upper (◸) and lower (◿) triangles and is a simple tabulation exercise. Calculagtion of thee risk time in these triangles reqires some care.

5.1 Risk time calculations

The following is based on material from lecture notes by Sverdrup [28]. To our knowledge this has not been published elsewhere, despite its obvious relevance in descriptive epidemiology. The paper by Hoem [8] and the correction note [9] has a reference to a similar result from an earlier version of Sverdrup's notes.

5.1.1 Census data

First, consider for the sake of simplicity the division of the Lexis-diagram in 1-year classes by age, calendar time and date of birth, and suppose that population figures are available in 1-year classes each year, as will be the situation for most areas where regular censuses are done. The situation is illustrated in figure 5.1. The target is to construct estimates of population risk time for each of the areas **A** and **B**.

In the following we let a refer to age, p to calendar time (period), and c to date of birth (cohort), and we let $\ell_{a,p}$ represent the population size in age a at the beginning of the year p.

If no deaths or migrations occurred in the population, we would have that $\ell_{a,p} = \ell_{a+1,p+1}$.

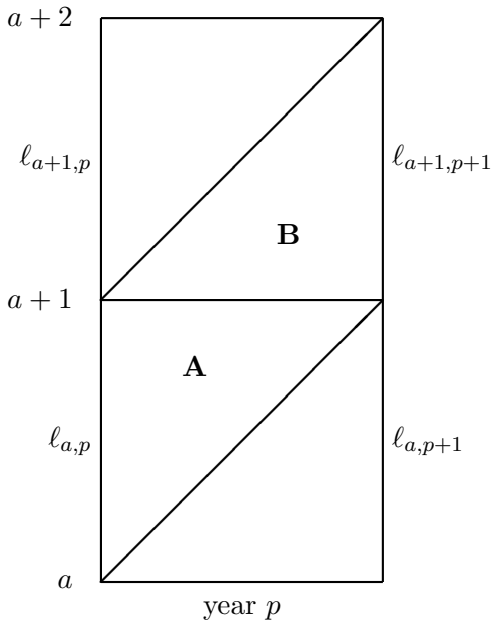


Figure 5.1: *Section of a Lexis diagram showing basic triangles and population figures (the ℓs) assumed known.*

In presence of mortality[1] we can at least infer that the survivors $\ell_{a+1,p+1}$ have been at risk throughout the year p. Assuming that the persons are uniformly distributed within the age-classes, the average risk time contribution of a survivor will be $\frac{1}{2}$ year to each of the triangles **A** and **B**.

In order to work out the contribution of risk time of those dying during the year p, we assume that the deaths are uniformly distributed over **A** and **B**[2]. This means that the total amount of risk time contributed to **A** and **B** by those dying in **A** and **B** is $(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{2}$y.

Those who die in **A** contribute no risk time to **B**. In **A** their average contribution can be computed by integration over the triangle **A**. The mean contribution must be calculated as an average w.r.t. to the uniform measure on **A**. The area of **A** is $\frac{1}{2}(= \int_{p=0}^{p=1} \int_{a=p}^{a=1} 1 \, da \, dp)$, so the density of the uniform measure is 2.

For simplicity of notation it is assumed that age and date range from 0 to 1 in all of the calculations below.

A person dying in age a at date p in **A** contributes p risk time, so the average will be:

$$\int_{p=0}^{p=1} \int_{a=p}^{a=1} 2p \, da \, dp = \int_{p=0}^{p=1} 2p(1-p) \, dp = \left[p^2 - \frac{2}{3}p^3 \right]_{p=0}^{p=1} = \frac{1}{3}$$

Those who die in **B** contribute risk time in both **A** and **B**. If death occurs in age a at date p the person has contributed $p - a$ person-years in **A** and a person-years in **B**.

[1] Immigration and emigration can be treated as negative and positive mortality respectively, and does not alter the results derived here, provided the assumptions made for the mortality pattern also holds for the migration patterns in the population.

[2] Note that this may a unrealistic assumption for age-classes of length 5 years or more.

Hence the average amount contributed in **A** is:

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2(p-a)\,da\,dp = \int_{p=0}^{p=1}\left[2pa-a^2\right]_{a=0}^{a=p} dp = \int_{p=0}^{p=1} p^2\,dp = \frac{1}{3}$$

and in **B**:

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2a\,da\,dp = \int_{p=0}^{p=1} p^2\,dp = \frac{1}{3}$$

Collecting these observations gives the following risk time in **A** and **B**:

	A:	**B**:
Survivors:	$\ell_{a+1,p+1} \times \frac{1}{2}$y	$\ell_{a+1,p+1} \times \frac{1}{2}$y
Dead in **A**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	
Dead in **B**:	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y	$\frac{1}{2}(\ell_{a,p} - \ell_{a+1,p+1}) \times \frac{1}{3}$y
$\sum$	$(\frac{1}{3}\ell_{a,p} + \frac{1}{6}\ell_{a+1,p+1}) \times 1$y	$(\frac{1}{6}\ell_{a,p} + \frac{1}{3}\ell_{a+1,p+1}) \times 1$y

The risk among 0-year olds in year p, born in year p can be computed by requiring that the total risk time among 0-years olds in year p should equal 1y$\times$ the average of the population sizes in age 0 at the beginning and end of year p, i.e.:

$$\frac{1}{2}(\ell_{0,p} + \ell_{0,p+1}) \times 1\text{y} - (\frac{1}{3}\ell_{0,p} + \frac{1}{6}\ell_{1,p+1}) \times 1\text{y} = (\frac{1}{6}\ell_{0,p} + \frac{1}{2}\ell_{0,p+1} - \frac{1}{6}\ell_{1,p+1}) \times 1\text{y}$$

A similar procedure can be applied in the last non-open age-class (usually 89). It has little meaning to try to subdivide open age-classes by date of birth.

5.1.2 Estimation of population size in one-year age classes from 5-year classified data and annual birth data.

For some populations the only available breakdown of population data is in 5-year age-classes. If annual figures of births are available, then the population in 5-year age-classes can be proportionally distributed in one-year age-classes according to the number of births in each of the relevant birth-years.

Consider the Lexis diagram in figure 5.2

The computations could be along two different lines:

- Use only the presence data and the birth cohort sizes, and neglect mortality. Use the presence data (in 5-year classes) to compute person-years at risk in 5 by 5 by 5 year triangles. Within each triangle the 5 birth cohorts contribute observation in 9, 7, 5, 3 and 1 small triangles, respectively. If there were no mortality the person-years in each small triangle could be computed on the basis of the birth figures alone. We then adjust the person-years computed this way for all 25 small with the same factor so that the person-years fits for the large 5 by 5 by 5 year traingle.

  This is a very simple procedure but it has the effect that the number of person years in each small traingle within each large one will be the same in the same cohort.

- The basis for deriving the formula for risk time in triangles based on presence data was an assumption of the number of deaths (and immigrations) being evenly distributed in each of the B-sets (classified by period and cohort).

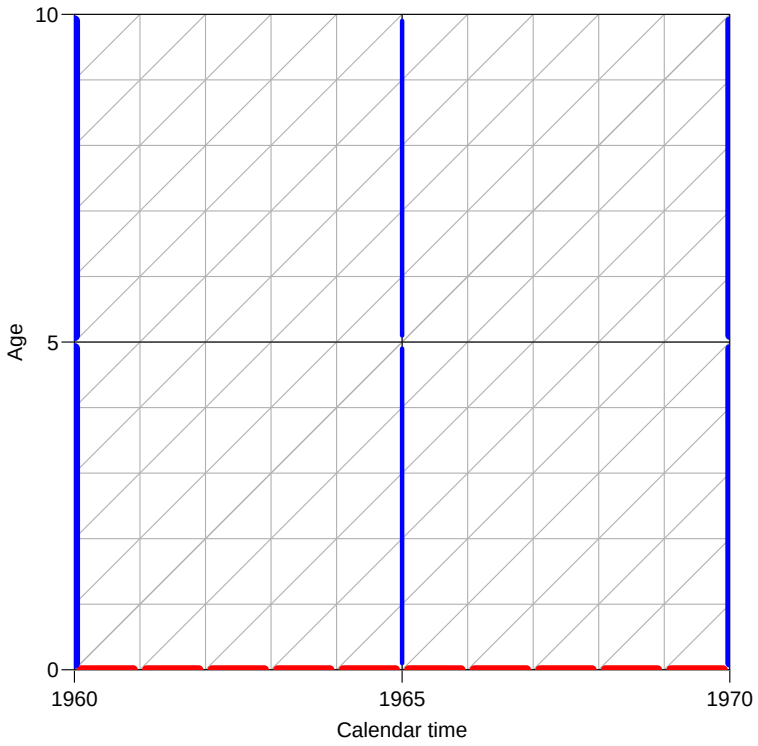


Figure 5.2: *Illustration of commonly available presence data (vertical, thick blue lines) and birth cohort (horizontal, thick red lines).*

If we have both birth data for a 5-year period and the presence data for the 0–5 year olds at the end of the period, we can compute the number of deaths in the triangle. Distributing these evenly over the small triangles would then enable computation of the person-years in each. A straightforward refinement would be to distribute the deaths evenly in triangles for each cohort in such a way that the number of deaths in the first lower triangle for each birth cohort were proportional to the size of the birth cohort.

In these calculations we must use the derivations from section 5.1.1 where each dead person in a triangle on average contributes 1/3 person-year, and each survivor 1/2 person-year.

This will give person-years in the 25 sub-triangles of the first lower triangles (0–5 years) as well as estimates of the population size in each of the single-year age-classes. The total number of deaths in the following B-set can now be distributed uniformly in all of the 10 triangles for each birth cohort, and person-years computed.

This can be further refined by assuming the mortality to be higher of 0-year olds and hence to allocate a somewhat larger proportion of the deaths to this age-class. The general version of this would be to use some (known) mortality function (of age) and then allocate the known number of deaths for a 5 by 5-year B-set (or first year lower triangle) as proportional to the size of the birth cohort times the mortality rates at the average age for the small triangle. Note that if we assume that mortality for men and women are proportional we will only need one function in the computations, as the absolute size of the function is irrelevant.

This approach assumes that migrations in and out of the population is not too concentrated

in specific birth-cohorts and age-classes.

5.2 Means for subsets of the Lexis diagram

5.2.1 The mean age, period and birth-cohort in triangular subsets of the Lexis diagram

When modelling the effect of age and date of event and date of birth for data classified by age, period and cohort, it is necessary to assign a value of these variables to each of the resulting triangles in the Lexis diagram. This is actually necessary for any sub-classification of the Lexis diagram used as basis for modelling rates. These mean values must obey $a = p - c$.

For the specific subdivision of the Lexis diagram into triangles these means have been quoted in at least two papers in the literature [29, 22], but none has given the mathematical derivation below.

Referring to figure 5.1, we can compute the average age and date of event (or rather exposure) in the triangles **A** and **B**. Again we use the convention that $a = 0$ and $p = 0$ in the lower left corner of the triangle, regardless of whether we consider **A** or **B**:

$$\begin{aligned}
\mathrm{E}_{\mathbf{A}}(a) &= \int_{p=0}^{p=1}\int_{a=p}^{a=1} 2a\,da\,dp = \int_{p=0}^{p=1} 1-p^2\,dp &&= \tfrac{2}{3}\\
\mathrm{E}_{\mathbf{A}}(p) &= \int_{a=0}^{a=1}\int_{p=0}^{p=a} 2p\,dp\,da = \int_{a=0}^{a=1} a^2\,dp &&= \tfrac{1}{3}\\
\mathrm{E}_{\mathbf{A}}(c) &= \tfrac{1}{3}-\tfrac{2}{3} &&= -\tfrac{1}{3}\\
\mathrm{E}_{\mathbf{B}}(a) &= \int_{p=0}^{p=1}\int_{a=0}^{a=p} 2a\,da\,dp = \int_{p=0}^{p=1} p^2\,dp &&= \tfrac{1}{3}\\
\mathrm{E}_{\mathbf{B}}(p) &= \int_{a=0}^{a=1}\int_{p=a}^{p=1} 2p\,dp\,da = \int_{a=0}^{a=1} 1-a^2\,dp &&= \tfrac{2}{3}\\
\mathrm{E}_{\mathbf{B}}(c) &= \tfrac{2}{3}-\tfrac{1}{3} &&= \tfrac{1}{3}
\end{aligned}$$

These formulae has the implication that in a tabulation in one-year classes of age and date of event and date of birth, the mean ages represented will be $\frac{1}{3}, \frac{2}{3}, \frac{4}{3}, \frac{5}{3}, \frac{7}{3}, \frac{8}{3}, \ldots$, the mean dates of event will be $\frac{1}{3}, \frac{2}{3}, \frac{4}{3}, \frac{5}{3}, \frac{7}{3}, \frac{8}{3}, \ldots$, and the mean dates of birth will be similarly irregularly spaced, as illustrated in figure 5.3.

5.3 Modelling data from triangular subsets

If modelling of rates is be based on data tabulated by age, period *and* cohort, the mean values of the three variables should be used, since any part of the diagram must have age, period and cohort assigned which obey that $a = p - c$. If this is ignored as in repeated papers by Boyle and Robertson [24, 1, 25, 26] one will introduce strange constraints on parameters on the model parameters.

However, another anomaly arises when traditional age-period-cohort models are applied to triangular data: The log-likelihood will be the sum of two parts each with a distinct set of

parameters, one part with only parameters and data from upper triangles, and one part with parameters and data from lower triangles. This is seen by noting (for example from figure 5.3), that mean age, period and cohort in any upper triangles nowhere appears an a lower triangle and vice versa.

So effectively an age-period cohort model for data from a triangular tabulation will be two separate age-period-cohort models, one based on the upper triangles and one based on the lower ones. This was already noted by Osmond & Gardner [22]. The implication is that the parameter estimates from the upper and lower triangles do not necessarily fit nicely together, in particular because the parameter constraints we have discussed so far will constrain two different cohorts (one from upper and one from lower triangles) to be 0, and we will have to choose 4 periods (two from upper and two from lower triangles) to be 0.

Thus the tabulation itself will cause severe problems for a sensible reporting of the results. As we shall see in the next section it is not really the tabulation but rather the parametrization that causes problems. Osmond & Gardner [22] noted the problem, but did not provide any satisfactory solution to it. We shall return to this in the next chapter.

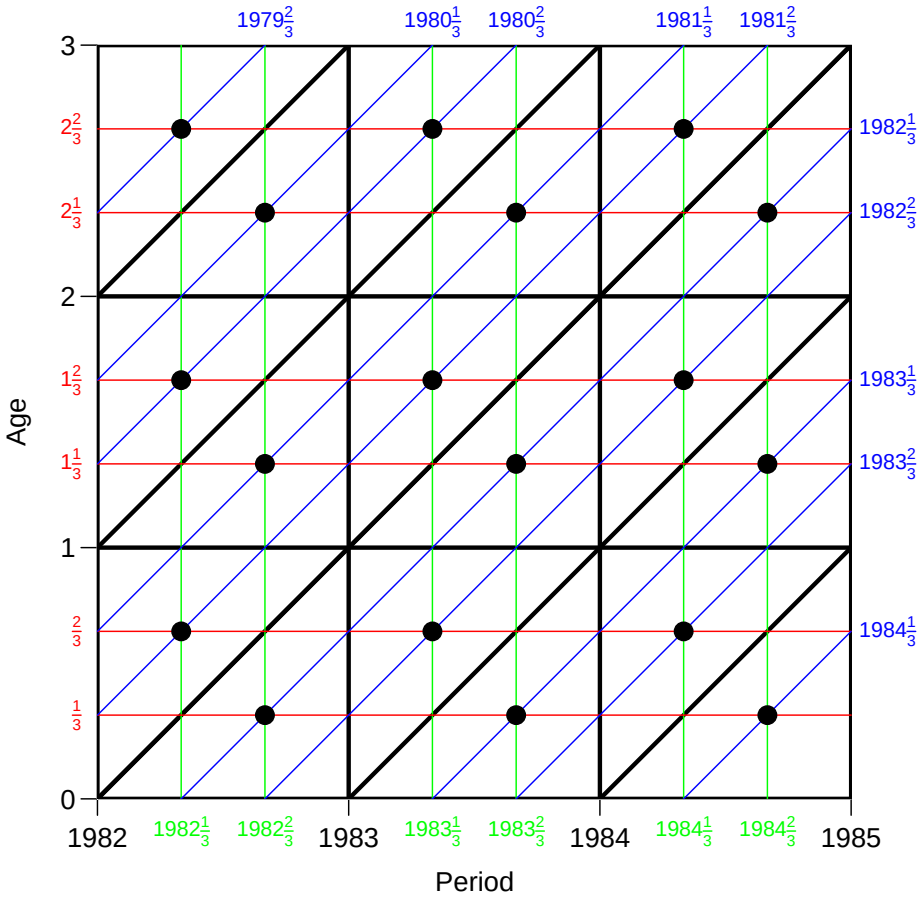


Figure 5.3: *A Lexis diagram showing mean age, period and birth date (top and right) for triangular sections of the diagram.*

Chapter 6

The age-period cohort model in a general setting

In this chapter we formulate the age-period-cohort model in a general setting, allowing any kind of tabulation of data, not only by age and date of event. We will pay particular attention to the case where data are tabulated by age and date of event as well as by date of birth, corresponding to triangles in the Lexis diagram. But this chapter will be relevant to observations of rates in any kind of subsets of a Lexis diagram.

6.1 Identifiable parameters

A model for rates in rectangles of the Lexis diagram where all three variables are modelled as factors with one parameter per level is:

$$\log[\mathrm{E}(D_{ap})] = \log(Y_{ap}) + \alpha_a + \beta_p + \gamma_c$$

The relation $c = p - a$ produces an identifiability / parametrization constraint, so the model has dimension $1+(A-1)+(P-1)+(C-1)-1 = A+P+C-3$, one less than a full three-way factor model has. As we saw in chapter **??** this can be accommodated by constraining the first and the last or the 2nd and the penultimate parameters of either period or cohort terms to be 0, and in this way producing parameters easy to graph with clearly recognizable constraints. But these constraints are arbitrary and does not give the resulting age- period- and cohort-parameters any easy interpretation.

In the factor model only the second order differences between parameters for levels (curvature) are uniquely identifiable:

$$\alpha_a - 2\alpha_{a+1} + \alpha_{a+2}$$

These can be explicitly modelled by using a parametrization as given by the following contrast matrix:

age	contrast matrix						linear predictor
	μ	δ	ζ_2	ζ_3	ζ_4	ζ_5	
1	1	0	0	0	0	0	$\eta_1 = \mu$
2	1	1	0	0	0	0	$\eta_2 = \mu + 1\delta$
3	1	2	1	0	0	0	$\eta_3 = \mu + 2\delta + 1\zeta_2$
4	1	3	2	1	0	0	$\eta_4 = \mu + 3\delta + 2\zeta_2 + 1\zeta_3$
5	1	4	3	2	1	0	$\eta_5 = \mu + 4\delta + 3\zeta_2 + 2\zeta_3 + 1\zeta_4$
6	1	5	4	3	2	1	$\eta_6 = \mu + 5\delta + 4\zeta_2 + 3\zeta_3 + 2\zeta_4 + 1\zeta_5$

(6.1)

From this we can see that:

$$\eta_4 - 2\eta_5 + \eta_6 = \zeta_5, \quad \eta_3 - 2\eta_4 + \eta_5 = \zeta_4, \text{etc.}$$

so the parameters ζ_2, ζ_e etc. are indeed the 2nd order differences of the age-parameters.

The factor model thus has $A - 2 + P - 2 + C - 2$ uniquely identifiable parameters, the remaining three are the general level and the two linear effects of the time-scales, neither of which can be uniquely determined without further constraints.

The problem with the 2nd order difference parameters is that despite their uniqueness they are not easily interpretable — differences of parameters are log rate-ratios but 2nd order differences are log rate-ratio ratios!

6.1.1 Linear trends

Holford [10] suggested to extract the linear trends from the age-, period- and cohort- parameters respectively, by regressing each set of estimates on the mean age, period and cohort, and then report the residuals as age, period and cohort effects. This would give a display of the identifiable quantities on a recognizable scale.

By regressing the estimates for the age-, period- and cohort classes on age, period and cohort respectively he can obtain a set of parameters (functions) $\tilde{f}$, $\tilde{g}$ $\tilde{h}$ that are 0 on average and are connected to the original parameters by:

$$\begin{aligned} f(a) &= \tilde{f}(a) + \mu_a + \delta_a a \\ g(p) &= \tilde{g}(p) + \mu_p + \delta_p p \\ h(c) &= \tilde{h}(c) + \mu_c + \delta_c c \end{aligned}$$

Holford notes that $\delta_a + \delta_p$ and $\delta_p + \delta_c$ are invariants in the sense that any parametrization by f, g and h will yield the same value of this. This is because any parametrization (f, g, h) can be obtained from another ($\tilde{f}$, $\tilde{g}$, $\tilde{h}$) as:

$$\begin{aligned} f(a) &= \tilde{f}(a) - \mu_p - \mu_c + \gamma a \\ g(p) &= \tilde{g}(p) + \mu_p \qquad - \gamma p \\ h(c) &= \tilde{h}(c) \qquad + \mu_c + \gamma c \end{aligned}$$

where it is easily verified that $f(a)+g(p)+h(c) = \tilde{f}(a)+\tilde{g}(p)+\tilde{h}(p)$ for all values of $(a, p, c = p-a)$. Hence, any change in parametrization will change the regression slopes of f, g and h with the same numerical quantity, but leave the resulting slopes extracted by regression unchanged.

Implementation of Holford's approach

Holford proposed to make a parametrization of the effects by using parameters corresponding to the orthogonal complement to the linear trend. To clarify this idea consider a parametrization

of for example the period effect as follows:

period	contrast matrix						linear predictor
p	μ	δ	β_2	β_3	β_4	β_5	
1	1	0					$\eta_1 = \mu + 0\delta + \sum_{j=2}^{5} \beta_j x_{1j}$
2	1	1					$\eta_2 = \mu + 1\delta + \sum_{j=2}^{5} \beta_j x_{2j}$
3	1	2		(x_{pj})			$\eta_3 = \mu + 2\delta + \sum_{j=2}^{5} \beta_j x_{3j}$
4	1	3					$\eta_4 = \mu + 3\delta + \sum_{j=2}^{5} \beta_j x_{4j}$
5	1	4					$\eta_5 = \mu + 4\delta + \sum_{j=2}^{5} \beta_j x_{5j}$
6	1	5					$\eta_6 = \mu + 5\delta + \sum_{j=2}^{5} \beta_j x_{6j}$

(6.2)

The columns for the βs are chosen so that they span a space which is orthogonal to the space spanned by the first two columns, intercept and drift. Since we are not interested in the single parameters β but only in their sum, it is immaterial how the the columns are chosen.

This has the effect that the (in this example) 6 (= P) parameter functions $\sum_{j=2}^{5} \beta_j x_{pj}$ represent "residuals" around the period drift. If similar orthogonal columns are used for the cohort effect the estimated drift parameter will be unique.

However the uniqueness depends on the definition of "orthogonal". The formulae devised by Holford are based orthogonality w.r.t. the usual inner product:

$$< x \,|\, y >= \sum x_i y_i$$

However this is not necessarily a sensible way of defining the drift. It might be equally sensible to use an inner product of the type:

$$< x \,|\, y >= \sum w_i x_i y_i$$

with some pre-defined weights. Holford's proposal is to use $w_i = 1$ for all i, but it might be equally sensible to use a different inner product for cohort and period contrasts, with w_i proportional to the number of cases in each period/cohort. This would correspond to regressing the estimates on the period using the number of cases in each period as weights.

Hence, the resulting "unique" linear components devised by Holford are just as arbitrary as any other set of extracted linear trends. The size of extracted linear trends are not a feature of the *model*; they are a feature of the model *and* the chosen method for extracting the linear trend. At least three different options for extracting a linear trend comes readily to mind:

- Simple linear regression of the estimated values at a prespecified set of points. This cooresponds to Holford's proposal.
- Weighted linear regression of the estimated values at the points where observations are. Weights could be the number of cases, arguing that the variance of the parameters (log-rates) is inversely proportional to the number of cases.
- Weighted linear regression as above but using the inverse of the floated variances as weights.

The latter will only work for models where effects of the three variables are modelled as factors, whereas the other two approaches can be used to extract trends from models where effects are modelled by prametric functions.

These three approaches are pursued in the classical analysis of the lung cancer data as well as in the analysis of the triangular tabulated lung cancer data in the programs listed below.

It is seen that the results for the so called identifiable trend is somewhat dependent on the choice of regression technique, giving an overall trend of 3.2% or 1.9% per year for lung cancer in Denmark over the period 1943–1996, the first being Holford's suggestion, the latter the weighted approach. For comparison, the slope estimate from the age-drift model is 2.3% per year.

(Strictly speaking we should take the exponential of these figures and subtract one in order to get the correct estimates but it makes little difference in this case because the numbers are so small.)

6.2 The curse of the tabulation

If the tabulation of data becomes increasingly fine, the age- period- cohort modelling by one factor level for each distinct value of the three factors becomes infeasible. This problem in age-period- cohort modelling emerges because the "factor" approach insists that effects be modelled by one separate parameter for each distinct value of the tabulation factors age, period and cohort. The factor models are effectively models that let the tabulation induce the model.

The classical approach emerging from cancer epidemiology has been to define a tabulation sufficiently coarse to avoid an excess amount of parameters in the modelling. Thus a natural and understandable wish to keep the number of parameters of the models at a reasonable level has lead to a coarse tabulation of data.

Thus there has been an unproductive feed-back loop between the tabulation of data and the modelling approach based on the concept of piecewise constant rates to be modelled individually — the"factor"-modelling approach. One may speculate whether this has been induced by limited availability of population figures or by limited computing capacity (initially the need to compute standardized rates by hand before the advent of proper modelling hard- and software).

6.3 Sensible parametrizations

If the problem is seen as modelling rates observed in some subset of the Lexis diagram, the logical approach would be to formulate the problem in continuous time, i.e. by a model for the rate at any point in the diagram:

$$\log[\lambda_{ap}] = f(a) + g(p) + h(c)$$

for three suitably smooth functions. This would in principle predict the rates at any point in the Lexis diagram, independently of the data available.

In order to accommodate this we would tabulate data as finely as possible and then use the model to describe age- period- and cohort- effects in as much detail as possible or desired within the limits set by the information content in the data (i.e. the number of events), that is deciding on the number of parameters we want to use for f, g and h. A separate issue will be the precise form of parametrization to choose for the functions.

For any tabulation of data in subsets of Lexis diagram an age- period- cohort model can be viewed as a split of the rate-function in three descriptive parts:

$$\log[\lambda(a,p)] = f(a) + g(p) + h(c = p - a) \tag{6.3}$$

i.e. as functions of the mean (a, p, c) in each of the subsets of the Lexis diagram considered.

The challenge is to choose parametrizations of these three functions in a way that is

1. meaningful,

2. understandable and recognizable,

3. practically estimable by standard software.

In particular the emphasis should be on the *estimation* of the effects and not on the testing of whether a particular effect is present or not. The emphasis should be on whether a deviation from linearity is clinically or epidemiologically relevant and not whether it is statistically significant. This requires standard errors of estimates to replace overall tests for cohort or period effects.

The only components of the model (6.3) that can be uniquely determined are the second derivatives of the three functions, and yet the relevant representation of the model is by graphs of three functions f, g and h that sum to the predicted log-rates. The first derivatives as well as the absolute levels can be moved around between the functions. An obvious choice could then be only to show the second derivatives, but as there are not on an easily understandable scale for these this is not an option in practise.

Instead f, g and h should be shown as functions constrained in a way that makes it clear what constraints have been chosen.

6.3.1 A suggestion for choice of parameters

Consider for a moment the age-cohort model:

$$\log[\lambda(a,c)] = f(a) + h(c)$$

In this model we face the classical problem that only the *first* derivative of f and h are identifiable. This is traditionally fixed by choosing a reference cohort c_0, say, and constrain $h(c_0) = 0$. This will make $f(a)$ interpretable as the age-specific log-rates in cohort c_0 and $h(c)$ as the log rate ratio of cohort c compared to cohort c_0. This was the approach we chose when we discussed the age-cohort model in section 3.3.

The formalism behind this is to write:

$$\ln(\lambda(a,c)) = f(a) + h(c) = (f(a) + \mu) + (h(c) - \mu)$$

and by choosing $\mu = h(c_0)$ we get the desired functions as:

$$\tilde{f}(a) = f(a) - h(c_0) \qquad \tilde{h}(c) = h(c) - h(c_0)$$

which indeed has the property that $h(c_0) = 0$. In practical terms this can be implemented by choosing the parametrization of the model carefully.

A similar machinery can be invoked to explicitly move the unidentifiables in an age-period-cohort model around according to desired constraints by extracting linear parts of the functions:

$$\begin{aligned} \log(\lambda(a,p)) &= \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c) \\ &= f(a) - (\mu_a + \beta_a a) + \\ &\quad g(p) - (\mu_p + \beta_p p) + \\ &\quad h(c) - (\mu_c + \beta_c c) \end{aligned}$$

The terms can now be moved between the age, period and cohort terms, in order to make these comply to certain constraints.

In the approach to the age-cohort model where only the *first* derivatives of the age and cohort effects are identifiable, we chose an arbitrary fix-point (the *zero*th derivative) for one

of them, making age-parameters interpretable as age-specific rate in a particular cohort and cohort parameters as rate-ratios relative to this. In the age-period-cohort where only the *second* derivatives of the effects are identifiable, we will also need to choose a fixpoint for the *first* derivative for one of them.

This can be done by fixing the slope between two points on the period scale or by requiring that the overall slope of the period parameters in some sense has a given value (0, for example). So we propose the following constraints:

- The age-function is interpretable as age-specific rates in cohort c_0 after adjustment for the period effect.
- The cohort function is 0 at a reference cohort c_0, interpretable as log-RR relative to cohort c_0.
- The period function is 0 on average, interpretable as log-RR relative to the age-cohort prediction.

This way of distributing effects between the three variables depends on:

- The choice of reference cohort c_0.
- The choice of how to define "0 on average" for $\tilde{g}$.

For practical computations where functions $\hat{f}$, $\hat{g}$ and $\hat{h}$ have been estimated we extract the linear part from $\hat{g}$, for example by regressing the values of $\hat{g}(p)$ for all occurring values of p on p using number of cases in each category as weights (corresponding to weighting by an approximate measure of inverse variance):

$$\hat{g}(p) = \tilde{g}(p) + (\mu + \beta p)$$

and then get following three functions: h

$$\begin{aligned}
\ln(\lambda(a,p)) &= \hat{f}(a) + \hat{g}(p) + \hat{h}(c) \\
&= \hat{f}(a) + \\
&\quad \tilde{g}(p) + (\mu + \beta p) + \\
&\quad \hat{h}(c) \\
&= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\
&\quad \tilde{g}(p) + \\
&\quad \hat{h}(c) - \hat{h}(c_0) + \beta(c - c_0)
\end{aligned}$$

The practical implementation of the procedure is therefore as follows:

1. Fit the age-period-cohort model with any parametrization, and obtain $\hat{f}$, $\hat{g}$ and $\hat{h}$.
2. Regress $\hat{g}$ on p (for example using D as weights), obtain intercept μ_p, slope β_p and residual $\tilde{g}(p)$: $\hat{g}(p) = \mu_p + \beta_p p + \tilde{g}(p)$
3. Report the effects as:

Age-specific incidence rates in cohort c_0:

$$\exp[\hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0]$$

Rate-ratio relative to cohort c_0:

$$\exp[\hat{h}(c) - \hat{h}(c_0) + \beta(c - c_0)]$$

Rate-ratio by period controlled for age and cohort:

$$\exp[\tilde{g}(p)]$$

This will produce *one* set of functions whose product is the maximum likelihood estimates of the rates. This is however but one set. The regression approach advocated here has the disadvantage that the parametrization is a function of data, and hence that standard errors of the resulting parameters are effectively intractable.

The procedure we adopted for the age-period-cohort model in section **??**, where penultimate periods were fixed to 0, is merely a variant of this where $\tilde{g}$ is defined as the function that is 0 at these two periods.

The approach used here implicitly assumes that the cohort is a more important scale than than period. However the role of the two can easily be interchanged using exactly the same procedure as outlined above.

Extraction of linear trends for lung cancer data in 5×5 year classes

The following example uses male lung cancer data from Denmark to illustrate this way of extracting the linear trend, and how various choices of the "0 on average" affects the extracted trend.

The models used are a model for the rectangular data using factor coding of age-classes and periods and synthetic cohorts.

It is seen from the results in the section at the end that the effect of using the naïve regression approach is that the drift (in this case) is overestimated. Whether the spline estimate resulting from the weighted regression is more sensible than the one obtained from the age-drift model is not clear.

```
 R 2.1.0
 ------------------------------------------------
 Program:  lung-trend-x.R
  Folder:  C:\Bendix\Undervis\APC\r
 Started:  onsdag 13. juli 2005, 18:02:46
 ------------------------------------------------
> library( Epi )
> # Read the 5 by 5 by 5 year tabulated data on lung cancer
> #
> lu <- read.table( "../data/lung5-Mc.txt", header=T )
> lu[1:10,]
   A5   P5   C5  D        Y up       Ax       Px       Cx
1  40 1943 1898 52 336233.8  1 43.33333 1944.667 1901.333
2  40 1943 1903 28 357812.7  0 41.66667 1946.333 1904.667
3  40 1948 1903 51 363783.7  1 43.33333 1949.667 1906.333
4  40 1948 1908 30 390985.8  0 41.66667 1951.333 1909.667
5  40 1953 1908 50 391925.3  1 43.33333 1954.667 1911.333
6  40 1953 1913 23 377515.3  0 41.66667 1956.333 1914.667
7  40 1958 1913 56 365575.5  1 43.33333 1959.667 1916.333
8  40 1958 1918 43 383689.0  0 41.66667 1961.333 1919.667
9  40 1963 1918 44 385878.5  1 43.33333 1964.667 1921.333
10 40 1963 1923 38 371361.5  0 41.66667 1966.333 1924.667
> attach( lu )
>
> # Fit models
> # First the drift model
```

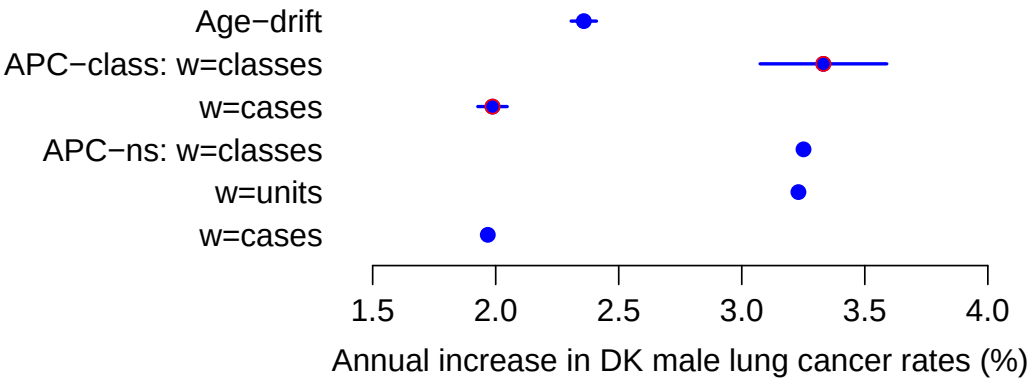
```
> ad  <- glm( D ~ factor( A5 ) + I(P5-A5) + offset( log( Y ) ), family=poisson )
> # then the full apc-model
> apc <- glm( D ~ factor( A5 ) + factor( P5-A5 ) + factor( P5 ) +
+                 offset( log( Y ) ), family=poisson )
> summary( apc )

Call:
glm(formula = D ~ factor(A5) + factor(P5 - A5) + factor(P5) +
    offset(log(Y)), family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-5.28485  -1.49866  -0.05384   1.54503   4.75866

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -11.968761   0.381235 -31.395  < 2e-16
factor(A5)45          0.994171   0.038120  26.080  < 2e-16
factor(A5)50          1.873728   0.036406  51.468  < 2e-16
factor(A5)55          2.559631   0.036144  70.817  < 2e-16
factor(A5)60          3.087159   0.036576  84.404  < 2e-16
factor(A5)65          3.479003   0.037513  92.742  < 2e-16
factor(A5)70          3.760496   0.038829  96.847  < 2e-16
factor(A5)75          3.888688   0.040667  95.623  < 2e-16
factor(A5)80          3.903883   0.043468  89.811  < 2e-16
factor(A5)85          3.801889   0.049829  76.299  < 2e-16
factor(P5 - A5)1863  -0.006614   0.420362  -0.016 0.987447
factor(P5 - A5)1868   0.490330   0.388598   1.262 0.207023
factor(P5 - A5)1873   0.789467   0.382514   2.064 0.039028
factor(P5 - A5)1878   0.994787   0.380779   2.613 0.008988
factor(P5 - A5)1883   1.330029   0.379948   3.501 0.000464
factor(P5 - A5)1888   1.786722   0.379480   4.708 2.50e-06
factor(P5 - A5)1893   2.119172   0.379270   5.588 2.30e-08
factor(P5 - A5)1898   2.367836   0.379162   6.245 4.24e-10
factor(P5 - A5)1903   2.560970   0.379115   6.755 1.43e-11
factor(P5 - A5)1908   2.640060   0.379117   6.964 3.31e-12
factor(P5 - A5)1913   2.645518   0.379193   6.977 3.02e-12
factor(P5 - A5)1918   2.728574   0.379248   7.195 6.26e-13
factor(P5 - A5)1923   2.819711   0.379334   7.433 1.06e-13
factor(P5 - A5)1928   2.806058   0.379517   7.394 1.43e-13
factor(P5 - A5)1933   2.837759   0.379803   7.472 7.92e-14
factor(P5 - A5)1938   2.729073   0.380375   7.175 7.25e-13
factor(P5 - A5)1943   2.726104   0.381329   7.149 8.74e-13
factor(P5 - A5)1948   2.933442   0.383417   7.651 2.00e-14
factor(P5 - A5)1953   2.947867   0.395385   7.456 8.94e-14
factor(P5)1948        0.095424   0.034467   2.769 0.005630
factor(P5)1953        0.104771   0.030359   3.451 0.000558
factor(P5)1958        0.200248   0.026503   7.556 4.16e-14
factor(P5)1963        0.249105   0.023356  10.666  < 2e-16
factor(P5)1968        0.311059   0.020498  15.175  < 2e-16
factor(P5)1973        0.295911   0.018183  16.274  < 2e-16
factor(P5)1978        0.294441   0.016231  18.141  < 2e-16
factor(P5)1983        0.249025   0.014928  16.682  < 2e-16
factor(P5)1988        0.103123   0.014601   7.063 1.63e-12
factor(P5)1993              NA         NA      NA       NA

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72456.29  on 219  degrees of freedom
Residual deviance:   888.66  on 182  degrees of freedom
AIC: 2521.9

Number of Fisher Scoring iterations: 4

> cf <- coef( apc )
> cf[is.na(cf)] <- 0
>
> apar <- c(0,cf[grep( "\\(A5", names( cf ) )])
> ppar <- c(0,cf[grep( "P5\\)", names( cf ) )])
> cpar <- c(0,cf[grep( "P5 ", names( cf ) )])
>
> # Compute no. cases and use these as weights, and as points of evaluation
> acase <- tapply( D, A5, sum )
> pcase <- tapply( D, P5, sum )
> ccase <- tapply( D, P5-A5, sum )
>
> apt <- as.numeric( names( acase ) )
> ppt <- as.numeric( names( pcase ) )
> cpt <- as.numeric( names( ccase ) )
>
> # Naive regressions a.m. Holford
> delta0.a <- coef( lm( apar ~ apt ) )[2]
> delta0.p <- coef( lm( ppar ~ ppt ) )[2]
> delta0.c <- coef( lm( cpar ~ cpt ) )[2]
>
```

```
> # Weighted regression
> deltaw.a <- coef( lm( apar ~ apt, weights=as.vector( acase ) ) )[2]
> deltaw.p <- coef( lm( ppar ~ ppt, weights=as.vector( pcase ) ) )[2]
> deltaw.c <- coef( lm( cpar ~ cpt, weights=as.vector( ccase ) ) )[2]
>
> #-------------------------------------------------------------------------------
> # Here is the orthogonality machinery
> #-------------------------------------------------------------------------------
> # Define the contrasts orthogonal to the linear by using the
> # default polynomial contrasts as starting point.
> # Orthogonality to (1,t) is extended to be w.r.t. the weighted inner
> # product at the expense of losing the internal orthogonality
> contr.nlin <- function( n, w=rep(1,n) ) ( diag( 1/w ) %*% contr.poly( n ) )[,-1]
> # Note that the result is orthogonal to cbind(1,t), but only a
> # orhogonal basis, if the weights are 1.
> round( MP <- contr.nlin( 11 ), 3 )
          .Q     .C     ^4    ^5     ^6     ^7     ^8     ^9    ^10
 [1,]  0.512 -0.458  0.355 -0.24  0.142 -0.072  0.030 -0.010  0.002
 [2,]  0.205  0.092 -0.355  0.48 -0.453  0.330 -0.188  0.081 -0.023
 [3,] -0.034  0.336 -0.355  0.08  0.274 -0.473  0.443 -0.274  0.105
 [4,] -0.205  0.351 -0.059 -0.32  0.340  0.029 -0.413  0.487 -0.279
 [5,] -0.307  0.214  0.237 -0.32 -0.113  0.402 -0.085 -0.426  0.489
 [6,] -0.341  0.000  0.355  0.00 -0.378  0.000  0.425  0.000 -0.586
 [7,] -0.307 -0.214  0.237  0.32 -0.113 -0.402 -0.085  0.426  0.489
 [8,] -0.205 -0.351 -0.059  0.32  0.340 -0.029 -0.413 -0.487 -0.279
 [9,] -0.034 -0.336 -0.355 -0.08  0.274  0.473  0.443  0.274  0.105
[10,]  0.205 -0.092 -0.355 -0.48 -0.453 -0.330 -0.188 -0.081 -0.023
[11,]  0.512  0.458  0.355  0.24  0.142  0.072  0.030  0.010  0.002
> round( t(MP) %*% cbind( ".1"=1, .L=1:11, MP ), 3 )
    .1 .L .Q .C ^4 ^5 ^6 ^7 ^8 ^9 ^10
.Q   0  0  1  0  0  0  0  0  0  0   0
.C   0  0  0  1  0  0  0  0  0  0   0
^4   0  0  0  0  1  0  0  0  0  0   0
^5   0  0  0  0  0  1  0  0  0  0   0
^6   0  0  0  0  0  0  1  0  0  0   0
^7   0  0  0  0  0  0  0  1  0  0   0
^8   0  0  0  0  0  0  0  0  1  0   0
^9   0  0  0  0  0  0  0  0  0  1   0
^10  0  0  0  0  0  0  0  0  0  0   1
> # Also works with a set of weights as e.g. the number of cases,
> # but no more an orthogonal basis (however irrelevant for our purposes)
> wt <- tapply( D, P5, sum )
> round( MP <- contr.nlin( 11, wt ), 6 )
             .Q        .C        ^4        ^5        ^6        ^7        ^8        ^9      ^10
 [1,]  0.000420 -0.000376  0.000291 -0.000197  0.000116 -0.000059  0.000025 -0.000008  2.0e-06
 [2,]  0.000102  0.000045 -0.000176  0.000238 -0.000225  0.000164 -0.000093  0.000040 -1.2e-05
 [3,] -0.000012  0.000113 -0.000120  0.000027  0.000092 -0.000160  0.000149 -0.000092  3.5e-05
 [4,] -0.000045  0.000077 -0.000013 -0.000070  0.000075  0.000006 -0.000091  0.000107 -6.1e-05
 [5,] -0.000049  0.000034  0.000038 -0.000051 -0.000018  0.000064 -0.000013 -0.000068  7.7e-05
 [6,] -0.000041  0.000000  0.000042  0.000000 -0.000045  0.000000  0.000051  0.000000 -7.0e-05
 [7,] -0.000031 -0.000022  0.000024  0.000033 -0.000012 -0.000041 -0.000009  0.000043  5.0e-05
 [8,] -0.000019 -0.000032 -0.000005  0.000029  0.000031 -0.000003 -0.000037 -0.000044 -2.5e-05
 [9,] -0.000003 -0.000029 -0.000031 -0.000007  0.000024  0.000041  0.000039  0.000024  9.0e-06
[10,]  0.000019 -0.000009 -0.000034 -0.000046 -0.000043 -0.000031 -0.000018 -0.000008 -2.0e-06
[11,]  0.000064  0.000057  0.000045  0.000030  0.000018  0.000009  0.000004  0.000001  0.0e+00
> round( t(MP) %*% diag( wt ) %*% cbind( ".1"=1, .L=1:11, MP ), 6 )
    .1 .L        .Q        .C        ^4        ^5        ^6        ^7        ^8        ^9       ^10
.Q   0  0  0.000325 -0.000171  0.000104 -0.000033  0.000013 -0.000002  0.000001  0.000000 -0.000001
.C   0  0 -0.000171  0.000302 -0.000156  0.000089 -0.000025  0.000007  0.000001 -0.000001  0.000000
^4   0  0  0.000104 -0.000156  0.000278 -0.000137  0.000072 -0.000018  0.000003  0.000001  0.000000
^5   0  0 -0.000033  0.000089 -0.000137  0.000252 -0.000118  0.000058 -0.000014  0.000003  0.000000
^6   0  0  0.000013 -0.000025  0.000072 -0.000118  0.000228 -0.000101  0.000046 -0.000010  0.000001
^7   0  0 -0.000002  0.000007 -0.000018  0.000058 -0.000101  0.000207 -0.000082  0.000031 -0.000006
^8   0  0  0.000001  0.000001  0.000003 -0.000014  0.000046 -0.000082  0.000181 -0.000061  0.000018
^9   0  0  0.000000 -0.000001  0.000001  0.000003 -0.000010  0.000031 -0.000061  0.000157 -0.000037
^10  0  0 -0.000001  0.000000  0.000000  0.000000  0.000001 -0.000006  0.000018 -0.000037  0.000132
>
>
> # Use the orthogonal parametrizations
> apc.dr <- glm( D ~ factor( A5 ) - 1 + I( P5-A5 ) +
+                    C( factor( P5-A5 ), contr.nlin ) +
+                    C( factor( P5 ), contr.nlin ) +
+                    offset( log( Y ) ), family=poisson )
> round( ci.lin( apc.dr, subset="I", Exp=T ), 3 )
          Estimate StdErr      z P exp(Est.)  2.5% 97.5%
I(P5 - A5)    0.033  0.001 25.781 0     1.033 1.031 1.036
> apc.wd <- glm( D ~ factor( A5 ) - 1 + I( P5-A5 ) +
+                    C( factor( P5-A5 ), contr.nlin, w=ccase ) +
+                    C( factor( P5    ), contr.nlin, w=pcase ) +
+                    offset( log( Y ) ), family=poisson )
> round( ci.lin( apc.wd, subset="I", Exp=T ), 3 )
          Estimate StdErr      z P exp(Est.)  2.5% 97.5%
I(P5 - A5)     0.02      0 65.755 0      1.02 1.019  1.02
>
> # Extract the drift estimates from the models and
```

```
> # plot the regression slopes from the factor model to demonstrate
> # that the results are actually the same.
> d.est <- rbind(
+  "Age-drift"        = ci.lin( ad    , subset="I\\(", Exp=T )[,5:7],
+  "APC-class: w=classes" = ci.lin( apc.dr, subset="I\\(", Exp=T )[,5:7],
+                 " " = rep( exp(delta0.p + delta0.c), 3 ),
+  "w=cases"=ci.lin( apc.wd, subset="I\\(", Exp=T )[,5:7],
+                 " " = rep( exp(deltaw.p + deltaw.c), 3 ) )
> d.est
                     exp(Est.)     2.5%    97.5%
Age-drift             1.023580 1.023065 1.024096
APC-class: w=classes  1.033315 1.030744 1.035893
                      1.033315 1.033315 1.033315
w=cases               1.019870 1.019272 1.020468
                      1.019870 1.019870 1.019870
>
>
> res <-
+ cbind( c(delta0.a,
+          delta0.p,
+          delta0.c,
+          delta0.p + delta0.a,
+          delta0.p + delta0.c),
+        c(deltaw.a,
+          deltaw.p,
+          deltaw.c,
+          deltaw.p + deltaw.a,
+          deltaw.p + deltaw.c) )
> rownames( res ) <- c("A","P","C","A+P","P+C")
> colnames( res ) <- c("  Reg,w=1", "  Reg,w=D" )
> round( cbind( res, "ratio"=res[,1]/res[,2] ), 4 )
     Reg,w=1   Reg,w=D   ratio
A      0.0832    0.0767  1.0853
P      0.0013   -0.0022 -0.5683
C      0.0315    0.0219  1.4378
A+P    0.0845    0.0744  1.1349
P+C    0.0328    0.0197  1.6657
>
> #---------------------------------------------------------------------
> # Models with splines
> #---------------------------------------------------------------------
> # Fit models to the triangular data using splines
> library( splines )
> # First define the internal and boundary knots
> A.kn <- seq(  50,  80, 5) ; A.b <- c(  40,  90)
> C.kn <- seq(1880,1940,10) ; C.b <- c(1840,1960)
> P.kn <- seq(1960,1980, 5) ; P.b <- c(1940,2000)
> # Then fit the models:
> # First fit the drift model
> ad.s  <- glm( D ~ ns( Ax, knots=A.kn, Bo=A.b, intercept=T ) - 1 + Px +
+                   offset( log( Y ) ), family=poisson )
> # Then the full apc-model
> apc.s <- glm( D ~ ns( Ax, knots=A.kn, Bo=A.b ) +
+                   ns( Px, knots=P.kn, Bo=P.b ) +
+                   ns( Cx, knots=C.kn, Bo=C.b ) +
+                   offset( log( Y ) ), family=poisson )
> summary( apc.s )

Call:
glm(formula = D ~ ns(Ax, knots = A.kn, Bo = A.b) + ns(Px, knots = P.kn,
    Bo = P.b) + ns(Cx, knots = C.kn, Bo = C.b) + offset(log(Y)),
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.55134  -0.96146   0.06885   0.85844   3.62161

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -11.07576    0.33590 -32.974  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)1   2.32772    0.05423  42.927  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)2   2.75395    0.07056  39.028  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)3   3.06559    0.07013  43.715  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)4   3.21215    0.08065  39.828  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)5   3.29622    0.08591  38.370  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)6   2.60224    0.08773  29.662  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)7   4.16749    0.15422  27.022  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)8   1.89012    0.10461  18.069  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)1   1.08246    0.06839  15.828  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)2   1.33981    0.08298  16.145  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)3   1.44462    0.08746  16.518  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)4   1.57703    0.08788  17.946  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)5   2.39303    0.17638  13.567  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)6   1.47720    0.11426  12.928  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)1   0.76797    0.33801   2.272  0.02308
```

```
ns(Cx, knots = C.kn, Bo = C.b)2   1.12068    0.34319   3.265  0.00109
ns(Cx, knots = C.kn, Bo = C.b)3   0.82755    0.32435   2.551  0.01073
ns(Cx, knots = C.kn, Bo = C.b)4   0.63597    0.32151   1.978  0.04792
ns(Cx, knots = C.kn, Bo = C.b)5   0.45991    0.30557   1.505  0.13230
ns(Cx, knots = C.kn, Bo = C.b)6  -0.24018    0.20603  -1.166  0.24372
ns(Cx, knots = C.kn, Bo = C.b)7  -0.41552    0.64254  -0.647  0.51783
ns(Cx, knots = C.kn, Bo = C.b)8        NA         NA      NA       NA

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72456.29  on 219  degrees of freedom
Residual deviance:   396.19  on 198  degrees of freedom
AIC: 1997.5

Number of Fisher Scoring iterations: 4

> cf <- coef( apc.s )
> cf[is.na(cf)] <- 0
>
> # Predictions for each of the terms, omitting the offset,
> # giving log-rates and log-RRs
> pterm <- predict( apc.s, type="terms" )
> pterm[1:10,]
   ns(Ax, knots = A.kn, Bo = A.b) ns(Px, knots = P.kn, Bo = P.b) ns(Cx, knots = C.kn, Bo = C.b)
1                       -1.869605                    -0.99457093                     0.44143196
2                       -2.165685                    -0.92246206                     0.39914220
3                       -1.869605                    -0.77822182                     0.36388985
4                       -2.165685                    -0.70608681                     0.28247093
5                       -1.869605                    -0.56177794                     0.24289538
6                       -2.165685                    -0.48960044                     0.17084233
7                       -1.869605                    -0.34519025                     0.13757339
8                       -2.165685                    -0.27282651                     0.07455918
9                       -1.869605                    -0.12294696                     0.04388238
10                      -2.165685                    -0.04318111                    -0.02291254
>
> # Put the constant with the age-effect. This is the "hat" functions.
> # Note the weird place R puts the intercept for this parametrization
> apar <- pterm[,1] + attr( pterm, "cons" )
> ppar <- pterm[,2]
> cpar <- pterm[,3]
>
> # Naive regressions a.m. Holford weighted by no. observations.
> deltan.a <- coef( lm( apar ~ Ax ) )[2]
> deltan.p <- coef( lm( ppar ~ Px ) )[2]
> deltan.c <- coef( lm( cpar ~ Cx ) )[2]
>
> # Weighted regression using no. of cases as weights
> deltaw.a <- coef( lm( apar ~ Ax, weights=D ) )[2]
> deltaw.p <- coef( lm( ppar ~ Px, weights=D ) )[2]
> deltaw.c <- coef( lm( cpar ~ Cx, weights=D ) )[2]
>
> # Naive regressions a.m. Holford per estimate
> a.eff <- tapply( apar, Ax, mean )
> p.eff <- tapply( ppar, Px, mean )
> c.eff <- tapply( cpar, Cx, mean )
> delta0.a <- coef( lm( a.eff ~ as.numeric( names( a.eff ) ) ) )[2]
> delta0.p <- coef( lm( p.eff ~ as.numeric( names( p.eff ) ) ) )[2]
> delta0.c <- coef( lm( c.eff ~ as.numeric( names( c.eff ) ) ) )[2]
>
> res.s <-
+ cbind( c(delta0.a,
+          delta0.p,
+          delta0.c,
+          delta0.p + delta0.a,
+          delta0.p + delta0.c),
+        c(deltan.a,
+          deltan.p,
+          deltan.c,
+          deltan.p + deltan.a,
+          deltan.p + deltan.c),
+        c(deltaw.a,
+          deltaw.p,
+          deltaw.c,
+          deltaw.p + deltaw.a,
+          deltaw.p + deltaw.c) )
> rownames( res.s ) <- c("A","P","C","A+P","P+C")
> colnames( res.s ) <- c("APC-ns: w=classes", "w=units", "w=cases" )
>
> # Compare all approaches
> round( (exp( cbind( res, res.s ) ) - 1 ) * 100, 3 )
      Reg,w=1   Reg,w=D APC-ns: w=classes w=units w=cases
A       8.678     7.970            5.227   5.227   4.525
P       0.127    -0.223            3.385   3.385   3.046
C       3.200     2.215           -0.129  -0.150  -1.046
A+P     8.816     7.728            8.789   8.789   7.709
```

```
P+C     3.332     1.987             3.251   3.230   1.968
>
> # Plot together with drift estimates
> d.est <-
+ rbind( d.est,
+        exp(t(res.s[c(5,5,5),])) )
>
> plt( "Lung-drift", height=2.5 )
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.5 )
> plotEst( (d.est-1)*100, y=c( c(3,2,2,1,1)+3, 3:1 ),
+         pch=c(16,16,1,16,1,16,16,16),
+         col=c("blue","red")[c(1,1,2,1,2,1,1,1)], lwd=2,
+         xlab="Annual increase in DK male lung cancer rates (%)" )
>
>
>
 ---------------------------------------------
 Program: lung-trend-x.R
  Folder: C:\Bendix\Undervis\APC\r
   Ended: onsdag 13. juli 2005, 18:02:49
 Elapsed: 00:00:03
 ---------------------------------------------
```



6.3.2 Choice of f, g and h in practise.

Splines

Spline functions are functions that between pre-defined points, so called *knots*, are polynomials, and that are joined together at these points with continuous derivatives at the knots.

The simplest from of splines are *linear* splines, curves that are piecewise linear. One implementation of linear splines for a variable x is by including the variables x, $(x-k_1)_+$, $(x-k_2)_+$, etc. in the model. Here we use the notation:

$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \leq 0 \end{cases}$$

The matrix in (6.1) can be recognized as the contrast matrix for linear splines with equidistant knots, so a natural generalization of the factor model would be to use linear splines.

Splines are mostly used in the form of cubic splines, which are 3rd degree polynomials between the knots constrained to have continuous 1st and 2nd derivatives at the knots. These functions require $k+3$ parameters if we use k knots. One implementation is the truncated power basis:

$$x, \quad x^2, \quad x^3, \quad (x-k_1)^3_+, \quad (x-k_2)^3_+, \ldots$$

A modification of cubic splines are *restricted* cubic splines or *natural* splines, that are restricted to be linear outside the outer knots. This induces restrictions on the coefficients used to form the basis, so these only require $k-1$ parameters if we use k knots. The expression for the basis functions is somewhat complicated.

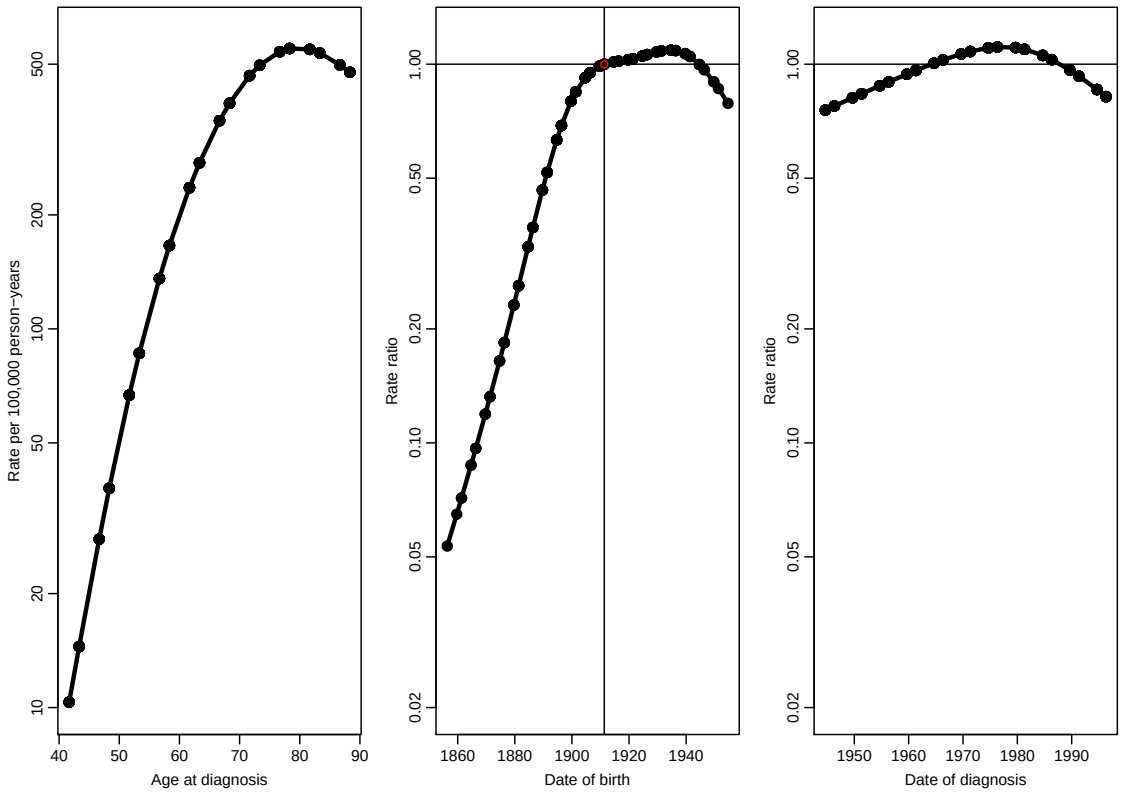


Figure 6.1: *Resulting figure from the* R*-program. Male lung cancer incidence in Denmark.*

An important point in implementation of splines is that the allocation of knots is required *a priori*. Also it is important to specify the boundary knots.

Splines in R

Both splines and natural splines are implemented in the R-package `splines` via the functions `bs()` and `ns()`, respectively.

One important point in the use of these is that the functions have a mechanism to allocate knots automatically based on the vector of xs supplied. Thus in order to make sure that allocation of knots is the same in the actual analysis where data is supplied and in prediction where a predefined range of the variable is supplied, it is essential that knots be assigned explicitly.

It is recommended to use natural splines in modelling because they are more stable than basic splines.

Example of application of natural splines in R.

The following program illustrates how to use natural splines in R and how to produce estimates constrained according to the suggestions above.

```
 R 1.9.0
 ------------------------------------------
 Program:  lung-spl-x.R
  Folder:  C:\Bendix\Undervis\APC\r
 Started:  tirsdag 20. juli 2004, 11:19:07
 ------------------------------------------
> # Read the 5 by 5 by 5 year tabulated data on lung cancer
> #
```

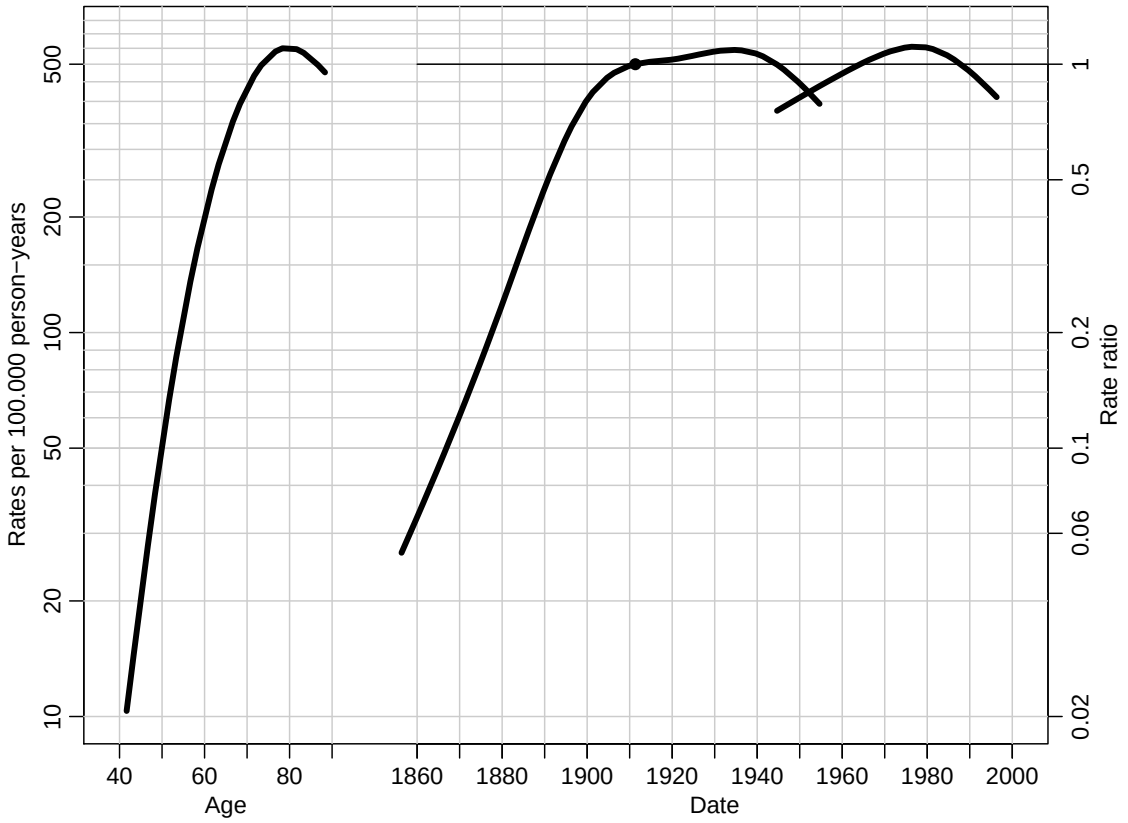


Figure 6.2: *Estimates from the male lung cancer age-period cohort model plotted using the same scales for all effects.*

```
> lu <- read.table( "../data/lung5-Mc.txt", header=T )
> lu[1:10,]
   A5   P5   C5  D        Y up       Ax       Px       Cx
1  40 1943 1898 52 336233.8  1 43.33333 1944.667 1901.333
2  40 1943 1903 28 357812.7  0 41.66667 1946.333 1904.667
3  40 1948 1903 51 363783.7  1 43.33333 1949.667 1906.333
4  40 1948 1908 30 390985.8  0 41.66667 1951.333 1909.667
5  40 1953 1908 50 391925.3  1 43.33333 1954.667 1911.333
6  40 1953 1913 23 377515.3  0 41.66667 1956.333 1914.667
7  40 1958 1913 56 365575.5  1 43.33333 1959.667 1916.333
8  40 1958 1918 43 383689.0  0 41.66667 1961.333 1919.667
9  40 1963 1918 44 385878.5  1 43.33333 1964.667 1921.333
10 40 1963 1923 38 371361.5  0 41.66667 1966.333 1924.667
> attach( lu )
>
> library( splines )
> # Fit model with defined knots
> # First the knots
> A.kn <- seq(  50,  80,10) ; A.b <- c(  40,  90)
> C.kn <- seq(1880,1940,20) ; C.b <- c(1840,1960)
> P.kn <- seq(1960,1980,10) ; P.b <- c(1940,2000)
> # Then fit the model
> apc <- glm( D ~ ns( Ax, knots=A.kn, Bo=A.b ) +
+                 ns( Cx, knots=C.kn, Bo=C.b ) +
+                 ns( Px, knots=P.kn, Bo=P.b ) + offset( log( Y ) ),
+             family=poisson )
> summary( apc )

Call:
glm(formula = D ~ ns(Ax, knots = A.kn, Bo = A.b) + ns(Cx, knots = C.kn,
    Bo = C.b) + ns(Px, knots = P.kn, Bo = P.b) + offset(log(Y)),
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-4.41354  -0.95933   0.09351   1.08558   3.41767

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -13.61998    0.27947 -48.735  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)1   3.36960    0.04261  79.077  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)2   4.11197    0.05058  81.300  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)3   3.77547    0.03651 103.399  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)4   5.86817    0.09849  59.581  < 2e-16
ns(Ax, knots = A.kn, Bo = A.b)5   2.90056    0.04124  70.342  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)1   3.81067    0.27123  14.050  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)2   3.69943    0.28611  12.930  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)3   3.37246    0.18566  18.164  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)4   5.01618    0.56172   8.930  < 2e-16
ns(Cx, knots = C.kn, Bo = C.b)5   2.24084    0.15801  14.182  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)1   0.47285    0.03640  12.992  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)2   0.47452    0.03109  15.265  < 2e-16
ns(Px, knots = P.kn, Bo = P.b)3   0.55846    0.08896   6.278 3.44e-10
ns(Px, knots = P.kn, Bo = P.b)4        NA         NA      NA       NA

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72456.29  on 219  degrees of freedom
Residual deviance:   506.03  on 206  degrees of freedom
AIC: 2091.3

Number of Fisher Scoring iterations: 4

>
> # Predictions for each of the terms, omitting the offset,
> # giving log-rates and log-RRs
> pterm <- predict( apc, type="terms" )
> pterm[1:10,]
   ns(Ax, knots = A.kn, Bo = A.b) ns(Cx, knots = C.kn, Bo = C.b) ns(Px, knots = P.kn, Bo = P.b)
1                      -2.468064                      0.3028151                   -0.270133532
2                      -2.803517                      0.3833639                   -0.242541173
3                      -2.468064                      0.4107853                   -0.187742251
4                      -2.803517                      0.4455021                   -0.160597912
5                      -2.468064                      0.4550294                   -0.106975044
6                      -2.803517                      0.4640021                   -0.080558739
7                      -2.468064                      0.4656794                   -0.028671949
8                      -2.803517                      0.4689960                   -0.003284331
9                      -2.468064                      0.4727176                    0.045442023
10                     -2.803517                      0.4841977                    0.068235089
>
> # Put the constant with the age-effect. This is the "hat" functions.
> # Note the weird place R puts the intercept for this parametrization
> a.eff <- pterm[,1] + attr( pterm, "cons" )
> c.eff <- pterm[,2]
> p.eff <- pterm[,3]
>
> # Define reference cohort as the one with maximal no. cases
> tc <- tapply( D, Cx, sum )
> c0 <- as.numeric( names( tc )[tc==max(tc)][1] )
>
> # Get the period effect and regress on period using no. cases as weights
> p.lm <- lm( p.eff ~ Px, weight=D )
> mu   <- coef( p.lm )[1]
> beta <- coef( p.lm )[2]
> # Compte the cohort effect at cohort c0
> h.c0 <- mean( c.eff[Cx==c0] )
>
> # Now we can make the three terms
> a.rates <- exp( a.eff + mu + beta * Ax + h.c0 + beta * c0 )
> c.RR    <- exp( c.eff      + beta * Cx - h.c0 - beta * c0 )
> p.RR    <- exp( p.eff - mu - beta * Px )
>
> # Now plot the three effects. Order effects when they are plotted
> pdf( "../graph/lung-spl-3.pdf", width=8, height=8/sqrt(2) )
> par( mfrow=c(1,3), mgp=c(3,1,0)/1.6, mar=c(3,3,1,1) )
> plot( Ax[order(Ax)], a.rates[order(Ax)]*10^5, log="y", pch=16, type="o", lwd=3,
+       xlab="Age at diagnosis", ylab="Rate per 100,000 person-years",
+       ylim=c(10,600) )
> plot( Cx[order(Cx)], c.RR[order(Cx)], log="y", pch=16, type="o", lwd=3,
+       xlab="Date of birth", ylab="Rate ratio", ylim=0.02*c(1,60) )
> abline( h=1, v=c0 )
> points( c0, 1, col="red" )
> plot( Px[order(Px)], p.RR[order(Px)], log="y", pch=16, type="o", lwd=3,
+       xlab="Date of diagnosis", ylab="Rate ratio", ylim=0.02*c(1,60) )
> abline( h=1 )
>
 -------------------------------------------
 Program: lung-spl-x.R
  Folder: C:\Bendix\Undervis\APC\r
   Ended: tirsdag 20. juli 2004, 11:19:08
```

```
Elapsed: 00:00:01
-----------------------------------------------
```

Chapter 7

Age-period cohort models for multiple datasets

When several sets of rates are observed in a Lexis diagram it is of course possible to fit separate age-period-cohort models for each set of rates. But depending on the context different approaches to modelling will be appropriate:

1. Rates of the same disease in different populations:

   (a) Men and women:

   Differences in age-specific rates and possibly also in cohort effects, whereas period effecs may be taken to be similar between sexes (s):

   $$\log(\lambda(a,p,s)) = f_s(a) + g(p) + h_s(c)$$

   The identification problem is the same in this model as in the model for a single set of rates, the linear trend can be moved beween the three terms as before, except that the change in age-trends alway will be the same for males and females. Hence the ratio of rates between males and females, $\exp\big((f_{\mathrm{m}}(a) - f_{\mathrm{f}}(a))\big)$ is identifiable.

   Likewise the ratio of cohort effects will be identifiable.

   However, if both age-effects and cohort effects are assumed to depend on sex, the rate-ratios are only determined up to a constant which can be moved between the ratio of age effecst and the ratio of cohort effects.

   (b) Different nations or ethnic groups (same sex):

   Age-specific rates will be assumed to be similar (proportional) between nations, whereas differences in cohort and period effects are of interest:

   $$\log(\lambda(a,p,n)) = f(a) + \alpha_n + g_n(p) + h_n(c)$$

   Rate-ratios of period and cohort effects between nations are only determined up to a constant. The α_n terms represent rate-ratios between populations, but referring to a specific chosen value of age, period and cohort.

   The identification problem is also present here when an age-curve is presented, it will have to refer to some chosen constarint on (one of) the period and cohort effects.

2. Rates of different diseases in the *same* population, for example rates different histological subtypes or different subsites.

Formally, this is a competing risk problem, but a s long as the rates are small, the problem can safely be treated as separate rates with the same population basis.

In this case interest be in differences between subtypes in all three effects, so effectively requiring fitting separate age-period-cohort models to the separate sets of rates:

$$\log(\lambda(a,p,t)) = f_t(a) + g_t(p) + h_t(c)$$

If we want to compare cohort effects between two subsites say, the obvoius quantity to consider would be $\exp\big((h_1(c) - h_2(c)\big))$. However there is an arbitrary linear trend involved in this comparison, since we effectively are fitting separate APC-models for the two subsites.

If external considerations do not suggest a model with common age or period effects the only feasible solution would be to applythe same set of constraints on the model for each subsite and then graph them together.

For all the models above it will be perfectly feasible to circumvent the identifiability problem by first fitting a model with (marginal) effects of age and cohort say, and subsequenttly fitting a model with (conditional) effects of period. These would pose no technincal identifiability problems.

7.1 Example: Male and female lung cancer in Denmark

7.2 Example: Cervical cancer in European populations

7.3 Example: Histological subtypes of testis cancer in Denmark

Chapter 8

Using the age-period-cohort model for prediction of future rates

Having fitted an age-period-cohort model we may want to use it to predict future rates. Predicting rates in periods where data are not (yet) available will require future values for the period parameters and cohort parameters be guessed in some way. The age-parameters need not be extrapolated, as predictions outside the age-range of observations will be of little interest.

Since the parametrization of the age-period-cohort model is arbitrary, we will only want extrapolations of future period and cohort effects that will give the same predicted rates regardless of parameter constraints chosen for the model fitted to data.

In any practical situation we will only want predictions where:

1. The age-range is the same as where we already have observations.

2. Extensions of period and cohort effects are coninuous extrapolations of estimated effects.

The following will assume that these to are to be met.

Any parametrization of the model by functions $f(a)$, $g(p)$ and $h(c)$ can obtained from a given parametrization by the functions $\tilde{f}(a)$, $\tilde{g}(p)$ and $\tilde{h}(c)$ by choosing numbers μ_p, μ_c and γ:

$$\begin{aligned} f(a) + g(p) + h(c) = \tilde{f}(a) &- \mu_p - \mu_c + \gamma a + \\ \tilde{g}(p) &+ \mu_p \quad - \gamma p + \\ \tilde{h}(c) &\quad + \mu_c + \gamma c \end{aligned}$$

These two parametrizations has the property that $f(a) + g(p) + h(c) = \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c)$.

The reqirement is that a prediction rule applied to g and h, should, if applied to $\tilde{g}$ and $\tilde{h}$, produce the same set of predicted rates.

Since a prediction is a continuous extension of the functions g and h based on estimated values of g and h, the device for producing these must be devised so that a change of $g(p)$ by $\mu_p + \gamma p$, and of $h(c)$ by $\mu_c + \gamma c$, (and hence of $f(a)$ by $-(\mu_p + \mu_c) + \gamma a$), must incur the same changes to the predicted values of g and h. Otherwise the predicted values would depend on the parametrization.

The mapping of estimated values $g(p)$ and $h(c)$ to predicted values must therefore transfer affine changes of these functions untouched.

One proposal for such predictions is a linear regression of the last couple of point estimates of the period/cohort effects as proposed by Osmond [21]. In fact, any extrapolations of g and h that is a *linear regression* on values at specific points will have the desired property.

In last chapter we proposed to use natural splines (restricted cubic splines), which has the property of being linear beyond the boundary knots. Since the basis for such splines (and basic splines too) includes intercept and the variable, these functions will be usable as prediction functions. The predictions will be invariant under reparametrizations.

In particular, the arbitrary default parametrization obtained by fitting the model can be used for predictions. Using cubic splines for predictions would probably be a bad idea, because they tend to highly unstable beyond the datapoints. Natural splines that are constrained to be linear beyond the boundary knots are likely to be more stable, in particular if the upper boundary knots are chosen inside the range of the data for period and cohort.

8.1 Prediction dependence on model

In the classical approach of Osmond [21], interest is restricted to the "factor" model, that is the model with one parameter per obserevd age, period and cohort. When we use a tabulation of data in very small subsets of the Lexis diagram, we are forced to do a smoothing of some kind. If we use natural splines for smoothing, this is done by choosing a set of knots. Each of these choices will give different models, different fitted values and different predicted values.

Hence the predictions depend on the model. This seems to be non-issue in the classical litterature, but this is only so because only *one* model is discussed. The model used in the classical approach can be viewed as a special case of "smoothing"; one is assuming that f and g are step-functions of a and p, and that h is a step function of the (a, p)-steps.

Hence the prediction of rates depends on the way we have chosen to tabulate data and on which kind of model we have decided to use for modelling.

But as in the classical approach a simple and probably also more robust prediction algorithm would be to extend the functions g and h linearly beyond a given point by using the slope from a regressing a prespecified number of values of $g(p)$ on p and $h(c)$ on c.

8.2 Practical approches

Møller *et al.* [18] has analysed several ways of predicting cancer incidence rates based on age-period-cohort models or variants thereof. Not all of the prediction algorithms they propose possess the invariance. However, if a prediciton algorithm is defined from a specific parametrization of the APC-model, the considersations above will automatically be met since any specific parametrization will be invariant under change of the initial parametrization.

The approaches of Møller *et al.* [18] use the suggestion by Holford [10] to get a uniquely defined secular trend. This is then used to produce the predicted rates. However, the uniquely defined trend is not a feature of the APC-model but a feature of the model *and* the chosen definition of "0 on average", c.f. section 6.1.

8.2.1 Software practicalities

The following section shows how to implement a linear spline estimation for the fitting in R(S-Plus) and SAS respectively.

The assumption is that we have a dataset with number of cases `D`, person-years `Y`, mean age `A`, mean period `P` and mean cohort `C`, where `C=P-A`.

R

Assume `spl` is a function that generates a spline basis for a variable, e.g.:

```
spl <-
function( x, k )
{
cbind( x, outer( x, k, function( x, k ) pmax( x-k, 0 ) ) )
}
```

The succession of models is then fitted by specifying knots for the linear splines on the three timescales in `k.A`, `k.P` and `k.C`:

```
m.drift <- glm( D ~ spl( A, k.A ) + I( C-c0 ) + offset( log( Y ) ), family=poisson )
f.ad    <- fitted( m.drift )
m.coh   <- glm( D ~ spl( C, k.C ) + offset( log( f.ad ) ), family=poisson )
f.adc   <- fitted( m.coh )
m.per   <- glm( D ~ spl( P, k.P ) + offset( log( f.adc ) ), family=poisson )
```

The rest is plotting of the estimates from these models:

```
new.A <- sort( c( seq( min(A), max(A),, 100 ), k.A ) )
pr.ad <- predict( m.drift, newdata=data.frame( A=new.A,
                                               C=rep( c0, length( new.A ) ),
                                               Y=rep( 10^5, length( new.A ) ) ),
                  type="link", se.fit=T )
matplot( new.A, exp( cbind( pr.ad$fit, pr.ad$se.fit ) %*%
                     rbind( c(1,1,1), c(0,-1,1)*qnorm(  0.975 ) ) ),
         type="l", log="y", lwd=c(3,1,1), lty=rep( 1, 3 ) )

new.C <- sort( c( seq( min(C), max(C),, 100 ), k.C ) )
pr.c <- predict.glm( m.coh, newdata=data.frame( C=new.C,
                                                f.ad=rep( 1, length( new.C ) ) ),
                     type="link", se.fit=T )
matplot( new.C, exp( cbind( pr.c$fit, pr.c$se.fit ) %*%
                     rbind( c(1,1,1), c(0,-1,1)*qnorm(0.975) ) ),
         type="l", log="y", lwd=c(3,1,1), lty=rep( 1, 3 ) )

new.P <- sort( c( seq( min(P), max(P),, 100 ), k.P ) )
pr.p <- predict.glm( m.per, newdata=data.frame( P=new.P,
                                                f.adc=rep( 1, length( new.C ) ) ),
                     type="link", se.fit=T )
matplot( new.P, exp( cbind( pr.p$fit, pr.p$se.fit ) %*%
                     rbind( c(1,1,1), c(0,-1,1)*qnorm(0.975) ) ),
         type="l", log="y", lwd=c(3,1,1), lty=rep( 1, 3 ) )
```

SAS

`a a1 a2 a3 a4 a5` a spline basis for age etc.:

```
proc genmod data = APC ;
 model D = a a1 a2 a3 a4 a5 c_c0 / dist=poisson offset=lpy ;
 output out = ad  xbeta = lfitad ;
run ;

proc genmod data = ad ;
 model D = c c1 c2 c3 c4 c5 / dist=poisson offset=lfitad ;
 output out = adc  xbeta = lfitc;
run ;

proc genmod data = adc ;
 model D = p p1 p2 p3 p4 p5 / dist=poisson offset=lfitc ;
run ;
```

Chapter 9

Reporting of results

9.1 Data

Clearly state the source of the data, the extent of the Lexis diagram covered, and the fineness of the tabulation.

9.2 Estimates

9.3 Graphs

Remember to use logaritmic y-axes. And remember to scale them to the same number of decader per cm for all graphs.

Likewise it may be useful to scale all x-axes to the same number of years per cm.

9.4 Tests

Don't.

Bibliography

[1] P Boyle and C Robertson. Statistical modelling of lung cancer and laryngeal cancer incidence in Scotland 1960–1979. *American Journal of Epidemiology*, 125(4):731–744, Apr 1987.

[2] P Boyle and C Robertson. Re:"statistical modelling of lung cancer and laryngel cancer incidence in scotland 1960–1979". *American Journal of Epidemiology*, 129(1):225–226, 1989.

[3] 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.

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

[5] Paul W Dickman, Andy Sloggett, Michael Hills, and Timo Hakulinen. Regression models for relative survival. *Statistics in Medicine*, 23:51–64, 2004.

[6] F Ederer, LM Axtell, and SJ Cutler. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph*, 6:101–121, 1961.

[7] MJ Gardner and C Osmond. Interpretation of time trends in disease rates in the presence of generation effects. *Statistics in Medicine*, 3:113–130, 1984.

[8] Jan M. Hoem. Fertility rates and reproduction rates in a probabilistic setting. *Biométrie-Praximétrie*, 10:38–66, 1969.

[9] Jan M. Hoem. Correction note. *Biométrie-Praximétrie*, 11:20, 1970.

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

[11] T.R. Holford. An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases*, 38(10):831–836, 1985.

[12] R Jacobsen, N Keiding, and E Lynge. Long term mortality trends behind low life expectancy of danish women. *Journal of Epidemiology and Community Health*, 56:205–208, 2002.

[13] Niels Keiding. The method of expected number of deaths, 1786-1886-1986. *International Statistical Review*, 55(1):1–20, 1987.

[14] Niels Keiding. Statistical inference in the Lexis diagram. *Phil. Trans R. Soc. London A*, 332:487–509, 1990.

[15] LB Knudsen and MJ Murphy. Registers as data source in studies of reproductive behaviour. Technical report, Danish Center for Demographic Research. Odense., 1999.

[16] W Lexis. *Einleitung in die Theorie der Bevölkerungsstatistik*. Karl J Trübner, Strassburg, 1875.

[17] Shilian Liu, Robert Semenchiw, Chris Waters, Shi Wu Wen, Leslie S. Mery, and Yang Mao. Clues to the aetological heterogneity af testicular seminomas and non-seminomas: time trends and age-period-cohort effects. *International Journal of Epidemiology*, 29:826–831, 2000.

[18] B Møller, H Fekjær, T Hakulinen, H Sigvaldason, HH Storm, M Talbäck, and T Haldorsen. Prediction of cancer incidence in the Nordic contries: empirical comparison of different approaches. *Statistics in Medicine*, 22:2751–2766, 2003.

[19] H. Møller. Trends in incidence of testicular cancer and prostate cancer in denmark. *Human Reproduction*, 16(5):1007–1011, 2001.

[20] Y. Ogata, K. Katsura, N. Keiding, C. Holst, and A. Green. Empirical bayes age-period-cohort analysis of retrospective incidence data. *Scandinavian Journal of Statistics*, 27:415–432, 2000.

[21] C Osmond. Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, 14(1):124–129, 1985.

[22] C Osmond and MJ Gardner. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. *American Journal of Epidemiology*, 129(1):31–35, 1989.

[23] M Rewers, RA Stone, RE La Porte, AL Drash, DJ Becker, M Walczak, and LH Kuller. Poisson regression modelling of temporal variation in incidence of childhood insulin-dependent diabetes mellitus in allegheny county, pennsylvania, and wielkopolska, poland, 1970–1985. *American Journal of Epidemiology*, 129(3):569–581, 1989.

[24] C Robertson and P Boyle. Age, period and cohort models: The use of individual records. *Statistics in Medicine*, 5:527–538, 1986.

[25] C. Robertson and P. Boyle. Age-period-cohort analysis of chronic disease rates. I: Modelling approach. *Statistics in Medicine*, 17:1305–1323, 1998.

[26] C. Robertson and P. Boyle. Age-period-cohort analysis of chronic disease rates. II: Graphical approaches. *Statistics in Medicine*, 17:1325–1340, 1998.

[27] E Schifflers, M Smans, and CS Muir. Birth cohort analysis using irregular cross-sectional data: A technical note. *Statistics in Medicine.*, 4:63–75, 1985.

[28] Erling Sverdrup. Statistiske metoder ved dødelikhetsundersøkelser. Statistical memoirs, Institute of Matematics, University of Oslo, 1967.

[29] Toshiro Tango. Re: "Statistical modelling of lung cancer and laryngeal cancer incidence in scotland 1960–1979". *American Journal of Epidemiology*, 127(3):677–678, 1988.