2021 SISCER: Age-Period-Cohort Modeling and Analysis
Lecture 1: Preliminaries

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Logistics

APC R Packages

There are now a number of packages for carrying out APC modeling.

We will illustrate modeling using:

- Within the `Epi` package (Bendix Carstensen).
- Via the `apc` package (Bent Nielsen).
- Bayesian modeling can be carried out in the `BAPC` package (Andrea Riebler and Leonhard Held).

Rosenberg and Anderson (2011); Rosenberg (2018) also describe a Webtool for fitting APC models, with R code in the background.

Thanks to: Bendix Carstensen, Bent Nielsen, Andrea Riebler and Theresa Smith for sharing thoughts and materials.

My Statistical and Epidemiological Life

From 1990-1995, I was in the Department of Mathematics at Imperial College, London.

From 1996-1999, I was in the Department of Epidemiology and Public Health at Imperial College, London.

Since 1999, I've been a Professor in the Departments of Statistics and Biostatistics at the University of Washington.

Since 1996, I've worked on spatio-temporal models for health data, initially for cancer, and more recently with applications in demography, small area estimation and infectious diseases.

When one considers temporal modeling of chronic diseases, Age-Period-Cohort modeling is an interesting topic...

Motivation

Age-Period-Cohort Modeling

Epidemiologist, demographers and social scientists have a great interest in time related changes in an event of interest (death, disease incidence,...).

From Berzuini and Clayton (1994), "...time itself does not cause disease events: it is simply the scale along which other causes operate. When we model the dependence of rates upon time we are attempting to allow for the effects of other variables which we have failed to measure: the time scale is a surrogate or proxy measure for other influences."

If we add relevant covariates to the model, we would expect the strength of temporal dependence to decrease.

We discuss three time scales that are often relevant for studying disease development.

The Three Time Scales

Age effects are due to growing old (independently of exposures). More specifically, these effects can arise from physiological changes or accumulation of social experiences; for health, these can be referred to as wear and tear. From Yang and Land (2013): “Age effects therefore represent biological and social processes of aging internal to individuals and represent development changes across the life course”.

Period effects affect all age groups simultaneously. Examples: world wars, recessions and booms, famine, infectious disease pandemics, public health interventions, technological advances. Also, changes in disease classification.

Cohort effects occur because of specific events occurring at a particular time to a particular group, e.g., the introduction of a new medical antenatal practice at some time, which affects babies who are subsequently born.

The Three Time Scales: Examples

In the context of respiratory illness:

- Age effects: Loss of elasticity in lung tissue.
- Period effects: A severe air pollution incident.
- Cohort effects: Smoking.

In the context of cancer:

- Age effects: Accumulation of mutations.
- Period effects: Introduction of PSA testing, breast cancer screening.
- Cohort effects: Smoking, HPV vaccine.

The Lexis Diagram

- Cohort analyses have been a tool used by demographers and sociologists since the late 19th and early 20th centuries when cohort started to be recognized as a key time scale in tracking vital statistics (e.g., fertility and mortality).
- In 1875 Wilhelm Lexis introduced a cohort-age diagram for representing the time scales along which we calculate vital rates (Lexis, 1875; Keiding, 1990).
- The Lexis diagram is a coordinate system based on calendar time (period) and age, in which individual's trajectories are drawn as lines of unit slope joining birth to current time, or the event of interest (e.g., death).

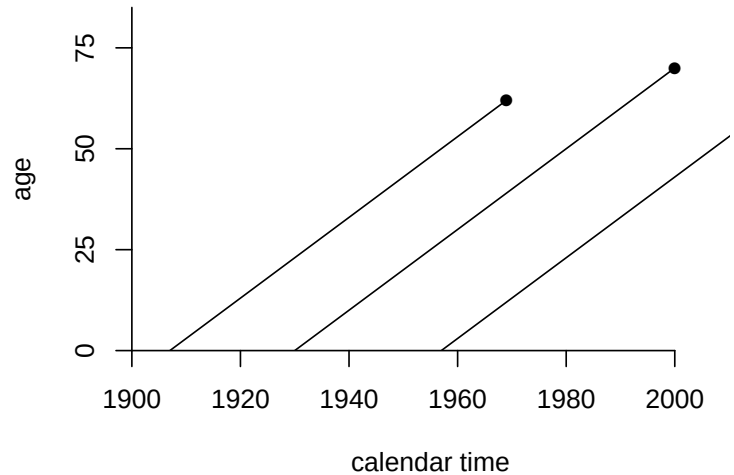


Figure 1: The Lexis diagram. Each line represents a person, with solid dots indicating death or the event of interest.

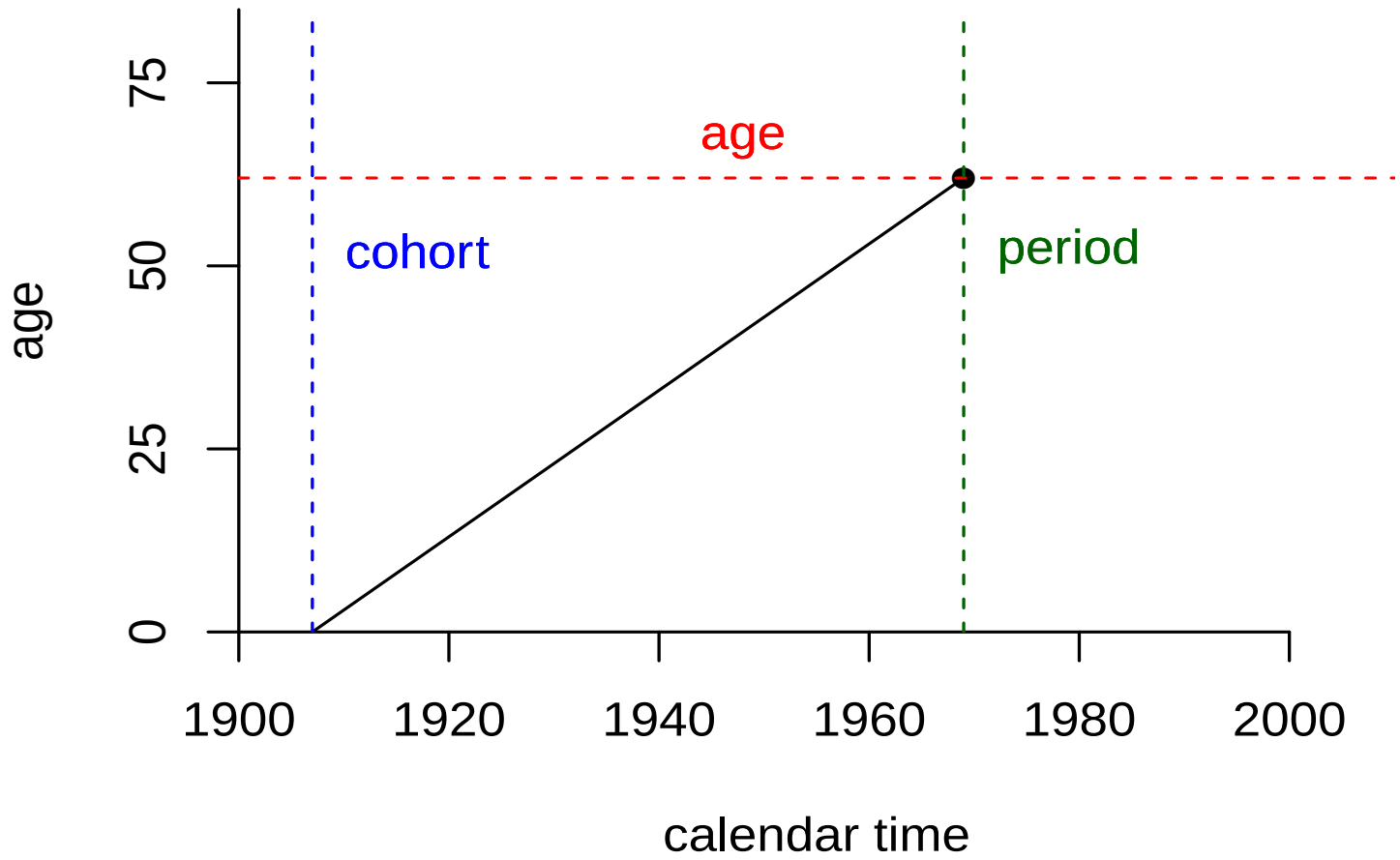


Figure 2: Age-period-cohort interpretation of the Lexis diagram.

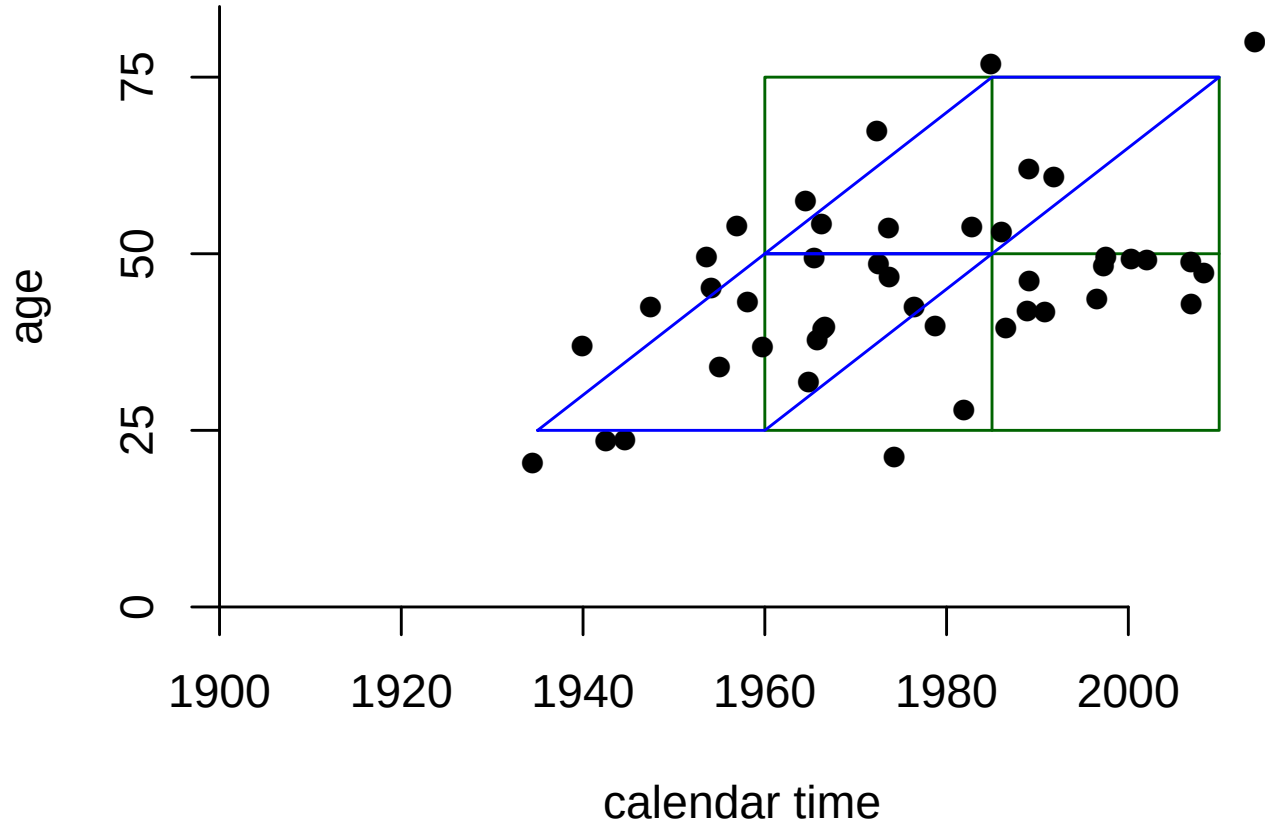


Figure 3: Aggregated data summarize the Lexis diagram, later we will discuss how data can be aggregated. Green rectangles are age-period classifications, blue parallelograms are age-cohort classifications.

This picture suggests we have an underlying temporal point process, see Keiding (1990) for more discussion of this aspect, we do not consider it further (though it's very interesting!).

The Lexis Diagram

The sum of all the life line lengths in a particular portion of the diagram represents person-years lived or exposure (risk) time in that portion.

Life lines and events can be considered from a cohort or period perspective.

Upper triangular individuals were in the age band to the left on the y-axis, when the period starts, while lower triangular individuals were in the previous age band.

For example, consider the bottom square in Figure 4: the 52 cases in the bottom left upper triangle were all in the age group 40–45 in 1943, while the 28 cases in the right lower triangle were all in the age group 35–40 in 1943.

Motivating Data: Male Lung Cancer in Denmark

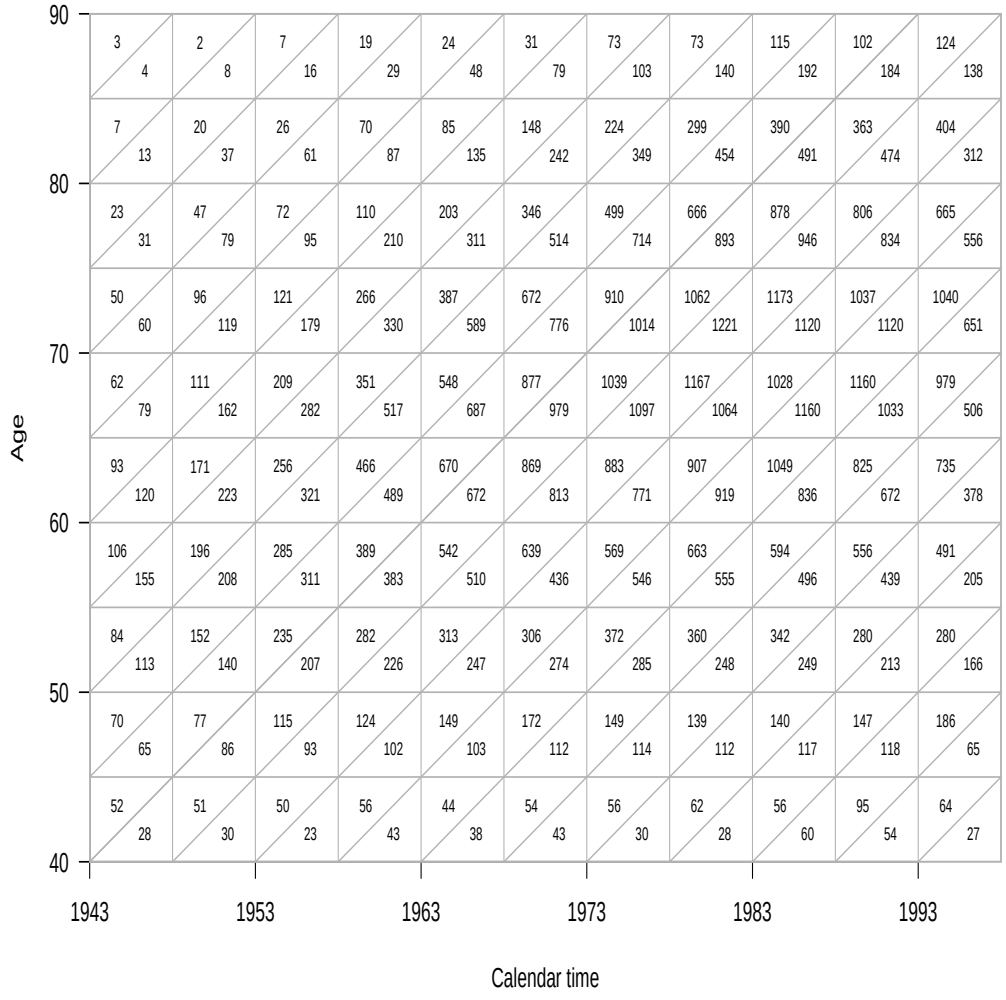


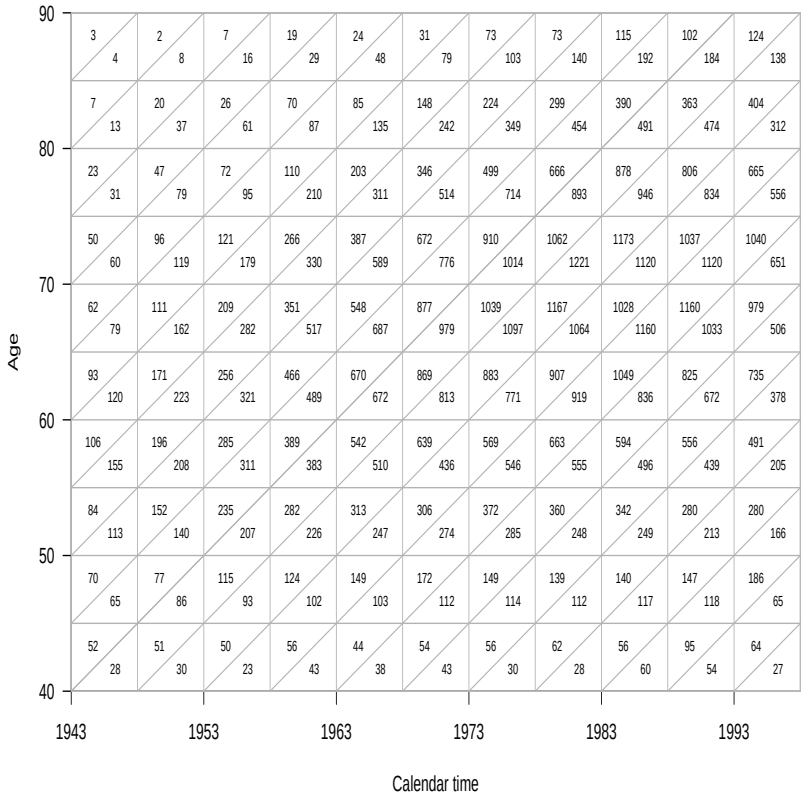
Figure 4: Lexis diagram of male lung cancer cases in Denmark, age range 40–89, over the period 1943–1996, these data are in the `Epi` package, and were collected by The Danish Cancer Registry and Statistics Denmark.

Motivating Data: Male Lung Cancer in Denmark

Horizontal $=$, vertical $\shortparallel$ and diagonal $/$ lines represent age, period and cohort. Individuals within these various lines can be characterized as falling within the relevant boundaries.

To tabulate data by

- age and period, sum over $\square$ in the Lexis diagram.
- period and cohort, sum over $\diagup\!\square$ in the Lexis diagram.
- age and cohort, sum over $\square\!\diagup$ in the Lexis diagram.



The case data are combined with population data, summed in the same way.

Estimating the Time at Risk

- Following Carstensen (2007, Section A.1) we describe how population at risk time can be estimated from census data.
- We estimate at risk times in the triangles A and B, and these can be summed up to give population estimates over the regions needed for age-period, period-cohort, age-cohort, if needed.
- Let $L_{a,p}$ represent the population size in age group a at the beginning of year p.

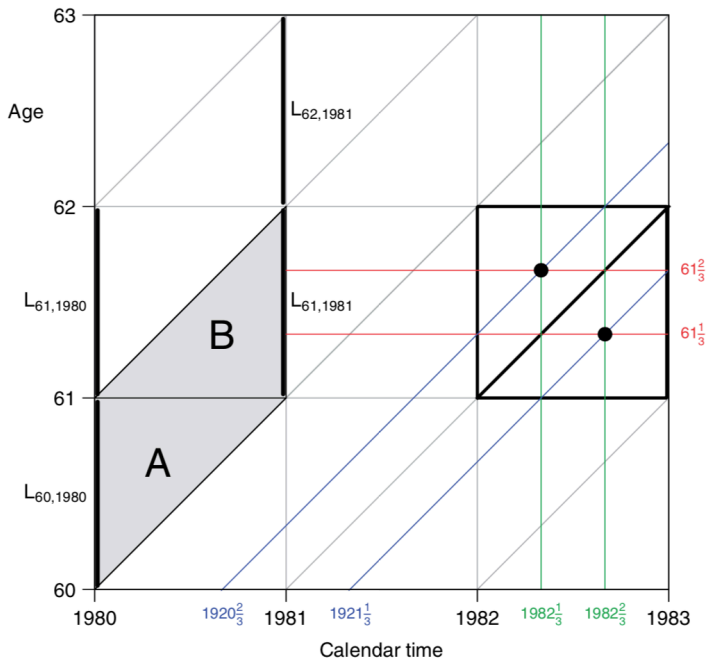


Figure 3. Lexis diagram. The thick lines in the left part show the population figures at the beginning of 1980 and 1981 necessary to estimate the population risk time in the triangles A and B. The right part of the diagram shows the mean age, period and cohort in the triangular subsets of a Lexis diagram. Note the connection between age, period and cohort: $p = c + a$: $1982\frac{1}{3} = 1920\frac{2}{3} + 61\frac{2}{3}$ and $1982\frac{2}{3} = 1921\frac{1}{3} + 61\frac{1}{3}$.

Figure 5: From Carstensen (2007).

Estimating the Time at Risk

For simplicity, we estimate the at risk time when the $L_{a,p}$ are available for 1-year age groups and periods (as is often the case).

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

As the $L_{a,p}$ population slides up the parallelogram ▱ they are subject to mortality, and we want to estimate the numbers dying in A and B, to estimate the populations.

In the presence of mortality, we know the survivors have been at risk throughout the year, so they contribute an amount,

$$\frac{1}{2} L_{a+1,p+1},$$

to each of A and B – this is the risk time we have to apportion between A and B.

Estimating the Time at Risk

Under the assumption that the the deaths in $A \cup B$ are uniformly distributed, we can obtain the expected person-time contributions to the risk time of individuals who die in A or B.

Assuming deaths occur uniiformly over time, the total amount of risk time contributed to A and B by those dying is

$$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}).$$

Estimating the Time at Risk

Those who die in A contribute no risk time to B and their average contribution to A can be calculated by integration over the triangle A; for simplicity assume age and period range from 0 to 1.

A person dying at age a, time p contributes a risk time of p ($0 \leq p \leq 1$), so the contribution (using the density function which is of height 2) is:

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

Hence, the contributed risk time is:

$$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3}.$$

Estimating the Time at Risk

Those who die in B contribute risk in both A and B — if death occurs at age a and at date p, the person has contributed $(p - a)$ person years in A and a person years in B.

The average amount contributed to A is, therefore,

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

Hence, the contributed risk time is:

$$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3}$$

Similarly, in B,

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

And the contributed risk time is:

$$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3}$$

Estimating the Time at Risk

Hence:

- given age-period population estimates, from the census or another source such as World Population Prospects (`https://population.un.org/wpp/`), to give $L_{a,p}$, $L_{a+1,p+1}$, or the Human Mortality Database.
- one can calculate the risk time for each A and B in the Lexis diagram using the last line of the table (these are age-class cohort risk times).

	A	B
Survivors	$\frac{1}{2}L_{a+1,p+1}$	$\frac{1}{2}L_{a+1,p+1}$
Dead in A	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}) \times \frac{1}{3}$	No contribution
Dead in B	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3}$	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3}$
Sum	$\frac{1}{3}L_{a,p} + \frac{1}{6}L_{a+1,p+1}$	$\frac{1}{6}L_{a,p} + \frac{1}{3}L_{a+1,p+1}$

- We can estimate the age-class a period p (◪) risk times by:

$$\frac{1}{3}L_{a,p} + \frac{1}{6}L_{a+1,p+1} + \frac{1}{6}L_{a-1,p} + \frac{1}{3}L_{a,p+1}$$

(B contribution replaces a by $a-1$).

	A	B
Survivors	$\frac{1}{2}L_{a+1,p+1} = 470$	$\frac{1}{2}L_{a+1,p+1} = 470$
Dead in A	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}) \times \frac{1}{3} = 10$	No contribution
Dead in B	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3} = 10$	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1})\frac{1}{3} = 10$
Sum	$\frac{1}{3}L_{a,p} + \frac{1}{6}L_{a+1,p+1} = 490$	$\frac{1}{6}L_{a,p} + \frac{1}{3}L_{a+1,p+1} = 480$

Table 1: An example of person-year contributions to the risk set when $L_{a,p} = 1000$ and $L_{a+1,p+1} = 940$. The sum of the risk times is 970, which is 940+(1000-970)/2.

APC Modeling: A Brief History

For a fascinating review of the early developments of APC mortality models in the 19th century, see Keiding (2011).

In the mid-20th century, studying cohort effects through simple techniques such as plots of age-specific mortality by birth year helped medical researchers understand age-time interactions for diseases such as tuberculosis (Frost, 1939; Springett, 1950).

Greenberg *et al.* (1950) introduced a primitive APC model with log incidence rates of syphilis regressed on a non-linear function of continuous age and categorical period and cohort effects.

APC Modeling: A Brief History

The application of APC models to cancer incidence and mortality gained steam in the 1980s with a series of seminal papers including Osmond and Gardner (1982); Holford (1983); Clayton and Schifflers (1987a,b).

Interest in the estimation and identification issues inherent in APC models:

- Dates back at least as far as the sociology literature in the 1970s (Mason *et al.*, 1973; Fienberg and Mason, 1979).
- Specific applications in cancer epidemiology began in the 1980s (Holford, 1983).

Overview of APC Analysis

The first step is to collate the health, population (and possibly exposures/covaraites of interest) data, and then massage into a form convenient for analysis, which will often involve aggregation.

Next, present initial tabulations of the data, and visualizations of rates as a function of the three time scales; this includes Lexis diagrams.

Age is virtually always included in a disease model and so the Age Only Model is often the null (starting) model.

Age-Period and Age-Cohort models may be fitted, if supported by the context, followed by Age-Period-Cohort models, if supported by the data.

Modeling of APC data, requires an understanding of the inherent non-identifiability, due to the interplay between the time scales.

Overview of APC Analysis

Hence, one potential sequence of models we might consider is:

- Age only.
- Age+Linear Period.
- Age+Linear Cohort.
- Age+Period.
- Age+Cohort.
- Age+Period+Cohort.

We will fit the above sequence of models to Danish male lung cancer incidence data; the analysis results are reproduced in the accompanying `R` notes.

Overview of APC Analysis

Age, period and cohort effects may be modeled in different ways including:

- Discrete time models: using a categorical classification of time (i.e., defining factors) with no constraints[1] (and often a likelihood analysis) or placing random walk priors (which encourage smoothness) on the levels and carrying out a Bayesian analysis.
- Continuous time models: Splines.

The fundamental identifiability of APC models is that when all of age, period and cohort are included in the model, the same fit to the data is obtained if we shift up/down or linearly transform ("tilt") the three sets of estimates in a compensatory fashion (if we know any 2 of age, period and cohort, the third is a linear combination).

The curvature (non-linear) aspects are estimable.

[1] for example, default models do not enforce smoothness which is often unappealing for time variables

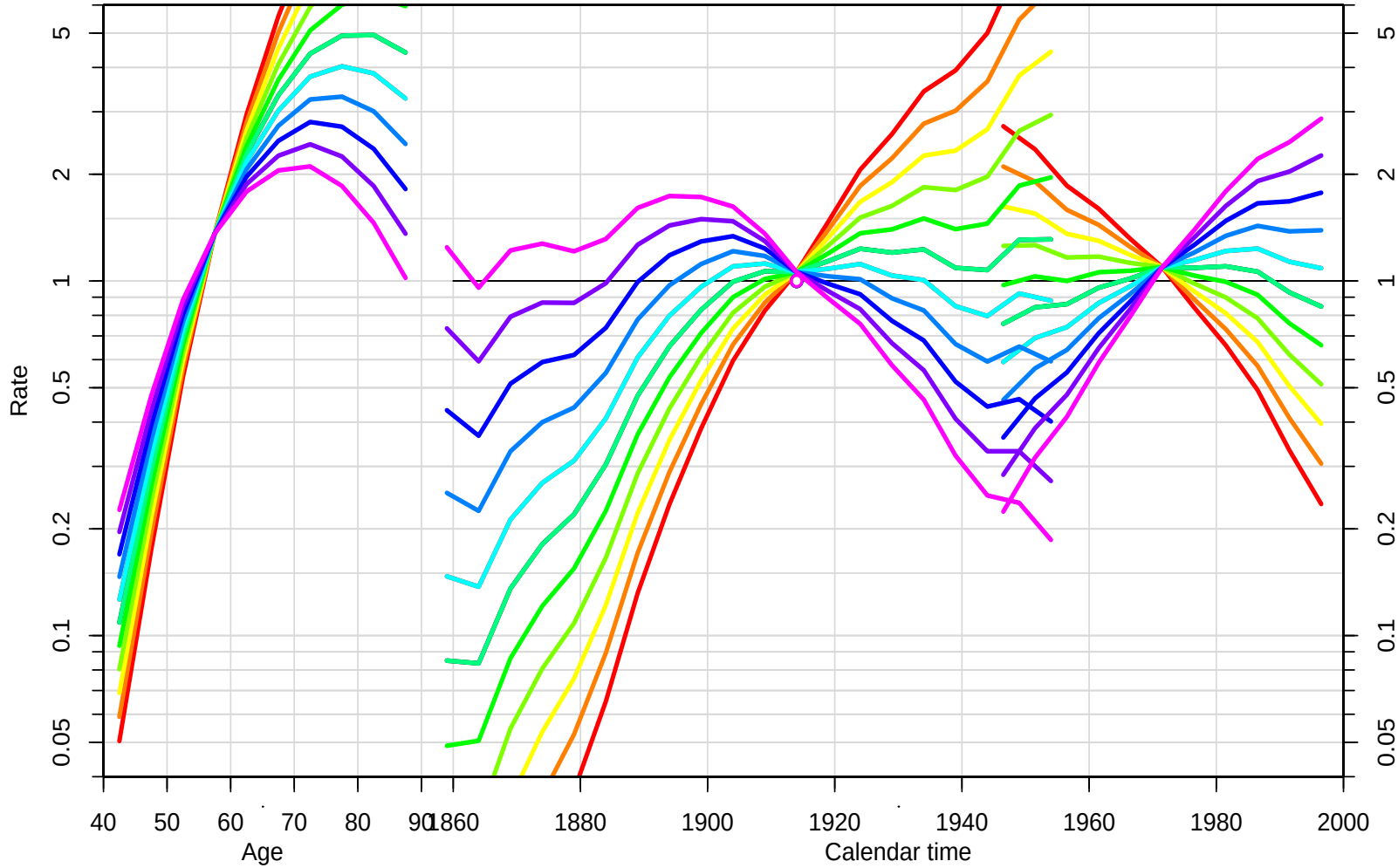


Figure 6: Illustration of the fundamental identifiability for the Danish lung cancer data. The curves of like colors provide the same fit, i.e., the fitted values for a particular set of age, period and cohort values are identical.

Some Key Statistical References

Clayton and Schifflers (1987a) look at age-period and age-cohort models in detail.

Clayton and Schifflers (1987b) examine the age-period-cohort model in detail; issues of model selection and non-identifiability also considered.

Carstensen (2007) provides a comprehensive, accessible review, and also functions for fitting and display of APC data in the `Epi` package in `R`. Includes detail on how to aggregate the data for analysis and a discussion of identifiability and parameterization.

Kuang *et al.* (2008), Nielsen and Nielsen (2014) and Martínez Miranda *et al.* (2015) discuss an appealing parameterization.

Smith and Wakefield (2016) discuss and compare various Bayesian models, including a Bayesian version of the aforementioned parameterization.

Initial Data Examinations

Tabulations and Graphs

It is always worth tabulating the data to check for errors, and see the magnitude of the counts.

Graphs are more informative for getting an initial idea of the associations between the disease rates and the the three time scales.

Table of Person Years (divided by $\times 10^5$)

Age \ Per	1943	1948	1953	1958	1963	1968	1973	1978	1983	1988	1993
40	69.4	75.5	76.9	74.9	75.7	71.0	69.5	75.6	94.1	102.6	75.3
45	62.2	67.7	73.8	75.4	73.7	74.7	69.8	68.1	74.2	92.4	82.1
50	53.9	60.1	65.4	71.6	73.4	71.8	72.5	67.5	65.9	72.0	70.1
55	47.1	51.2	57.1	62.2	68.1	69.9	68.3	68.7	64.1	62.6	54.4
60	40.3	43.5	47.4	52.8	57.3	62.7	64.4	62.8	63.0	59.1	46.3
65	32.9	35.8	38.6	42.0	46.3	50.1	54.8	56.4	54.9	55.3	42.1
70	23.0	26.9	29.5	31.7	34.1	37.4	40.4	44.3	45.9	44.9	36.6
75	14.0	16.7	19.6	21.5	22.9	24.6	26.8	29.0	31.9	33.6	26.3
80	6.8	8.1	9.9	11.6	12.6	13.7	15.0	16.3	17.6	19.6	16.8
85	2.5	2.8	3.4	4.2	4.9	5.6	6.4	7.1	7.8	8.5	7.5

- Decreasing person years with age, which has implications for estimation.

Table of Cases

Age \ Per	1943	1948	1953	1958	1963	1968	1973	1978	1983	1988	1993
40	80	81	73	99	82	97	86	90	116	149	91
45	135	163	208	226	252	284	263	251	257	265	251
50	197	292	442	508	560	580	657	608	591	493	446
55	261	404	596	772	1052	1075	1115	1218	1090	995	696
60	213	394	577	955	1342	1682	1654	1826	1885	1497	1113
65	141	273	491	868	1235	1856	2136	2231	2188	2193	1485
70	110	215	300	596	976	1448	1924	2283	2293	2157	1691
75	54	126	167	320	514	860	1213	1559	1824	1640	1221
80	20	57	87	157	220	390	573	753	881	837	716
85	7	10	23	48	72	110	176	213	307	286	262

- Once the data are aggregated, there are four recommended plots (Clayton and Schifflers, 1987a,b; Carstensen, 2007).

Table of Rates (Cases/Time at Risk) ($\times 10^4$)

Age \ Per	1943	1948	1953	1958	1963	1968	1973	1978	1983	1988	1993
40	12	11	9	13	11	14	12	12	12	15	12
45	22	24	28	30	34	38	38	37	35	29	31
50	37	49	68	71	76	81	91	90	90	69	64
55	55	79	104	124	154	154	163	177	170	159	128
60	53	91	122	181	234	268	257	291	299	253	240
65	43	76	127	207	267	370	389	395	399	396	352
70	48	80	102	188	286	388	476	515	500	480	462
75	39	76	85	149	225	350	452	537	572	487	464
80	30	71	88	135	175	285	382	461	501	426	426
85	28	35	67	114	146	196	276	299	396	335	351

- Age and period midpoints are used for row and column labels.
- There is clearly a relationship between the rates and age, with a general increase and then more gradual decline.
- There is a similar pattern with period, but the magnitude of change is smaller (maximum row totals are in red).

Initial Plots: Rates are on log scale

- Plot 1: Rates versus age, with responses in the same period connected (log rate scale), i.e., cross-sectional (in each period) age-specific rates.
- Figure 7 illustrates. For all periods there is a relatively steep increase, followed by a more gradual decline.
- Plot 1 will exhibit parallel lines if age-specific rates are proportional between periods (i.e., follow an age–period main effects model).
- The curves are not parallel which suggests age and period main effects alone are not sufficient.

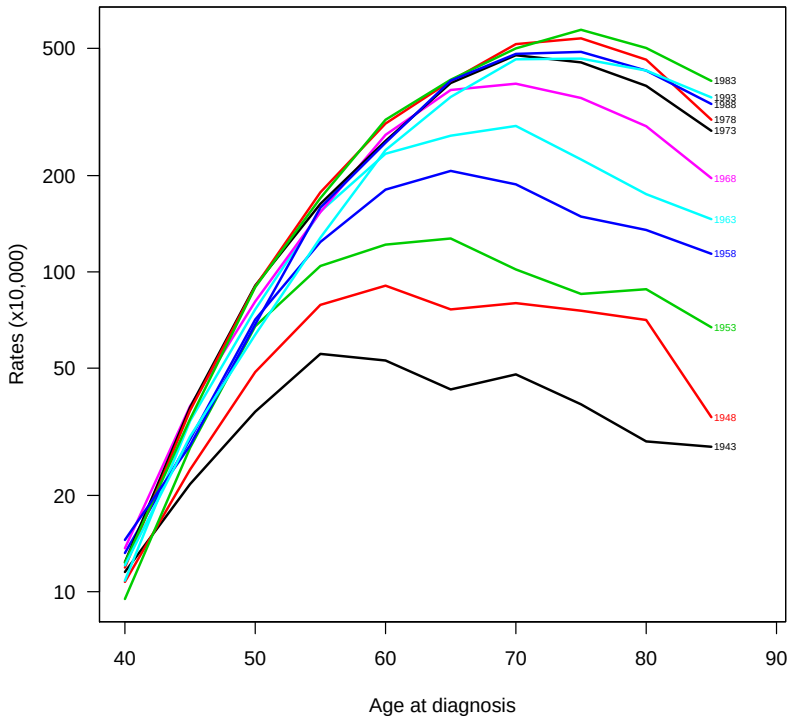


Figure 7: Plot 1: Danish lung cancer data: (log) rates against age with each curve representing one period. Each period (line) offers a cross-sectional look at the rates by age.

Initial Plots: Rates are on log scale

- Plot 2: Rates versus age, with responses in the same cohort connected, i.e., longitudinal age-specific rates.
- Figure 8 illustrates.
- Plot 2 will exhibit parallel lines if age-specific rates are proportional between cohorts (i.e., follow an age–cohort main effects model).
- The curves are not parallel which suggests age and cohort main effects alone are not sufficient.

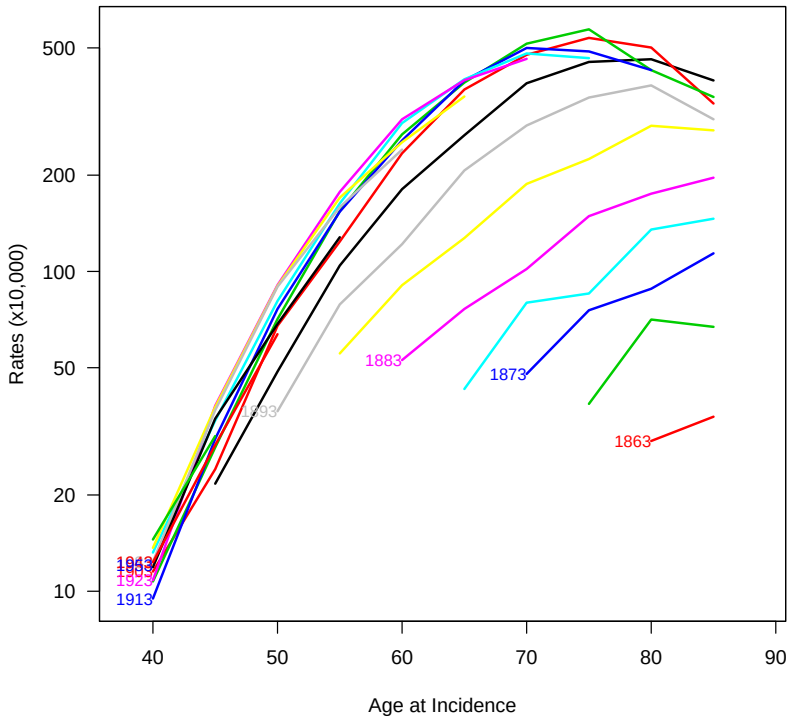


Figure 8: Plot 2: Danish lung cancer data: (log) rates against age with each curve representing one cohort. Each cohort (line) offers a longitudinal look at the rates by age.

Initial Plots: Rates are on log scale

- Plot 3: Rates versus period, with responses in the same age group connected.
- Figure 9 illustrates.
- Plot 3 will exhibit parallel lines if age-specific rates are proportional between periods (i.e., follow an age–period main effects model).
- The curves are not parallel which suggests age and period main effects alone are not sufficient.

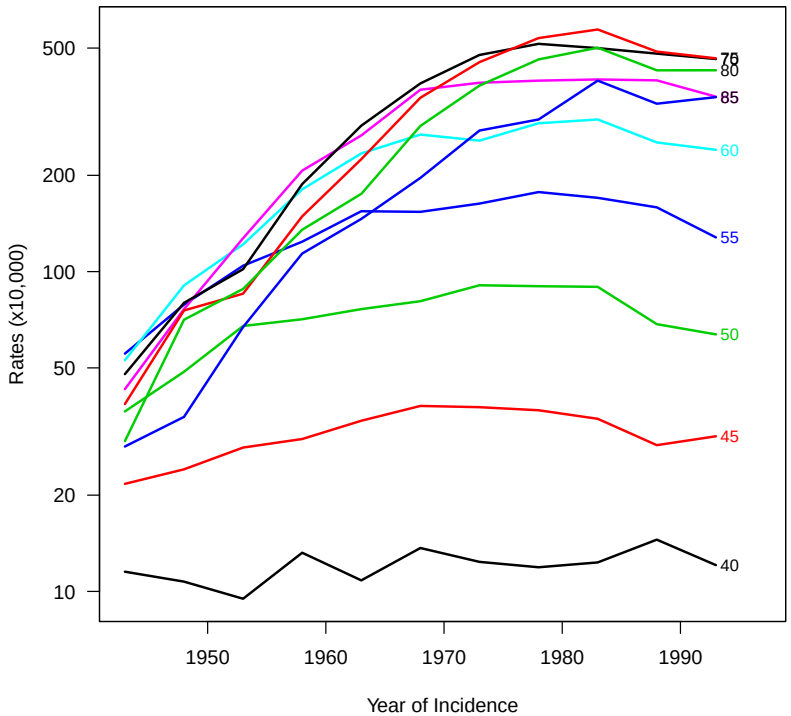


Figure 9: Plot 3: Danish lung cancer data: (log) rates against period with each curve representing one age group.

Initial Plots: Rates are on log scale

- Plot 4: Rates versus cohort, with responses of the same age connected.
- Figure 10 illustrates.
- Plot 4 will exhibit parallel lines if age-specific rates are proportional between cohorts (i.e., follow an age–cohort main effects model).
- The curves are not parallel which suggests age and cohort main effects alone are not sufficient.

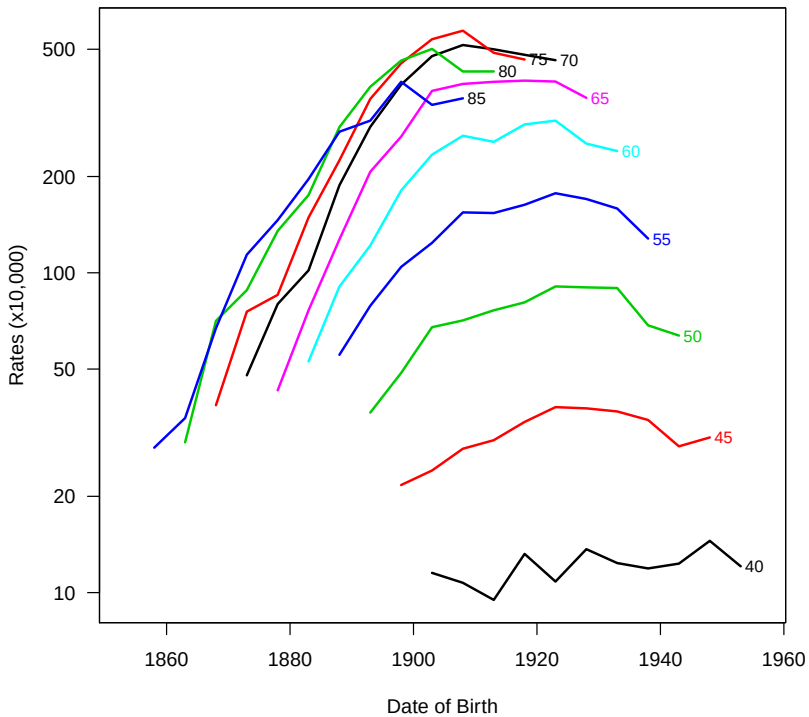


Figure 10: Plot 4: Danish lung cancer data: rates against cohort with each curve representing one age group.

Factor Models

One-Way Factor Models

We digress to discuss ANOVA models, since one approach to analysis of APC data treats each of age, period and cohort as factors.

Consideration of the ANOVA model also helps understand the non-identifiability associated with age-period-cohort models.

Suppose we have a factor with A levels and the simple linear model:

$$\mathrm{E}[Y_a] = \delta + \alpha_a,$$

for $a = 1, \ldots, A$.

No ordering or similarity assumed for $[\alpha_1, \ldots, \alpha_A]$; they are exchangeable – this is unappealing for an ordered time variable where we will often expect smoothness over time – later we will discuss random walk models that encourage smoothness.

The set of parameters $\boldsymbol{\theta} = [\delta, \alpha_1, \ldots, \alpha_A]^{\mathsf{T}}$ is not identifiable.

Identifiability in a Model with a Single Factor

Informal definition of identifiability: Identical description (fit) to a set of data arising from different sets of model parameters.

And the different parameter sets, may have very different interpretations.

We start with a simple example of non-identifiability that arises in any model with factors.

Identifiability in a model with a single factor

The model can be expressed as

$$\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{x}\boldsymbol{\theta} = \underbrace{\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}}_{A\times(A+1)} \underbrace{\begin{bmatrix} \delta \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_A \end{bmatrix}}_{(A+1)\times 1}$$

The LS estimator $\widehat{\theta} = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{Y}$ does not have a unique solution since $(\boldsymbol{x}^\top \boldsymbol{x})^{-1}$ is singular — we see that in $\boldsymbol{x}$ the sum of columns 2 to $(A + 1)$ is equal to column 1.

To gain an identifiable set of parameters, we require a single constraint.

The most common solutions are:

- Corner-Point Constraint: Set $\alpha_1 = 0$.
- Sum-to-Zero Constraint: Enforce $\sum_{a=1}^{A} \alpha_a = 0$.

Parameter interpretation of each constrained set is different.

Identifiability in a Model with a Single Factor

Identifiability corresponds to the same fit corresponding to different sets of parameters.

Recall $\boldsymbol{\theta} = [\, \delta, \alpha_1, \ldots, \alpha_A \,]$.

The identifiability problem occurs because we can add a constant, a, say to δ, and subtract that constant from each α_a, and obtain the same fit.

The transformation that represents the different sets of parameters is

$$G = \{g : g\boldsymbol{\theta} = (g\delta, g\boldsymbol{\alpha})\}$$

where

$$\begin{aligned} g\delta &= \delta - \mathrm{a} \\ g\boldsymbol{\alpha} &= \{\alpha_a + \mathrm{a}\}_{a=1}^{A} \end{aligned}$$

for any real number a.

Identifiability in a Model with a Single Factor

For simplicity, suppose $A = 2$.

The unconstrained (non-identifiable) set is

$$\underbrace{(\delta - \mathtt{a})}_{\delta^\star} + \underbrace{(\alpha_1 + \mathtt{a})}_{\alpha_1^\star} = \delta + \alpha_1$$
$$\underbrace{(\delta - \mathtt{a})}_{\delta^\star} + \underbrace{(\alpha_2 + \mathtt{a})}_{\alpha_2^\star} = \delta + \alpha_2$$

so that the fit of the parameter set $\boldsymbol{\theta} = [\delta, \alpha_1, \alpha_2]^\mathsf{T}$ is identical to the parameter set $\boldsymbol{\theta}^\star = [\delta^\star, \alpha_1^\star, \alpha_2^\star]^\mathsf{T}$, and $\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star$.

But note for any $\mathtt{a} \neq 0$:

$$\begin{aligned} \delta &\neq \delta^\star, \\ \alpha_1 &\neq \alpha_1^\star, \\ \alpha_2 &\neq \alpha_2^\star, \end{aligned}$$

$\boldsymbol{\theta} = [10, 5, 3]$ gives the same fit ($\mathtt{a}$=5) as $\boldsymbol{\theta}^\star = [5, 10, 8]$ with fitted values of $[15, 13]$ under both parameter sets.

Identifiability in a Model with Two Factors

In a two-way array of data Y_{ab}, with two factors and A and B with A and B levels, respectively, consider the main-effects only model,

$$\mathrm{E}[Y_{ab}] = \delta + \alpha_a + \beta_b,$$

for $a = 1, \ldots, A$, $b = 1, \ldots, B$; we have $1 + A + B$ parameters.

In the situation in which the two factors, we require two constraints to ensure identifiability.

Identifiability in a Model with Two Factors

The model is

$$\mathrm{E}[Y_{ab}] = \delta + \alpha_a + \beta_b,$$

for $a = 1, \ldots, A$, $b = 1, \ldots, B$.

The corner-point constraints are $\alpha_1 = \beta_1 = 0$.

The two sum-to-zero constraints are

$$\sum_{a=1}^{A} \alpha_a = \sum_{b=1}^{B} \beta_b = 0.$$

In both cases we have

$$1 + (A - 1) + (B - 1) = A + B - 1,$$

identifiable parameters.

The extension to more factors is immediate, so with 3 factors, we need 3 constraints, and the number of identifiable parameters is

$$1 + A - 1 + B - 1 + C - 1 = A + B + C - 2.$$

Interpretation for Single Factor Model

In the single factor model

$$\mathrm{E}[Y_a] = \delta + \alpha_a,$$

for $a = 1, \ldots, A$, with the

- corner-point parameterization, δ is the mean response at level 1, and α_a is the mean difference between level a, $a = 2, \ldots, A$, and level 1.
- sum-to-zero parameterization, δ is the overall mean response, and α_a is the difference between the mean for level a, $a = 1, \ldots, A$, and the overall mean.

Interpretation for Two Factor Model

Now we consider interpretation in the main effects only two factor model:

$$\mathrm{E}[Y_{ab}] = \delta + \alpha_a + \beta_b.$$

Under the:

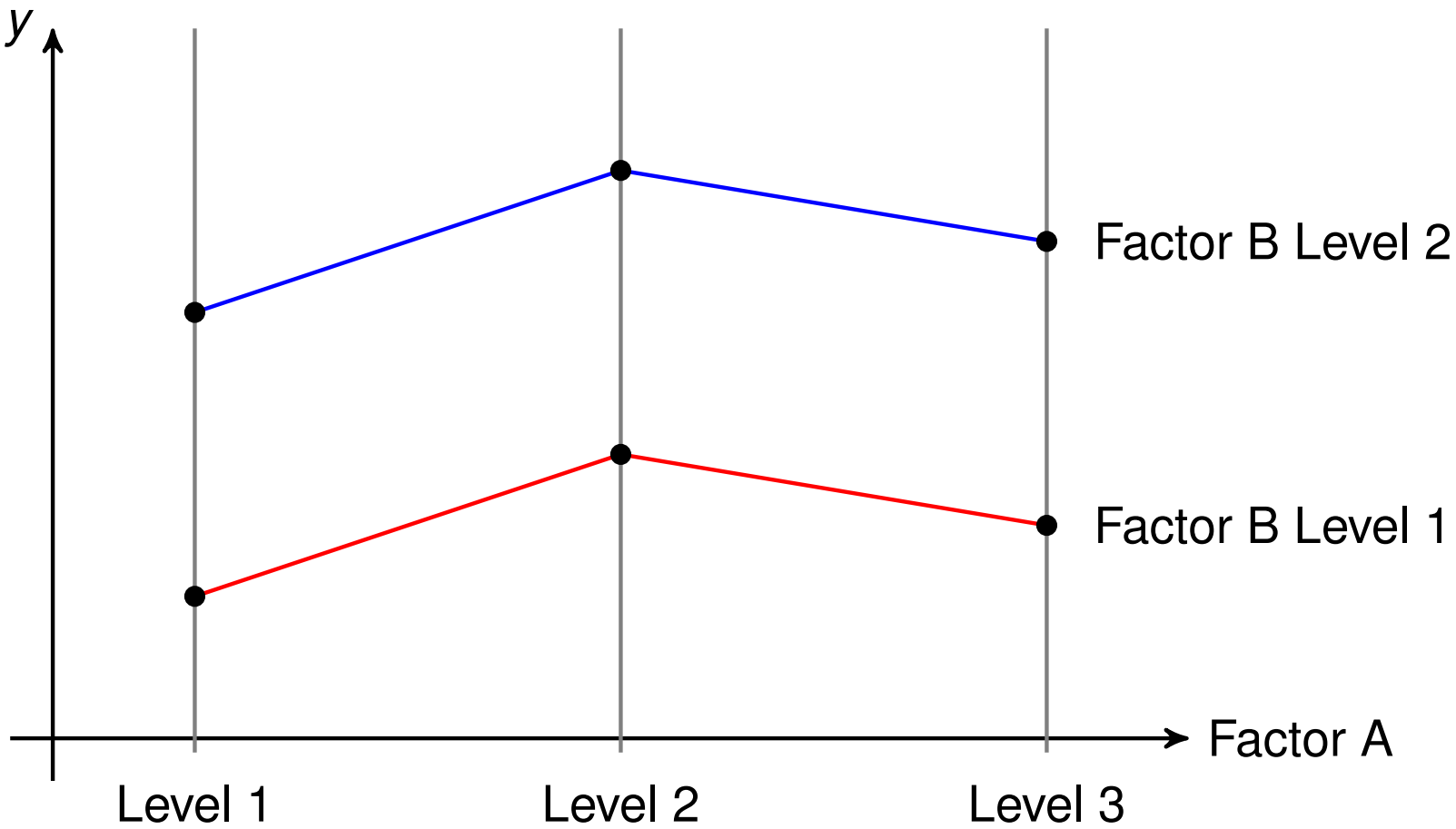
- Corner-point parameterization, δ is the mean response at the first level, and α_a is the mean difference between level a, $a = 2, \ldots, A$, and level 1 (regardless of the level of B) and β_b is mean difference between level b, $b = 2, \ldots, B$, and level 1 (regardless of the level of A).
- Sum-to-zero parameterization, δ is the overall mean response, and α_a is the difference between the mean for level a, $a = 1, \ldots, A$, and the overall mean (regardless of the level of B) and β_b is the difference between the mean for level b, $b = 1, \ldots, B$, and the overall mean (regardless of the level of A).

Interpretation for Two Factor Model

To emphasize: being in column b (a) is associated with a change in the mean response of β_b (α_a), regardless of the level of factor A (B), i.e., there is no interaction.

A plot of the responses versus factor A (say) levels, with separate lines for each level of factor B, across levels of factor A, will be parallel if the main effects model is appropriate.

Cartoon of No Interaction between Factors A and B



- The parallel lines indicate that a main effects only model is sufficient to describe these data.
- Could flip this around so Factor B on the x-axis — Initial Plots 1 and 3 do this with age and period as factors, and Initial Plots 2 and 4 do this with age and cohort as factors.

Concluding Remarks

- Begin by getting the required time at risk and tabulating the data.
- The Lexis diagram is a useful summary.
- The four plots we have described are initial explorations of whether rates are proportional between periods or cohorts.
- Identifiability is ubiquitous in factor models, but as we will see is much more difficult to conceptualize/deal with in APC models.

References

Berzuini, C. and Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, **13**, 823–838.

Carstensen, B. (2007). Age–period–cohort models for the lexis diagram. *Statistics in Medicine*, **26**, 3018–3045.

Clayton, D. and Schifflers, E. (1987a). Models for temporal variation in cancer rates. I: age–period and age–cohort models. *Statistics in medicine*, **6**, 449–467.

Clayton, D. and Schifflers, E. (1987b). Models for temporal variation in cancer rates. II: age–period–cohort models. *Statistics in medicine*, **6**, 469–481.

Fienberg, S. and Mason, W. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, pages 1–67.

Frost, W. (1939). The age selection of mortality from tuberculosis in successive decades. *American Journal of Hygiene*, **30**, 91–96. Reprinted in *American Journal of Epidemiology* 141:4–9, 1995.

Greenberg, B., Wright, J., and Sheps, C. (1950). A technique for analyzing some factors affecting the incidence of syphilis. *Journal of the American Statistical Association*, **45**, 373–399.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, **39**, 311–324.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, **332**, 487–509.

Keiding, N. (2011). Age–period–cohort analysis in the 1870s: Diagrams, stereograms, and the basic differential equation. *Canadian Journal of Statistics*, **39**, 405–420.

Kuang, D., Nielsen, B., and Nielsen, J. (2008). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, **95**, 979–986.

Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. K.J. Trübner.

Martínez Miranda, M., Nielsen, B., and Nielsen, J. (2015). Inference and forecasting in the age–period–cohort model with unknown exposure with an application to mesothelioma mortality. *Journal of the Royal Statistical Society: Series A*, **278**, 29–55.

Mason, K., Mason, W., Winsborough, H., and Poole, W. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, **38**, 242–258.

Nielsen, B. and Nielsen, J. (2014). Identification and forecasting in mortality models. *The Scientific World Journal*, **2014**.

Osmond, C. and Gardner, M. (1982). Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, **1**, 245–259.

Rosenberg, P. and Anderson, W. (2011). Age-period-cohort models in cancer surveillance research: ready for prime time? *Cancer Epidemiology Biomarkers & Prevention*, **20**, 1263–1268.

Rosenberg, P. S. (2018). A new age-period-cohort model for cancer surveillance research. *Statistical methods in medical research*, pages 1–29.

Smith, T. and Wakefield, J. (2016). A review and comparison of age-period-cohort models for cancer incidence. *Statistical Science*, **32**, 165–175.

Springett, V. (1950). A comparative study of tuberculosis mortality rates. *Journal of Hygiene*, **48**, 361–395.

Yang, Y. and Land, K. C. (2013). *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman and Hall/CRC Press.