2020 SISCER: Age-Period-Cohort Modeling and Analysis
Lecture 2: Identification and Modeling

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Notation and Preliminaries

APC Notation

Age is time from birth and is usually discretized into single years or five-year intervals.

Period is the time at which the event of interest (incidence/mortality) was recorded. Discretized into single years, or groups of years.

Cohort is the time at which birth occurred. Recorded as a single year, or a collection of years.

Notation

Let a, p, c index age, period, cohort.

A key observation is that given any two of age-period-cohort, the third can be derived.

Different authors use different indexing:

- We can index in terms of calendar dates and years, e.g., $a = 70$, $p = 2005$, $c = 1935$. In this case, $c = p - a$.
- More usually (and we use this notation), each of the factors is indexed from 1 onwards in the usual way, i.e., $a = 1, \ldots, A$, $p = 1, \ldots, P$, $c = 1, \ldots, C$. In this case, $c = A - a + p$.

Note that the number of cohort levels is $C = A + P - 1$.

Tabulation

		Period				
		1	2	3	4	5
	5	1	2	3	4	5
	4	2	3	4	5	6
Age	3	3	4	5	6	7
	2	4	5	6	7	8
	1	5	6	7	8	9

Table 1: Here we display $A = 5$ age groups and $P = 5$ periods. Indices of age, period, and cohort for equal-width age groups and time intervals. The first row and last column give the distinct cohorts; 5 is in both so there are $C = A + P - 1 = 9$ cohorts.

The cohorts (which are distinguished by different colors) proceed up the diagonals, as in the Lexis diagram.

Notice that the first and last cohorts (with indices 1 and 9, respectively) have a single observation only, which has implications for model fitting (sampling variability).

Forecasting

- Suppose we wish to forecast rates for time periods **6–8** using a model with age, period and cohort terms and no interactions.
- The indices in **bold green** indicate the cohort effects that need to be forecasted to generate predictions for these time periods.
- The period effects **6–8** need to be forecast also.
- The age effects are already present.

		Period							
		1	2	3	4	5	**6**	**7**	**8**
	5	1	2	3	4	5	6	7	8
	4	2	3	4	5	6	7	8	9
Age	3	3	4	5	6	7	8	9	**10**
	2	4	5	6	7	8	9	**10**	**11**
	1	5	6	7	8	9	**10**	**11**	**12**

Table 2: Indices of the age, period, and cohort parameters for equal-width age groups and time intervals.

Models

Due to the dependence between a, p and c, we only need to index data and model parameters by two of the three; a is usually taken along with one of p or c.

Let Y_{ap} be the number of disease counts observed in period p and age group a; cohort is found from $c = A - a + p$ – we assume data are available at a yearly time scale.

Suppose

$$\mathrm{E}[Y_{ap}] = N_{ap}\lambda_{ap},$$

denotes the mean, where

- N_{ap} is the number of person years at risk.
- λ_{ap} is the rate per year of new events.

Interpretation in Simple Models

A Poisson model[1] is a natural starting point for the modeling of rates for a rare disease – the rate parameter is positive, so loglinear models are common.

To tie down interpretation, suppose first we have the age main effects only model:

$$\log \lambda_{ap} = \delta + \alpha_a,$$

for $a = 1, \ldots, A$, with $\alpha_1 = 0$ (corner-point constraint) for identifiability.

Then,

$$\frac{\mathrm{E}[Y|a = a^\star]}{\mathrm{E}[Y|a = 1]} = \frac{\exp(\delta + \alpha_{a^\star})}{\exp(\delta)} = \exp(\alpha_{a^\star})$$

so that $\exp(\alpha_{a^\star})$ is the relative rate (or rate ratio) comparing the disease rate in age $a^\star$ to the disease rate in the first age group — doesn't tell us about the absolute level, only relative to baseline.

[1] or an overdispersed version

Interpretation in Simple Models

Again consider the model

$$\log \lambda_{ap} = \delta + \alpha_a,$$

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

In anticipation of models we examine later, consider the first and second differences:

$$\begin{aligned}
\Delta\alpha_a &= \alpha_a - \alpha_{a-1} \\
\Delta^2\alpha_a &= \Delta\alpha_a - \Delta\alpha_{a-1} = \underbrace{(\alpha_a - \alpha_{a-1})}_{\text{Slope } a-1 \to a} - \underbrace{(\alpha_{a-1} - \alpha_{a-2})}_{\text{Slope } a-2 \to a-1} \\
&= \alpha_a - 2\alpha_{a-1} - \alpha_{a-2}
\end{aligned}$$

The first differences tell us the change in the relative (log) rates between consecutive ages – if +ve/-ve, the rate is increasing/decreasing.

The second differences tell us how the local slopes are changing – if +ve/-ve, the difference in slopes is increasing/decreasing.

Interpretation in Simple Models

To tie down interpretation, suppose first we have the age and period main effects only model:

$$\log \lambda_a = \delta + \alpha_a + \beta_p,$$

for $a = 1, \ldots, A, p = 1, \ldots, P$, with $\alpha_1 = \beta_1 = 0$ for identifiability.

Then,

$$\frac{\mathrm{E}[Y|a = a^\star, p = p^\star]}{\mathrm{E}[Y|a = 1, p = p^\star]} = \frac{\exp(\delta + \alpha_{a^\star} + \beta_{p^\star})}{\exp(\delta + \beta_{p^\star})} = \exp(\alpha_{a^\star})$$

so that $\exp(\alpha_{a^\star})$ is the relative rate comparing the disease rate in age $a^\star$ to the disease rate in the first age group, with the period held constant – but with no interaction, this age association is the same for all periods.

Age-Period and Age-Cohort Models

Age-Period Models

We now discuss the fitting of age-period models.

Obvious but key point: These models assume no cohort effects.

Let Y_{ap} be the number of disease counts observed in period p of age a and $\mathrm{E}[Y_{ap}] = N_{ap}\lambda_{ap}$ denote the mean, where N_{ap} is the number of person years at risk.

Various plausible models are listed in Table 3.

Form of $\log \lambda_{ap}$	Description	No of Parameters
α_a	Age (factor) effects only	A
$\alpha_a + \beta_p$	Age and period (factor) main effects	$A + P - 1$
$\alpha_a + \beta \times p$	Age and linear period effects*	$A + 1$
$\alpha_a + g(p)$	Smoother in period	Depends on smoother
$f(a) + g(p)$	Smoothers in age and period	Depends on smoother
$\alpha_a + \beta_p + \gamma_{ap}$	Interaction (factor) model	$A \times P$

Table 3: Age-period models, *known as the drift model.

Identifiability in the Age-Period Model

Consider the model

$$\log \lambda_{ap} = f(a) + g(p).$$

The functions $f(a)$ and $g(p)$ are not identifiable since we can write

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

for some constant k.

Hence, only contrasts (i.e., differences, or first derivatives) of $f(\cdot)$ and $g(\cdot)$ are identifiable.

This isn't a big deal, since we can "tie down" one of the curves (see later).

Identifiability in the Age-Period Model

As an example, in the linear model with

$$f(a) = k_1 + \alpha a$$

and

$$g(p) = k_2 + \beta p$$

we obtain,

$$\log \lambda_{ap} = \delta + \alpha a + \beta p,$$

and we cannot uniquely identify $f(a)$ and $g(p)$, since $\delta = k_1 + k_2$).

But differences such as

$$f(4) - f(2) = 2\alpha$$

or

$$g(6) - g(1) = 5\beta,$$

are identifiable.

Identifiability in the Age-Period Model

In other words, there is no way to determine from the data alone the absolute levels of $f(\cdot)$ and $g(\cdot)$.

In the `apc.fit` function in the `Epi` package, various parameterizations are available.

Identifiability can be obtained by (say) picking a reference period p_0 and taking $g(p_0) = 0$.

With this choice $f(a)$ is interpreted as the age-specific log rate in period p_0 and $g(p)$ as log rate ratios of period p as compared to p_0.

As an alternative we could pick a reference age a_0, and look at rates relative to this point.

Identifiability in the Age-Period Model

```
APv2 <- glm( D ~ factor(A) + relevel( factor(P), "1971.5" )
             + offset( log(Y) ), family=poisson, data=dfEpi )
```

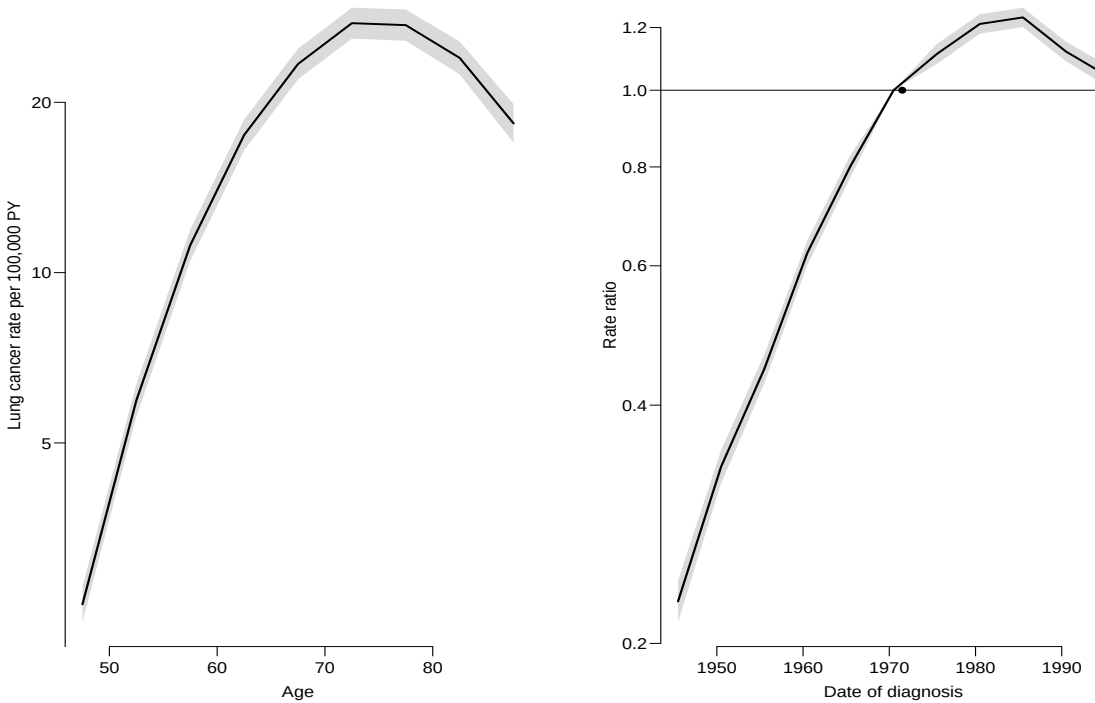

Figure 1: Left: Age levels interpretable as age-specific rates in period $p_0 = 1971.5$. Right: Period levels are interpretable as rate ratios of period p relative to $p_0 = 1971.5$. Shaded regions are 95% confidence intervals.

Age-Cohort Models

- Various age-cohort models can be fitted.
- Again obvious, but key point: These models assume no period effects.
- Let Y_{ac} be the number of disease counts observed in cohort c of age a and $\mathrm{E}[Y_{ac}] = N_{ac}\lambda_{ac}$ denote the mean, where N_{ac} is the number of person years at risk.
- Various plausible models are possible.

Form of $\log \lambda_{ac}$	Description	No of parameters
α_a	Age (factor) effects only	A
$\alpha_a + \gamma_c$	Age and period (factor) main effects	$A + C - 1$
$\alpha_a + \gamma \times c$	Age and linear period effects*	$A + 1$
$\alpha_a + h(c)$	Smoother cohort	Depends on smoother
$f(a) + h(c)$	Smoothers in age and cohort	Depends on smoother
$\alpha_a + \gamma_c + \nu_{ac}$	Interaction (factor) model	$A \times C$

Table 4: Age-cohort models, *known as the drift model.

Equivalence of Drift Models

Suppose the true model is linear in period p and cohort c:

$$\log \lambda_{ap} = \alpha_a + \beta p + \gamma c.$$

In an age-period model ($c = A - a + p$):

$$\begin{aligned} \log \lambda_{ap} &= \alpha_a + \beta p + \gamma(A - a + p) \\ &= \alpha_a + (\beta + \gamma)p + \gamma(A - a) \\ &= \underbrace{\alpha_a^\dagger}_{\alpha_a + \gamma(A-a)} + \underbrace{\beta^\dagger}_{\beta+\gamma} p \end{aligned}$$

In an age-cohort model ($p = c - A + a$):

$$\begin{aligned} \log \lambda_{ac} &= \alpha_a + \beta(c - A + a) + \gamma c \\ &= \alpha_a + (\beta + \gamma)c + \beta(A - a) \\ &= \underbrace{\alpha_a^\star}_{\alpha_a + \beta(A-a)} + \underbrace{\gamma^\star}_{\beta+\gamma} c \end{aligned}$$

So identical slopes are obtained since

$$\beta^\dagger = \gamma^\star = \beta + \gamma.$$

This is referred to as the net drift and is estimable.

Age-Period and Age-Cohort Models

Different age curves are generated by the two drift formulations, $\alpha_a^\dagger$ and $\alpha_a^\star$, so unless the age incidence relationship is known, the models are indistinguishable (Clayton and Schiffler, 1987, page 470).

Clayton and Schifflers (1987) refer to the age relationships estimated by the age-period model as cross-sectional age curves and the age-cohort model as longitudinal age curves.

A related discussion is the inability to estimate both cohort and longitudinal effects from cross-sectional data (Diggle *et al.*, 2002).

Note that the age-period and age-cohort models are not nested and so they cannot be tested against each other using likelihood ratio tests.

Drift Model for the Danish Male Lung Cancer Data

In the Danish lung cancer data there are 10 age groups, and 11 periods, hence 110 total cells.

The Age model has 10 distinct levels – age is given special status, since age is almost always a very important component, and a flexible model is used.

In Table 5 we take output from the `apc` function in the `Epi` package.

The drift model is a huge improvement over the Age only model.

	Resid. Df	Resid. Dev	Df	Deviance	p-value
Age	100	15103.0			
Age-drift	99	6417.4	1	8685.6	$< 2.2 \times 10^{-16}$

Table 5: Drift model for Danish male lung cancer data.

Period Drift Model for the Danish Lung Cancer Data

The fits, and residual deviance and degrees of freedom are identical under the two models (Age with Period Drift/Age with Cohort Drift), but the age effects are different in the two models[2].

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-55.0584	0.5093	-108.11	0.0000
as.factor(A)47.5	0.9497	0.0367	25.86	0.0000
as.factor(A)52.5	1.7936	0.0338	53.03	0.0000
as.factor(A)57.5	2.4405	0.0326	74.76	0.0000
as.factor(A)62.5	2.8947	0.0322	90.02	0.0000
as.factor(A)67.5	3.1809	0.0320	99.40	0.0000
as.factor(A)72.5	3.3428	0.0321	104.19	0.0000
as.factor(A)77.5	3.3315	0.0326	102.17	0.0000
as.factor(A)82.5	3.1951	0.0342	93.35	0.0000
as.factor(A)87.5	2.9305	0.0402	72.83	0.0000
P	0.0233	0.0003	90.70	0.0000

[2] and differences between age groups, so not just an intercept issue

Cohort Drift Model for the Danish Lung Cancer Data

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-54.0679	0.4984	-108.49	0.0000
as.factor(A)47.5	1.0663	0.0368	29.01	0.0000
as.factor(A)52.5	2.0266	0.0339	59.73	0.0000
as.factor(A)57.5	2.7901	0.0329	84.82	0.0000
as.factor(A)62.5	3.3608	0.0326	103.16	0.0000
as.factor(A)67.5	3.7636	0.0326	115.36	0.0000
as.factor(A)72.5	4.0420	0.0329	122.73	0.0000
as.factor(A)77.5	4.1472	0.0337	123.01	0.0000
as.factor(A)82.5	4.1274	0.0356	116.09	0.0000
as.factor(A)87.5	3.9793	0.0416	95.59	0.0000
Cterm	0.0233	0.0003	90.70	0.0000

Identifiability

Identifiability in the Age-Period-Cohort Model

Age, period, cohort effects can all be present in a particular population, the difficulty is in estimating the effects because of non-identifiability.

Non-identifiability occurs because, as already stated, given any two of age, period, cohort the third can be deduced.

The basic APC model is,

$$\log \lambda_{ap} = \eta_{ap} = \delta + \alpha_a + \beta_p + \gamma_c, \tag{1}$$

where the cohort index is $c = A - a + p$.

In this model, it is tempting to

- interpret δ as the overall log rate of incidence and
- to interpret differences in the age effects (α_a), period effects (β_p), or cohort effects (γ_c) as log relative rates.

However, direct interpretation of these effects is difficult because the model is over parameterized.

Models

A general main effects only (i.e., no interaction) model is

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

for functions $f(a)$, $g(p)$, $h(c)$, and where $c = A - a + p$.

Non-identifiable Slopes

As a hint of the identifiability problems to come, suppose we wish to estimate (log-)linear slopes:

$$\begin{aligned}
\log \lambda_{apc} &= \delta + \beta^{A} a + \beta^{P} p + \beta^{C} c \\
&= \delta + \beta^{A} a + \beta^{P} p + \beta^{C} (A - a + p) \\
&= \underbrace{\delta + \beta^{C} A}_{\text{Intercept}} + \underbrace{(\beta^{A} - \beta^{C})}_{\text{“Age” Slope}} a + \underbrace{(\beta^{P} + \beta^{C})}_{\text{“Period” Slope}} p \qquad (2)
\end{aligned}$$

so that the slopes β^{A}, β^{P} and β^{C} are not identifiable, due to the linear relationship between a, p, and c.

A Helpful Explanation?

Suppose we wish to estimate slopes and quadratic terms in the model:

$$
\begin{aligned}
\log \lambda_{apc} &= \delta + \beta^{A} a + \beta^{P} p + \beta^{C} c + \gamma^{A} a^2 + \gamma^{P} p^2 + \gamma^{C} c^2 \\
&= \delta + \beta^{A} a + \beta^{P} p + \beta^{C} (A - a + p) + \gamma^{A} a^2 + \gamma^{P} p^2 + \gamma^{C} (A - a + p)^2 \\
&= \underbrace{\delta + \beta^{C} A + \gamma^{C} A^2}_{\text{Intercept}} + \underbrace{(\beta^{A} - \beta^{C} - 2A\gamma^{C})}_{\text{“Age” Slope}} a + \underbrace{(\beta^{P} + \beta^{C} + 2A\gamma^{C})}_{\text{“Period” Slope}} p \\
&+ \underbrace{(\gamma^{A} + \gamma^{C})}_{\text{“Age” Quadratic}} a^2 + \underbrace{(\gamma^{P} + \gamma^{C})}_{\text{“Period” Quadratic}} p^2 - \underbrace{\gamma^{C}}_{\text{“Age-Period” Cross Term}} 2ap \qquad (3)
\end{aligned}
$$

so that the slopes β^{A}, β^{P} and β^{C} are not identifiable, due to the linear relationship between a, p, and c, but the quadratic terms γ^{A}, γ^{P} and γ^{C} terms are identifiable.

Because the cohort indices are a linear combination of the age and period indices, we can't identify linear terms in the model – more on this later, when we consider second difference models that are locally quadratic.

Identifiability in the Age-Period-Cohort Model

There are two sources of identifiability to consider.

The simpler one to account for is that which always occurs in models with factors: with an intercept in the model, we have one more level than is estimable and so three constraints are required.

A typical solution is to impose corner-point or sum-to-zero constraints.

The same thing happens with factors replaced by smoothers; Simon Wood recommends having an intercept and sum-to-zero constraints for each smoother:

`https://rdrr.io/cran/mgcv/man/identifiability.html`

Identifiability in the Age-Period-Cohort Model

The more insidious form of identifiability arises because of the linear dependence between the three factors:

$$c = A - a + p,$$

and there is no solution to this problem – in some situations in which non-identifibability arises (e.g., ecological inference) additional data can help, but not here...

Instead one must make assumptions if one wishes to directly interpret the parameters in equation (1), or not interpret these parameters, but only interpret those parameters that are identifiable.

Further, these assumptions are uncheckable from the raw data alone.

If we're just interested in forecasting then individual parameter interpretation not as critical but, as we will see, certain models produce forecasts that are invariant to the way that identifiability is overcome.

Identifiability in the Age-Period-Cohort Model

Suppose we group the intercept, age, period, and cohort effects into a single vector, $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta}^{\mathsf{T}} = [\, \delta, \boldsymbol{\alpha}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\gamma}^{\mathsf{T}} \,] = [\, \delta, \alpha_1, \ldots, \alpha_A, \beta_1, \ldots, \beta_P, \gamma_1, \ldots, \gamma_C \,], \qquad (4)$$

with $C = A + P - 1$.

We see that, for a suitably defined design matrix $\boldsymbol{x}$, the vector of log rates is $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\theta}$.

The matrix $\boldsymbol{x}$ is rank deficient in this case because the entries corresponding to the cohort effects are linearly dependent on the entries for the age and period effects and because of the general factor problem, as described above.

Identifiability in the Age-Period-Cohort Model

Thus, the full set of age, period, and cohort effects are not identifiable.

With respect to the parameter set

$$\boldsymbol{\theta}^{\mathsf{T}} = [\, \delta, \boldsymbol{\alpha}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\gamma}^{\mathsf{T}} \,] = [\, \delta, \alpha_1, \ldots, \alpha_A, \beta_1, \ldots, \beta_P, \gamma_1, \ldots, \gamma_C \,],$$

there are

$$\begin{aligned} 1 \quad &+ \quad A - 1 + P - 1 + C - 1 - 1 \qquad \text{the last because of } c = A - a + p \\ &= \quad 1 + A - 1 + P - 1 + (A + P - 1) - 1 - 1 \\ &= \quad 2(A + P) - 4 \end{aligned}$$

identifiable parameters.

Distinguish between ‘regular’ situations with 3 factors, e.g., age, gender, race.

Identifiability in the Age-Period-Cohort Model

Several authors, beginning with Fienberg and Mason (1979), have discussed the non-identifiability of the individual terms of the APC model.

Kuang *et al.* (2008b) and Nielsen and Nielsen (2014), following Carstensen (2007), define the identifiability issue from a group theoretic perspective.

Identifiability in the Age-Period-Cohort Model

The overall linear predictor

$$\eta_{ap} = \delta + \alpha_a + \beta_p + \gamma_c,$$

is

- invariant to a translation on each set of effects and
- addition of a linear trend in the age, period, and cohort parameters.

Identifiability in the Age-Period-Cohort Model

The group of transformations that give identical fits is

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

where

$$g\delta = \delta - \mathtt{a} - \mathtt{b} - \mathtt{c} - (A-1)\mathtt{d} \tag{5}$$

$$g\boldsymbol{\alpha} = \{\alpha_a + \mathtt{a} + (a-1)\mathtt{d}\}_{a=1}^{A} \tag{6}$$

$$g\boldsymbol{\beta} = \{\beta_p + \mathtt{b} - (p-1)\mathtt{d}\}_{p=1}^{P} \tag{7}$$

$$g\boldsymbol{\gamma} = \left\{ \gamma_{A-a+p} + \mathtt{c} + \underbrace{[(A-1) - (a-1) + (p-1)]}_{=A-a+p-1}\mathtt{d} \right\}_{a=1,p=1}^{a=A,p=P} \tag{8}$$

for any real numbers $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}$.

An interpretation of these numbers is that $\mathtt{a}$, $\mathtt{b}$, $\mathtt{c}$ are the overall levels of the age, period, cohort effects, respectively, and $\mathtt{d}$ is the linear trend.

Identifiability in the Age-Period-Cohort Model

The log rates are invariant with respect to these transformations.

Specifically, for any choices of the indices of the three factors a, b, c,

$$\begin{aligned}
\eta_{ap}(g\delta, g\alpha_a, g\beta_p, g\gamma_c) &= [\, \delta - \mathtt{a} - \mathtt{b} - \mathtt{c} - (A-1)\mathtt{d} \,] \\
&\quad + [\, \alpha_a + \mathtt{a} + (a-1)\mathtt{d} \,] \\
&\quad + [\, \beta_p + \mathtt{b} - (p-1)\mathtt{d} \,] \\
&\quad + \left[\, \gamma_{A-a+p} + \mathtt{c} + (A-a+p-1)\mathtt{d} \,\right] \\
&= \delta + \alpha_a + \beta_p + \gamma_{A-a+p} \\
&= \eta_{ap}(\delta, \alpha_a, \beta_p, \gamma_c).
\end{aligned}$$

Hence, for any g,

$$\begin{aligned}
\eta_{ap}(g\boldsymbol{\theta}) &= \eta_{ap}(g\delta, g\alpha_a, g\beta_p, g\gamma_c) \\
&= \eta_{ap}(\delta, \alpha_a, \beta_p, \gamma_c) \\
&= \eta_{ap}(\boldsymbol{\theta}).
\end{aligned}$$

Since the data likelihood only depends on the age, period, and cohort parameters through the log rates, it is also invariant to these groups of transformations.

Identifiability in the Age-Period-Cohort Model

To obtain identifiability, sum-to-zero constraints,

$$\sum_a \alpha_a = \sum_p \beta_p = \sum_c \gamma_c = 0,$$

are a common solution (another is corner point constraints).

The total number of non-identifiable parameters is 4, but the sum-to-zero constraints reduces this number by 3 (effectively fixing the values `a, b, c`).

This gives identifiability of the intercept δ.

But this does not solve the identifiability problem caused by the linear relationship between cohort, period and age.

This linear relationship means that separate linear associations with each of age, period and cohort are not identifiable.

We require one more constraint to produce an identifiable set.

Identifiability in the Age-Period-Cohort Model

One approach is to assume two period or two cohort effects are equal (Mason *et al.*, 1973).

For example, Clayton and Schifflers (1987) consider restricting the first differences of the period effects i.e.,

$$\beta_2 - \beta_1, \beta_3 - \beta_2, \ldots, \beta_P - \beta_{P-1}$$

to be zero on average, which is equivalent to the restriction $\beta_1 = \beta_P$.

Alternatively, one can restrict a sequential pair of effects to be equal (e.g., $\gamma_1 = \gamma_2$).

Holford (1991) rejects these approaches because the estimated effects will depend on which pair of effects are restricted, and generally there is no scientific rationale for choosing, say $\gamma_1 = \gamma_2$ over $\gamma_4 = \gamma_5$.

Identifiability of Age-Period-Cohort Model

One approach to the identifiability problem of APC models is to express the model only in terms of those functions of the age, period, and cohort parameters that are estimable.

For example, Holford (1983) partitions each set of effects into the linear effect and a curvature effect.

Linear combinations of the curvature effects (for example, the average) are estimable, and some functions of the slopes in the age, period, and cohort effects are estimable.

Identifiability of Age-Period-Cohort Model

Suppose again that β^A, β^P, and β^C, are the linear slopes of the age, period, and cohort effects.

Holford's model consists of factors for each of age, period and cohort – from these sets of factors, the estimable linear trends can be calculated.

The design matrix is then parameterized by the linear trends, and the remaining non-linear terms.

Under this parameterization, the curvature effects and linear combinations of the slopes of the form

$$\mathtt{u}\beta^A + \mathtt{v}\beta^P + (\mathtt{v} - \mathtt{u})\beta^C,$$

for different values of $\mathtt{u}, \mathtt{v}$ are estimable, as we show on the next slide.

As examples:

- setting $(\mathtt{u}, \mathtt{v}) = (1, 0)$ we see that $\beta^A - \beta^C$ is identifiable and
- setting $(\mathtt{u}, \mathtt{v}) = (0, 1)$ shows $\beta^P + \beta^C$ is identifiable,

as we saw earlier in (2).

Identifiability of Age-Period-Cohort Model

As discussed above, the log rates are invariant to the addition of a linear trend $\mathtt{d}$ to the age and cohort effects and subtraction of $\mathtt{d}$ from the period effects, i.e.,

$$\begin{aligned}\beta^{A^\star} &= \beta^A + \mathtt{d},\\ \beta^{P^\star} &= \beta^P - \mathtt{d},\\ \beta^{C^\star} &= \beta^C + \mathtt{d}\end{aligned}$$

Then,

$$\begin{aligned}\mathtt{u}\beta^{A^\star} + \mathtt{v}\beta^{P^\star} + (\mathtt{v}-\mathtt{u})\beta^{C^\star} &= \mathtt{u}\beta^A + \mathtt{v}\beta^P + (\mathtt{v}-\mathtt{u})\beta^C + \mathtt{u}\mathtt{d} - \mathtt{v}\mathtt{d} + (\mathtt{v}-\mathtt{u})\mathtt{d}\\ &= \mathtt{u}\beta^A + \mathtt{v}\beta^P + (\mathtt{v}-\mathtt{u})\beta^C.\end{aligned}$$

Hence, these functions of the linear slopes are invariant to any transformation g.

Identifiability of Age-Period-Cohort Model

Rosenberg and Anderson (2011) showed that many epidemiological summaries, such as longitudinal or cross sectional age trends, can be expressed as estimable functions of the parameters in Holford's APC model.

These summaries are available in a user-friendly web tool from the National Cancer Institute:

`http://analysistools.nci.nih.gov/apc/`

The `R` code is here:

`https://github.com/CBIIT/nci-webtools-dceg-age-period-cohort/blob/master/apc/apc.R`

Models for the Danish Male Lung Cancer Data

In the `Epi` package, factor or spline models can be fitted.

Table 6 gives summaries from the factor model, via the call:

```
apc.fit( dfEpi, model="factor",parm="ACP")
```

	Resid. Df	Resid. Dev	Df	Deviance	p-value
Age	100	15103.0			
Age-drift	99	6417.4	1	8685.6	$< 2.2 \times 10^{-16}$
Age-Cohort	81	829.6	18	5587.8	$< 2.2 \times 10^{-16}$
Age-Period-Cohort	72	208.5	9	621.1	$< 2.2 \times 10^{-16}$
Age-Period	90	2723.5	-18	-2514.9	$< 2.2 \times 10^{-16}$
Age-drift	99	6417.4	-9	-3693.9	$< 2.2 \times 10^{-16}$

Table 6: Factor models for Danish male lung cancer data.

Models for the Danish Male Lung Cancer Data

- Everything is significant when added – sample size is very big here, so complex models favored.
- We begin with Age as the null model, with period and cohort giving highly significant reductions in the deviance.
- Informally (AIC?) cohort appears to have a stronger association than period.
- Holford (1991) (among others) discusses cohort effects for lung cancer.
- Figure 2 clearly illustrates the identifiability associated with the APC models!

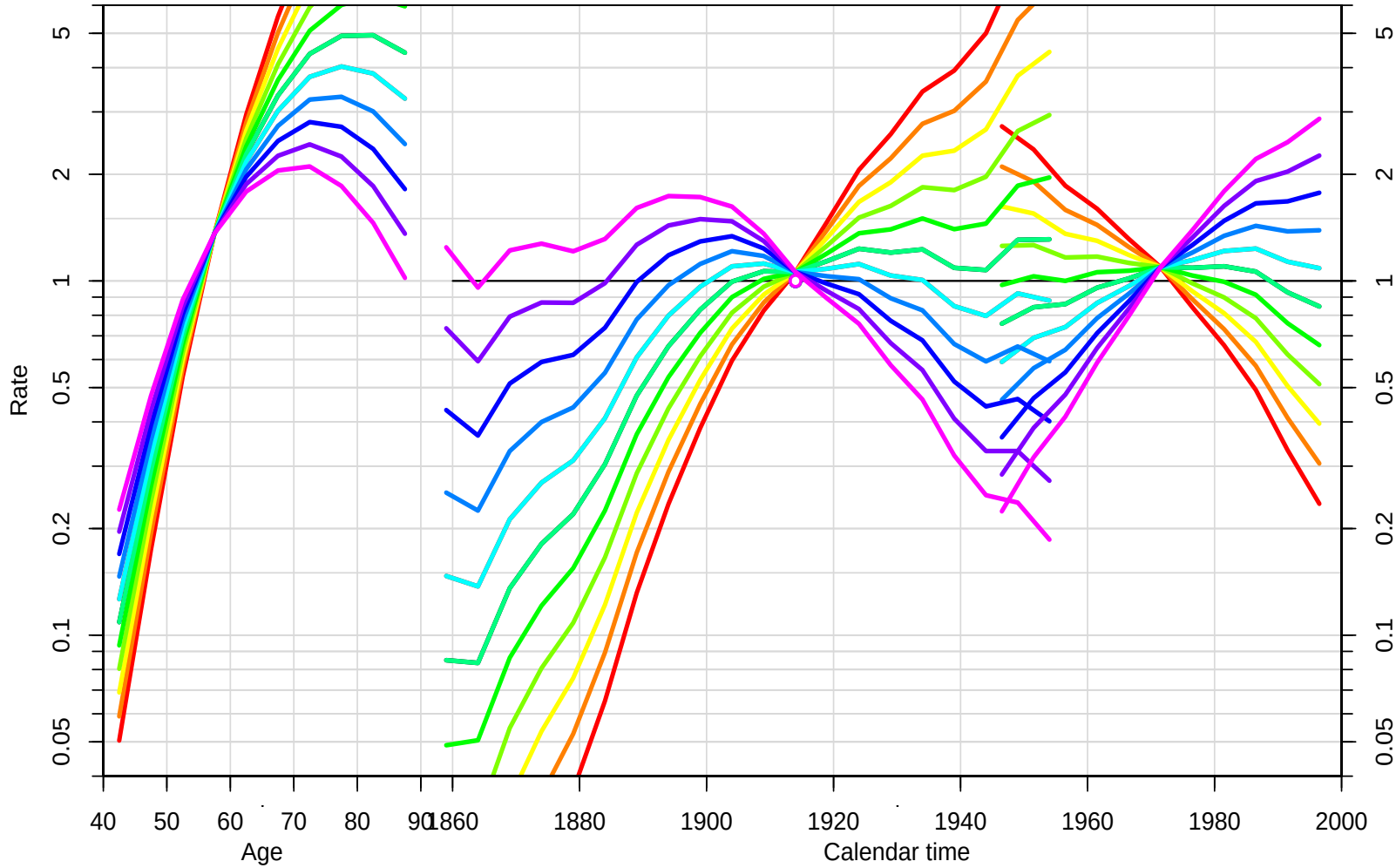

Figure 2: Age-period-cohort estimates from the factor model. Curves with added annual period drifts of -4%, -3%,. . . ,4% are also shown. The rates predicted from curves of like colors are the same.

Identifiability in the Age-Period-Cohort Model

Alternatives to using all of age, period, cohort include using simpler two-factor models or a two-factor model with a predictor ('characteristic') in place of the third factor (O'Brien, 2000).

For example, a plausible model for lung cancer rates may include age and period factors and the smoking rate as a linear predictor.

Replacing the effects of one time scale with an explanatory variable (here, smoking) is a good alternative to the full age-period-cohort model in simple problems where the disease generating process is well understood.

However, access to the relevant data may be an issue, and forecasting disease rates would require forecasts of explanatory variables.

A Canonical Parameterization

A Canonical Parameterization

Kuang *et al.* (2008b) suggested a related parameterization of APC models to that of Holford, which we now discuss.

Kuang *et al.* (2008a), Kuang *et al.* (2008b), Nielsen and Nielsen (2014) and Martínez Miranda *et al.* (2015) parameterize the APC model in terms of three initial log rates and the full set of second differences for data with equal-width age and time intervals.

Kuang *et al.* (2008b) construct a mapping from the rates at three initial time points using age-cohort indices (i.e., η_{ac} instead of η_{ap}).

We focus on a parametrization in Martínez Miranda, Nielsen and Nielsen (2014) based on age-period (*ap*) indexing.

We refer to this as the MMNN model.

A Canonical Parameterization

The parameter set consists of

- three sets of second differences, and
- three points η_{ap} that are chosen to identify the shared level and the linear trend.

Hence, the parameter set is

$$\begin{aligned}\boldsymbol{\theta} \quad = \quad & [\, \eta_{A1}, \eta_{A1} - \eta_{A-1,1}, \eta_{A2} - \eta_{A1}, \\ & \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \\ & \Delta^2\beta_3, \ldots, \Delta^2\beta_P, \\ & \Delta^2\gamma_3, \ldots, \Delta^2\gamma_{A+P-1} \,]\end{aligned}$$

which is of length $2(A + P) - 4$, as required.

Kuang *et al.* (2008b) show that the parameter $\boldsymbol{\theta}$ is identifiable in that

$$\eta(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta}^\star)$$

only if

$$\boldsymbol{\theta} = \boldsymbol{\theta}^\star.$$

A Canonical Parameterization

Next, the link between the parameter vector and the log rate is derived.

From Theorem 1 of Martínez Miranda *et al.* (2015), the overall log rate can be written as

$$\eta_{ap} = \underbrace{\eta_{A1}}_{\text{Overall Level}} + \underbrace{(a-A)(\eta_{A1}-\eta_{A-1,1})}_{\text{Linear Trend}} + \underbrace{(p-1)(\eta_{A2}-\eta_{A1})}_{\text{Linear Trend}} + \underbrace{\sum_{t=a}^{A-2}\sum_{s=t}^{A-2}\Delta^2\alpha_{s+2} + \sum_{t=3}^{p}\sum_{s=3}^{t}\Delta^2\beta_s + \sum_{t=3}^{A-a+p}\sum_{s=3}^{t}\Delta^2\gamma_s}_{\text{Time Effects}} \tag{9}$$

A Canonical Parameterization

This means that, for a design matrix $\boldsymbol{x}$, we can write $\boldsymbol{\eta} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is,

$$[\,\eta_{A1}, \eta_{A1} - \eta_{A-1,1}, \eta_{A2} - \eta_{A1}, \Delta^2\alpha_3, \ldots, \Delta^2\alpha_A, \Delta^2\beta_3, \ldots, \Delta^2\beta_P, \Delta^2\gamma_3, \ldots, \Delta^2\gamma_{A+P-1}\,].$$

The entries in $\boldsymbol{x}$ are equal to the multiplicative factors in (9).

For example, for $A = P = 3$ (so that $C = 5$), the mapping is

$$\begin{pmatrix} \eta_{11} \\ \eta_{12} \\ \eta_{13} \\ \eta_{21} \\ \eta_{22} \\ \eta_{23} \\ \eta_{31} \\ \eta_{32} \\ \eta_{33} \end{pmatrix} = \begin{pmatrix} 1 & -2 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & -2 & 1 & 1 & 0 & 2 & 1 & 0 \\ 1 & -2 & 2 & 1 & 1 & 3 & 2 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 2 & 0 & 1 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{31} \\ \eta_{31} - \eta_{21} \\ \eta_{32} - \eta_{31} \\ \Delta^2\alpha_3 \\ \Delta^2\beta_3 \\ \Delta^2\gamma_3 \\ \Delta^2\gamma_4 \\ \Delta^2\gamma_5 \end{pmatrix}.$$

A Canonical Parameterization

Note this parametrization has four fewer parameters than the model with an overall rate and the full set of age, period and cohort effects.

This makes sense because the group transformation g, given by (5)–(8), is defined by four real numbers.

The parameters θ are easily estimated via standard Poisson regression where the columns of $\boldsymbol{x}$ are treated as the predictors.

This can be done directly or using the `apc` package (Nielsen, 2014).

Just as the set of estimable functions for the linear trends in Holford's model is infinite, the choice of the three initial points in the Nielsen parameterizations (which are, equivalently, functions of first differences) is not unique.

A Canonical Parameterization

- Instead of

$$\{\eta_{A1}, \eta_{A-1,1}, \eta_{A2}\}$$

we can choose any three

$$\{\eta_{i_1 j_1}, \eta_{i_2 j_2}, \eta_{i_3 j_3}\}$$

as long as the indices of the three points are not linearly dependent (see Corollary 1 in Kuang *et al.* (2008b)).

- Nielsen and Nielsen (2014) choose initial points based on the median age and cohort levels and not on the extremes, as in earlier papers.

- Using this guideline, as an example, the baseline rates reflect the rates in the middle of Table 7 rather than the corners.

		Period				
		1	2	3	4	5
Age	5	1	2	3	4	5
	4	2	3	4	5	6
	3	3	4	5	6	7
	2	4	5	6	7	8
	1	5	6	7	8	9

Table 7: There are $A = 5$ age groups and $P = 5$ periods. Indices of age, period, and cohort for equal-width age groups and time intervals.

Identifiability of Age-Period-Cohort Model

The second differences (second derivatives, accelerations) in each of age, period and cohort are identifiable, see Holford (1983); Clayton and Schifflers (1987) and Kuang *et al.* (2008b).

This is somewhat surprising when first encountered!

We may write each of the functions as polynomials, for example:

$$f(a) = \underbrace{f(a_0)}_{\textit{Intercept}} + \underbrace{(a-a_0)f'(a_0)}_{\textit{Linear Term}} + \underbrace{(a-a_0)^2 f''(a_0)}_{\textit{Quadratic Term}} + \cdots$$

Identifiability of Age-Period-Cohort Model

Knowing the second derivatives[3], does not uniquely identify the linear part of the function, or the level.

For example, for the quadratic,

$$f(a) = Aa^2 + Ba + C,$$

the second derivative is

$$f''(a) = A,$$

which is consistent with any B and C, i.e., with any slope or intercept.

[3] or second differences

Identifiability of Age-Period-Cohort Model

Now suppose $f(a) = \alpha_a$.

The first differences are

$$\Delta\alpha_a = \alpha_a - \alpha_{a-1},$$

and eliminates (doesn't tell us anything about) the level (intercept).

The second differences are

$$\begin{aligned}
\Delta^2\alpha_a &= \Delta(\alpha_a - \alpha_{a-1}) \\
&= \Delta\alpha_a - \Delta\alpha_{a-1} \\
&= \underbrace{(\alpha_a - \alpha_{a-1})}_{\text{Slope is in here}} - \underbrace{(\alpha_{a-1} - \alpha_{a-2})}_{\text{Slope is in here}} \\
&= \alpha_a - 2\alpha_{a-1} + \alpha_{a-2}
\end{aligned}$$

and eliminates (doesn't tell us anything about) linear trends.

This helps to understand why the level and linear trends can be non-identifiable, but the second differences be estimable.

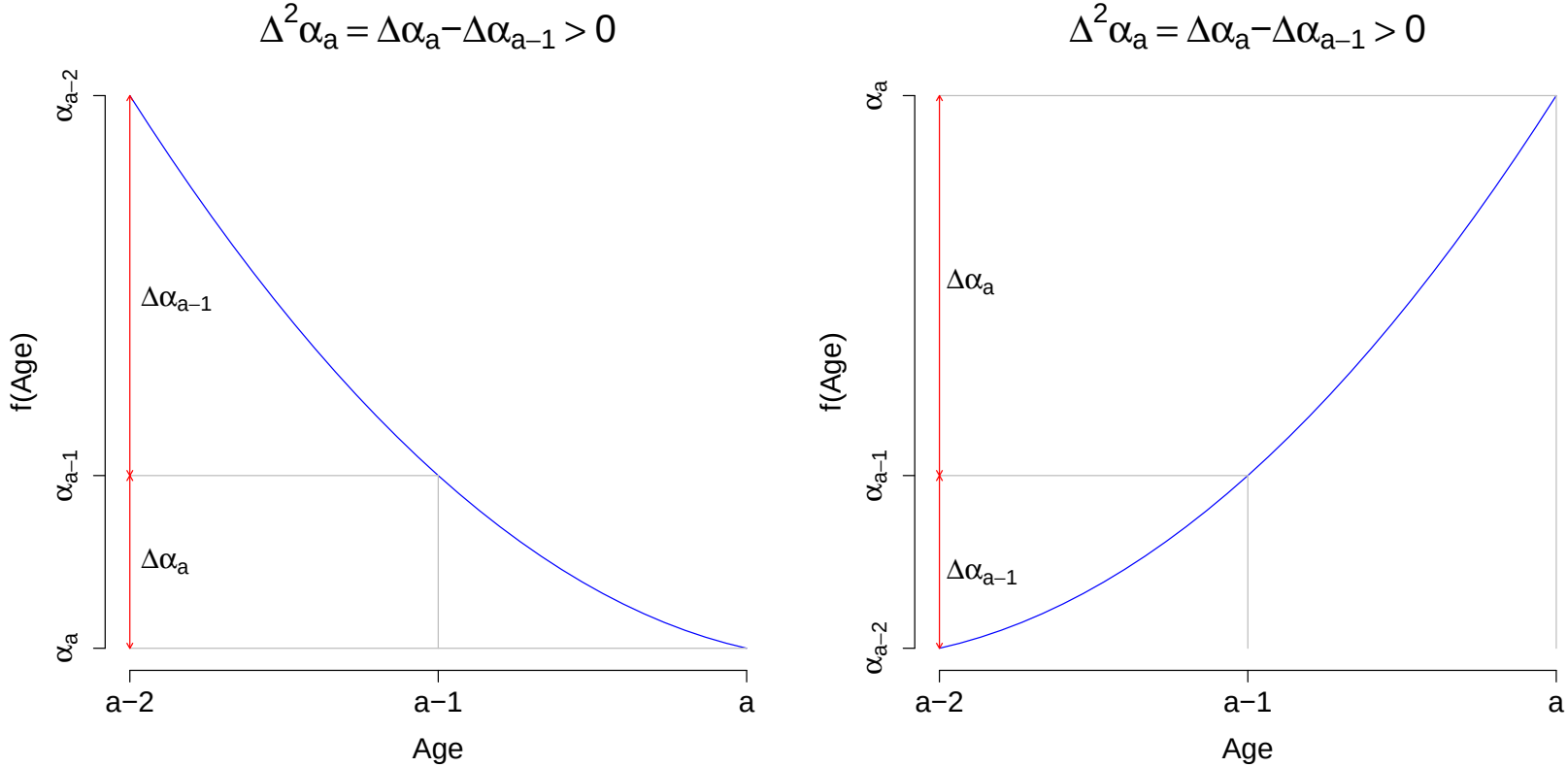

Figure 3: Curves with $\Delta^2\alpha_a > 0$.

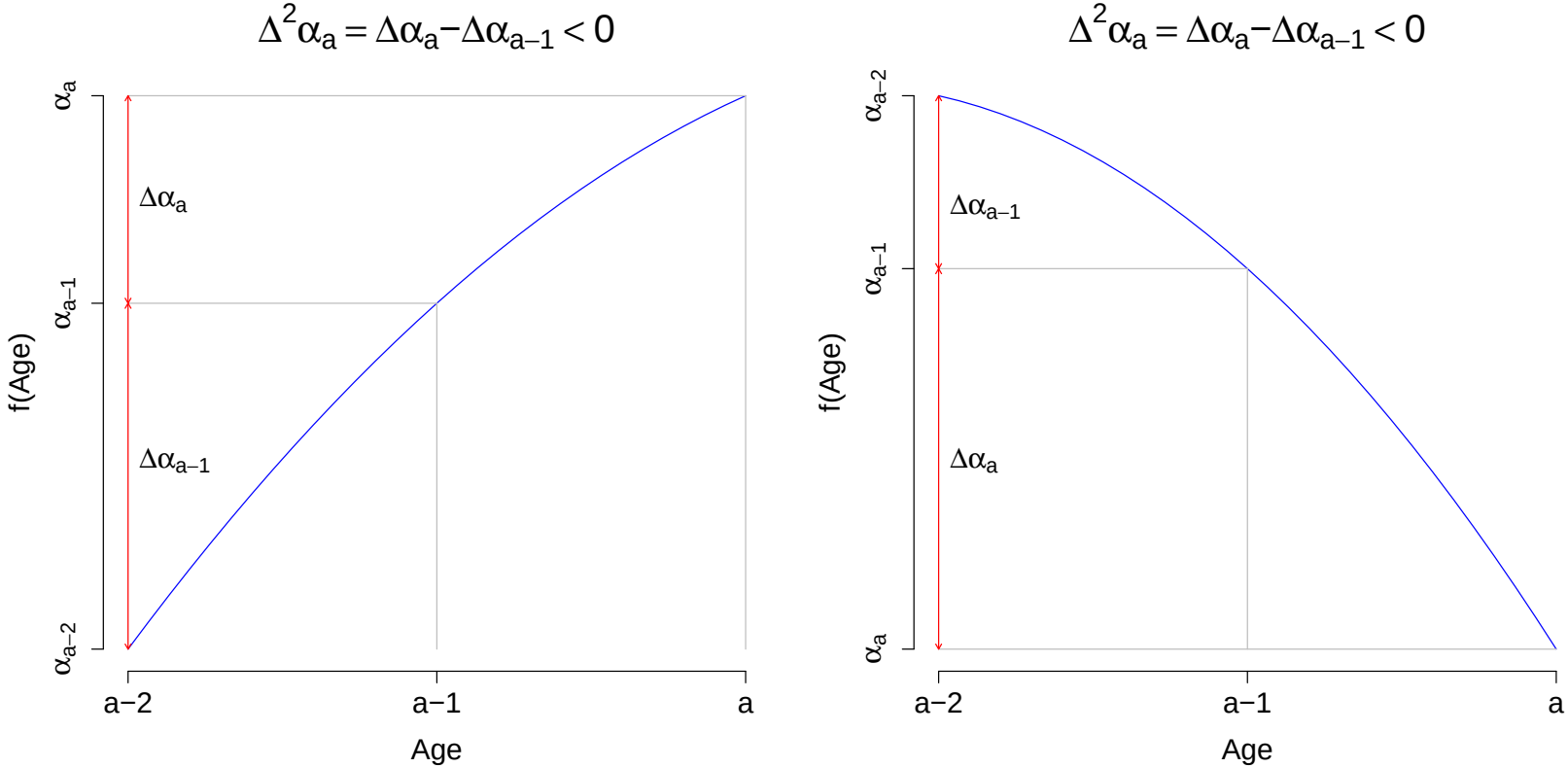

Figure 4: Curves with $\Delta^2\alpha_a < 0$.

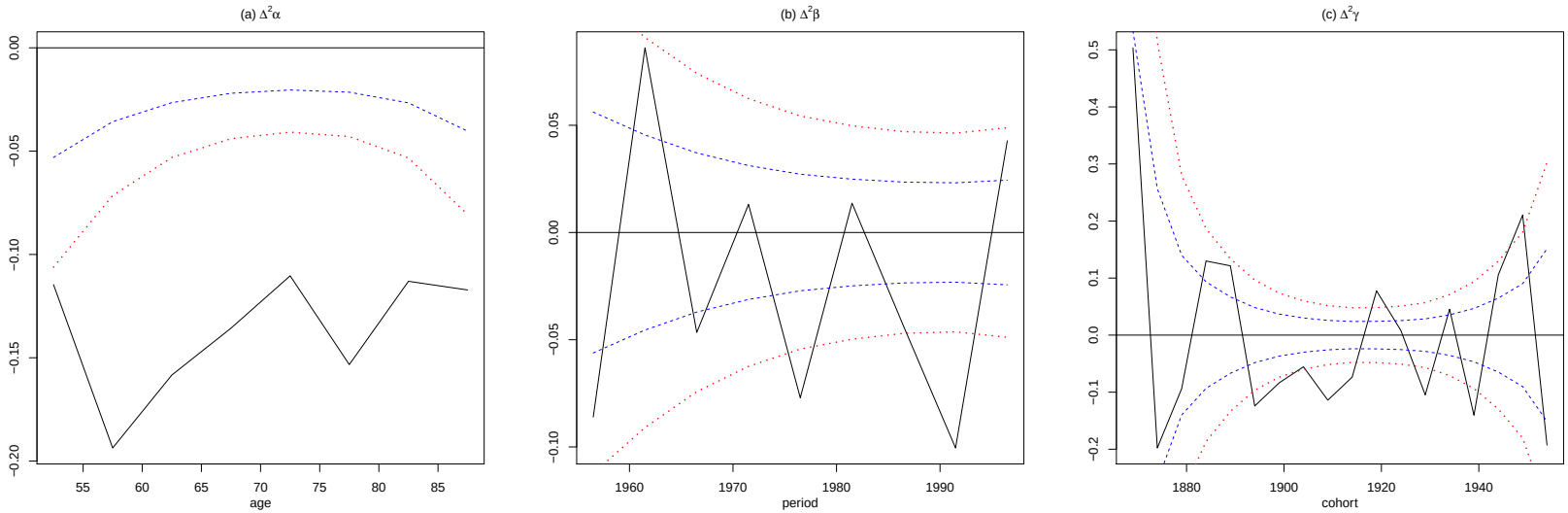

Figure 5: Estimated second differences for the Danish lung cancer data: (a) Age, (b) Period, (c) Cohort. The blue and red lines are at 1 and 2 standard errors from zero.

The period and cohort second differences are not so different from zero, while the age second derivatives are all negative which is consistent with a slowing down of the age effect.

Conclusions

- Age-Period-Cohort models suffer from serous identifiability problems which mean that unless uncheckable assumptions are made, levels and linear trends in each of the three components are not interpretable.
- APC models can be used for producing fits and forecasts, which are an important use.
- Second order terms are interpretable.
- The factor models we have seen so far do not acknowledge the temporal ordering of the levels.
- The Bayesian models we examine in the next lecture recognize the ordering.
- Spline models provide a parsimonious way of encouraging smoothness of rates that are close.

References

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

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

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, second edition.

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.

Holford, T. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health*, **12**, 425–457.

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

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics*, **7**, 401–416.

Kuang, D., Nielsen, B., and Nielsen, J. (2008a). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika*, **95**, 987–991.

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

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. (2014). *apc: A Package for Age-Period-Cohort Analysis*. R package version 1.0.

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

O'Brien, R. (2000). Age period cohort characteristic models. *Social Science Research*, **29**, 123–139.

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.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

Cubic Smoothing Splines

Proof: The proof has two parts, and is based on Green and Silverman (1994, Chapter 2).

We begin by showing that a natural cubic spline minimizes (**??**) amongst all interpolating functions, and then extend to non-interpolating functions.

Assume that $x_1 < ... < x_n$. We consider all functions that are continuous in $[x_1, x_n]$, with continuous first and second derivatives, and which interpolate (x_i, y_i), $i = 1, \ldots, n$.

Since the first term of (**??**) is zero we need to show that the natural cubic spline, $g(x)$, minimizes

$$\int_{x_1}^{x_n} f''(x)^2 dx.$$

Cubic Smoothing Splines

Let $\widetilde{g}(x)$ be another interpolant of (x_i, y_i), and define $h(x) = \widetilde{g}(x) - g(x)$. Then

$$\begin{aligned}\int_{x_1}^{x_n} \widetilde{g}\,''(x)^2 dx &= \int \{g''(x) + h''(x)\}^2 dx \\ &= \int g''(x)^2 dx + 2\int g''(x)h''(x)dx + \int h''(x)^2 dx.\end{aligned}$$

Applying integration by parts to the cross term:

$$\begin{aligned}\int_{x_1}^{x_n} g''(x)h''(x)dx &= \left[g''(x)h'(x)\right]_{x_1}^{x_n} - \int_{x_1}^{x_n} g'''(x)h'(x)dx \\ &= -\int_{x_1}^{x_n} g'''(x)h'(x)dx \quad \text{since } g''(x_1) = g''(x_n) = 0 \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+)\int_{x_i}^{x_{i+1}} h'(x)dx \\ &\quad \text{since } g'''(x) \text{ is constant in, and } x_i^+ \text{ is a point in, } (x_i, x_{i+1}) \\ &= -\sum_{i=1}^{n-1} g'''(x_i^+)\{h(x_{i+1}) - h(x_i)\} \\ &= 0\end{aligned}$$

since $h(x_{i+1}) = \widetilde{g}(x_{i+1}) - g(x_{i+1})$, and both are interpolants (also for $h(x_i)$).

Cubic Smoothing Splines

We have shown that

$$\begin{aligned}\int_{x_1}^{x_n} \widetilde{g}''(x)^2 dx &= \int_{x_1}^{x_n} g''(x)^2 dx + \int_{x_1}^{x_n} h''(x)^2 dx \\ &\geq \int_{x_1}^{x_n} g''(x)^2 dx\end{aligned}$$

with equality if and only if $h''(x) = 0$ for $x_1 < x < x_n$. The latter implies $h(x) = a + bx$, but $h(x_1) = h(x_n) = 0$ and so $a = b = 0$.

Hence, any interpolant that is not identical to $g(x)$ will have a higher integrated squared second derivative.

Therefore, the natural cubic spline with knots at the unique x values is the smoothest interpolant in the sense of minimizing $\int f''(x)^2 dx$. This is of use in, for example, numerical analysis, where interpolation of (x_i, y_i) is of interest.

But in statistical applications, the data are measured with error, and we typically do not wish to restrict attention to interpolating functions.

Cubic Smoothing Splines

We have shown that a natural cubic spline minimizes (**??**) amongst all interpolating functions, but the minimizing function need not necessarily be an interpolant, since an interpolating function may have a large associated penalty contribution.

The second part of the proof therefore considers functions that do not necessarily interpolate the data but have n free parameters $g(x_i)$ with the aim being minimization of (**??**). The resulting $g(x)$ is known as a *smoothing spline*.

Suppose some function, $f^\star(x)$, other than the cubic smoothing spline minimizes (**??**). Let $g(x)$ be the natural cubic spline that interpolates $(x_i, f^\star(x_i))$, $i = 1, ..., n$. Obviously $f^\star$ and g produce the same residual sum of squares in (**??**) since $f^\star(x_i) = g(x_i)$. But by the first part of the proof

$$\int f^{\star''}(x)^2 dx > \int g''(x)^2 dx.$$

Hence the natural cubic spline is the function that minimizes (**??**); this spline is known as a *cubic smoothing spline*.