Age-Period-Cohort model in a Dirichlet framework: A coherent causes of death estimation.

SEMinario in Scienze Economiche e Sociali.
DEMET, Univeristy of Foggia

Andrea Nigri [1],Rebecca Graziani [2], Marco Bonetti [2]

March 15, 2023

[1] Department of Economics, Management and Territory, University of Foggia
[2] Department of Social and Political Sciences, Bocconi University

Longevity risks modeling...The 'so-what?' question.

Steady improvement in mortality level

SPOILER: This is a methodological paper ... We will not be dealing with socio-economic implications!

let's start...Outline

Outline

- Background & Motivation
- APC - Methodological framework
- Model
- Data
- Results
- Conclusion and further directions

Background & Motivation

- Demographic studies often require the modeling of multiple outcomes.
- Many tasks need the different outcomes to be treated jointly in a unified framework.
- The overall mortality can be described as the composition of several causes of death (CoDs).
- Here we shall focus on the case of compositional data (non-negative proportions with unit-sum).

Our proposal

- For compositional data, Dirichlet distribution represents the best solution.
- Dirichlet regression provides a GLM-like framework that relates compositional data to other relevant variables.
- We propose an Age-Period-Cohort (A-P-C) model within the Dirichlet framework, with specific interest in its use for modeling longevity with multiple causes of death modeling.

APC - Metodologial framework

In the OLS regression we can write the APC model that uses categorical coding for ages,periods, and cohorts as follows:

$$Y_{ij} = \mu + a_i + p_j + c_k + \epsilon_{ij},$$

- Singular "design matrix" - cannot be inverted
- The lack of identifiability of the parameters in the classic APC model (Glenn (1976); Holford (1983); Wilmoth (1990); Y. Yang, Fu, and Land (2004); and Nielsen (2008); and O'Brien (2011))

The first attempts at fitting the full APC model to create an idetifiale parametrization using a sum-to-zero constrain:

$$\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} p_j = \sum_{k=1}^{K} c_k = 0.$$

The problem with such constraints is that the results can differ substantially depending on the constraint chosen. **Solution**: Mixed models.

Why use Mixed model

O'Brien, R. M. (2017) Mixed models, linear dependency, and identification in age-period-cohort models. Statististics in Medicine:
"Mixed models are identified, without introducing an additional constraint."

- Model identification by constraining the solution through the shrinkage associated with random effects.
- This shrinkage typically constrains the trend of one of the random factors to be near zero and that this determines the slope of the other trends
- Similar linear trends, no matter what combination of fixed and random factors are used.
- External constraint need to be justified on theoretical/substantive grounds.

Model

Dirichlet distribution

- Dirichlet distribution: the most natural distribution for working with compositions (Grunwald, Raftery and Guttorp (1993)).
- Working directly in the simplex allows for an easy interpretation of the behavior of the components - Unlike the log-ratio approaches in Aitchison (1986).

It is the generalization of the widely known Beta distribution, and it is defined by the following joint probability density function

$$p\left(y_1, \ldots, y_{D-1} \mid \boldsymbol{\alpha}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{d=1}^{D} y_d^{\alpha_d - 1} \tag{1}$$

$$\sum_{d=1}^{D} y_d = 1, \quad y_d \geqslant 0$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^T \in \mathbb{R}_{>0}^D$ is the vector of shape (or concentration) parameters for each category, $\sum_{d=1}^{D} y_d = 1, \quad y_d \geqslant 0$, where

Let $\boldsymbol{Y} \sim \mathcal{D}(\boldsymbol{\alpha})$ denote the random vector that follows the Dirichlet distribution (with $\boldsymbol{Y} = (Y_1, \ldots, Y_D)^T$). Then, the value y_d of Y_d can be interpreted as the probability that an event will fall in category d.

- The shape parameter $\boldsymbol{\alpha}$, controls both the location and dispersion of the distribution.
- To isolate the effects of the location and dispersion, we use the approach of Grunwald, Raftery and Guttorp (1993) by reparameterizing the distribution with location, $\boldsymbol{\theta}$, and scale, $\boldsymbol{\tau}$, parameters.

Let $\mathbf{Y} \sim \text{Dirichlet}(\boldsymbol{\alpha} = \boldsymbol{\tau}\boldsymbol{\theta})$, with

$$E[\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\tau}] = \boldsymbol{\theta} \quad \text{and} \quad \text{Var}[\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\tau}] = \boldsymbol{\theta}\boldsymbol{\theta}^T/(\boldsymbol{\tau} + 1).$$

- $\boldsymbol{\theta}$ determines the location (the mean) of the distribution of $\mathbf{Y}$ in the simplex (of dimension $d, \mathcal{S}^d$).
- The scale parameter $\boldsymbol{\tau}$ is a strictly positive value, has no influence on the expectation (it affects only the dispersion).

Regression

As in Grunwald, Raftery and Guttorp (1993), we assume $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\eta})$ where $E[\boldsymbol{\theta}] = \frac{\boldsymbol{\eta}}{\sum_{d=1}^{D}(\eta_d)}$, and we can model the location using a generalized linear model on the η parameter (Hijazi and Jernigan (2009)) as follow:

$$\log(\eta_d) = \beta_{d0} + \beta_{d1}X_1 + \beta_{d2}X_2 + \cdots + \beta_{dk}X_k,$$

The scale parameter τ in our model will be treated as a constant.

We can further generalize the model to a APC Dirichlet mixed model to be applied to the study of causes of death as follow:

$$\log(\eta_d) = \beta_{d0} + \beta_{d1}Age + p_d + c_d$$

We model the proportions of the CoDs considering as covariates Age, Period, and Cohort.

A-P-C Model specification

A-P-C Model specification

We chose to benchmark the canonical framework using constraints against two different models using, (i) random effect for period, and (ii) for both period and age.

Let $\mathcal{A} = \{a_0, a_1, \ldots, a_\omega\}$, $\mathcal{P} = \{p_0, p_1, \ldots, p_n\}$, and $\mathcal{C} = \{c_0, c_1, \ldots, c_m\}$ be the set of age, year and cohort categories, respectively. The A-P-C model describes the proportion of death at age $a \in \mathcal{A}$, time $p \in \mathcal{P}$, and cohort $c \in \mathcal{C}$, considering a multi dimensional output for each cause $d \in \mathcal{D}$, where $\mathcal{D} = \{d_0, d_1, \ldots, d_k\}$,

Constrained Model

We start by introducing the constrained model as the classic way and thus adapting the constraints to our multi causes framework in the following formulation.

$$\log(\eta_d) = \beta_{(d,a)} + \beta_{(d,p)} + \beta_{(d,c)} \tag{2}$$

where, βs are coefficients of fixed effects using categorical coding for Age, Period and Cohort respectively, and $\forall d \in D$ we impose the following constraints:

$$\begin{cases} \sum_{\omega=1}^{\Omega} \beta_{(d_1,a_\omega)} = \sum_{n=1}^{N} \beta_{(d_1,p_n)} = \sum_{m=1}^{M} \beta_{(d_1,c_m)} = 0 \\ \sum_{\omega=1}^{\Omega} \beta_{(d_2,a_\omega)} = \sum_{n=1}^{N} \beta_{(d_2,p_n)} = \sum_{m=1}^{M} \beta_{(d_2,c_m)} = 0 \\ \\ ..... \\ \\ \sum_{\omega=1}^{\Omega} \beta_{(d_k,a_\omega)} = \sum_{n=1}^{N} \beta_{(d_k,p_n)} = \sum_{m=1}^{M} \beta_{(d_k,c_m)} = 0 \end{cases} \tag{3}$$

Random Effects

$$\log(\eta_d) = \beta_{(d,a)} + \beta_{(d,p)} + \gamma_{d,c} \tag{4}$$

Where, $\gamma_{d,c}$ is the random effect for cohort level specific for each cause, and βs fixed affects using categorical coding for Age and Period.

$$\log(\eta_d) = \beta_{(d,a)} + \gamma_{d,p} + \gamma_{d,c} \tag{5}$$

In this case, $\gamma_{d,p}$ and $\gamma_{d,c}$ are random effects for period and cohort level (specific for each cause), and β provide fixed effects using categorical coding for Age. For all tree models we specify a $\tau \sim \Gamma(0.01, 0.01)$ for precision parameter in Dirichlet distribution. Furthermore, we use flat priors for random effects coefficients.

Model estimation

- Samples from the posterior distributions were drawn by using Hamiltonian Monte Carlo sampling and specifically using the stan software stan. Hamiltonian Monte Carlo simulates movement through the parameter space by analogy to a physical system where the potential energy is equal to negative log-posterior, and it is a special case of the more general Metropolis–Hastings algorithm for Markov chain Monte Carlo sampling.
- Four parallel chains were constructed and used to assess convergence to the posterior distribution, each with 8000 samples, and the first half of each chain was used as a warm-up period

Model comparison

- For the majority of model, Gelman-Rubin split $\hat{r}$ diagnostics are below the suggested 1.05 threshold and examination of trace-plots indicate sampler convergence to the target distribution.
- Separate models were therefore estimated and their accuracy was assessed by using the leave-one-out information criterion (LOOIC). The LOOIC is a measure of how well we might expect a model to perform in predicting a data point without including it in the data that are used to fit the model.

Data

Cause	ICD7	ICD8	ICD9	ICD10
Infectious	A001-A043, A104, A132, B001-B017, B043	A001-A044	B01-B07	A00-B99
Lung cancer	A050	A051	B101	C33-C34
Other Cancer	A044-A059, B018-B019	A045-A050, A052-A060	B08-B09, B100, B109, B11-B17	C except C33-C34
CVD	A070, A079-A086, B022, B024-B029	A080-A088	B25-B30	I00-I99
Respiratory	A087-A097, B030-B032	A089-A096	B31-B32	J00-98
Digestive	A098-A107, B033-B037	A097-A104	B33-B34	K00-K93
External	A138-A150, B047-B050	A138-A150	B47-B56	V00-Y89
Other	A137, B045, A060-A069, B020, A071-A078, A108-A137	A105-A137, B045, A061-A079	B46, B18-B24, B35-B46	R00-R99, D00-D48, D50-D89,E00-E88, F01-F99,G00-G98, H00-H57,H60-H93, K00-K92,L00-L98, M00-M99,N00-N98, O00-O99,P00-P96, Q00-Q99,R00-R99

Table 1: WHO: ICD codes and classification

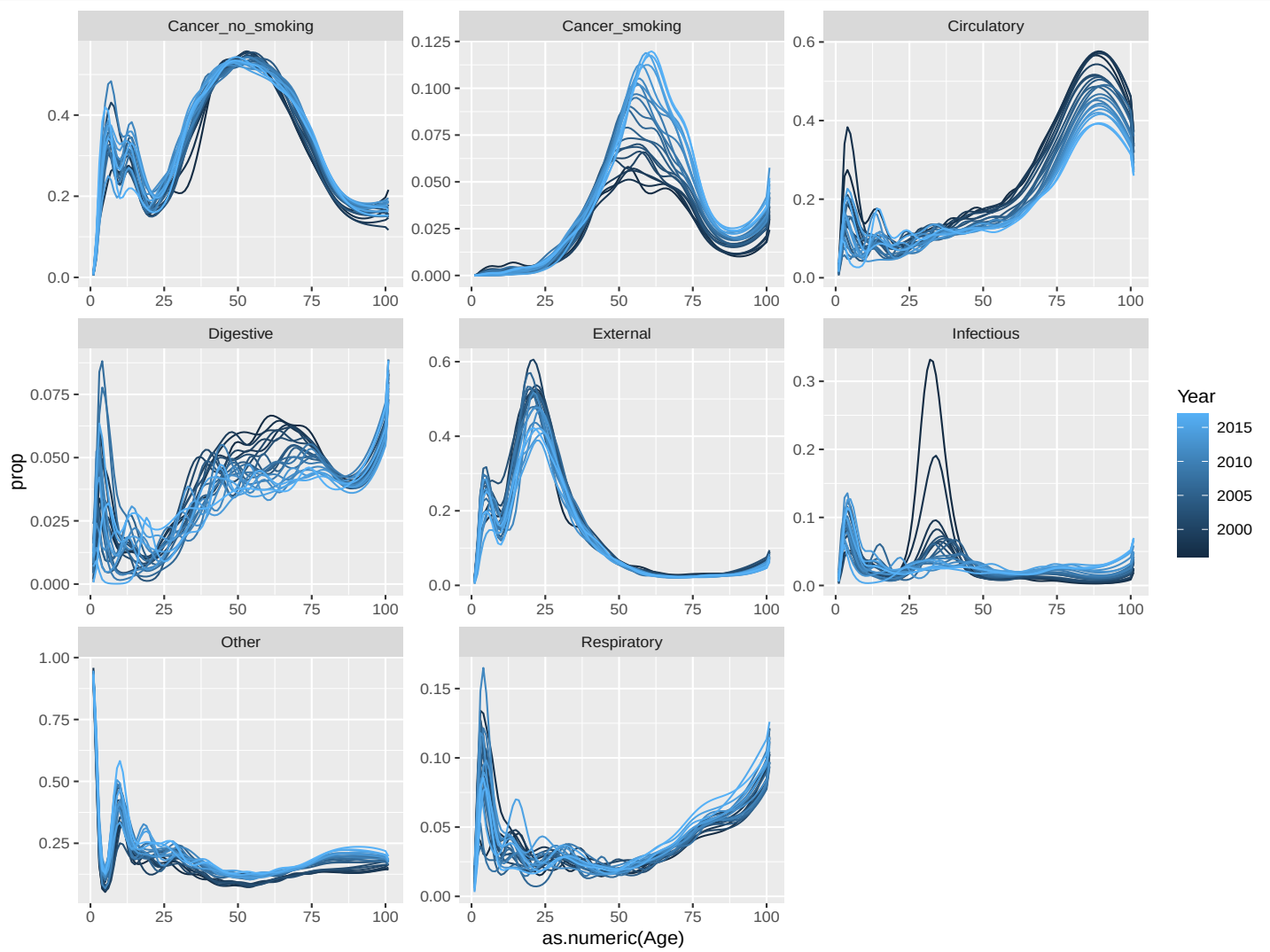
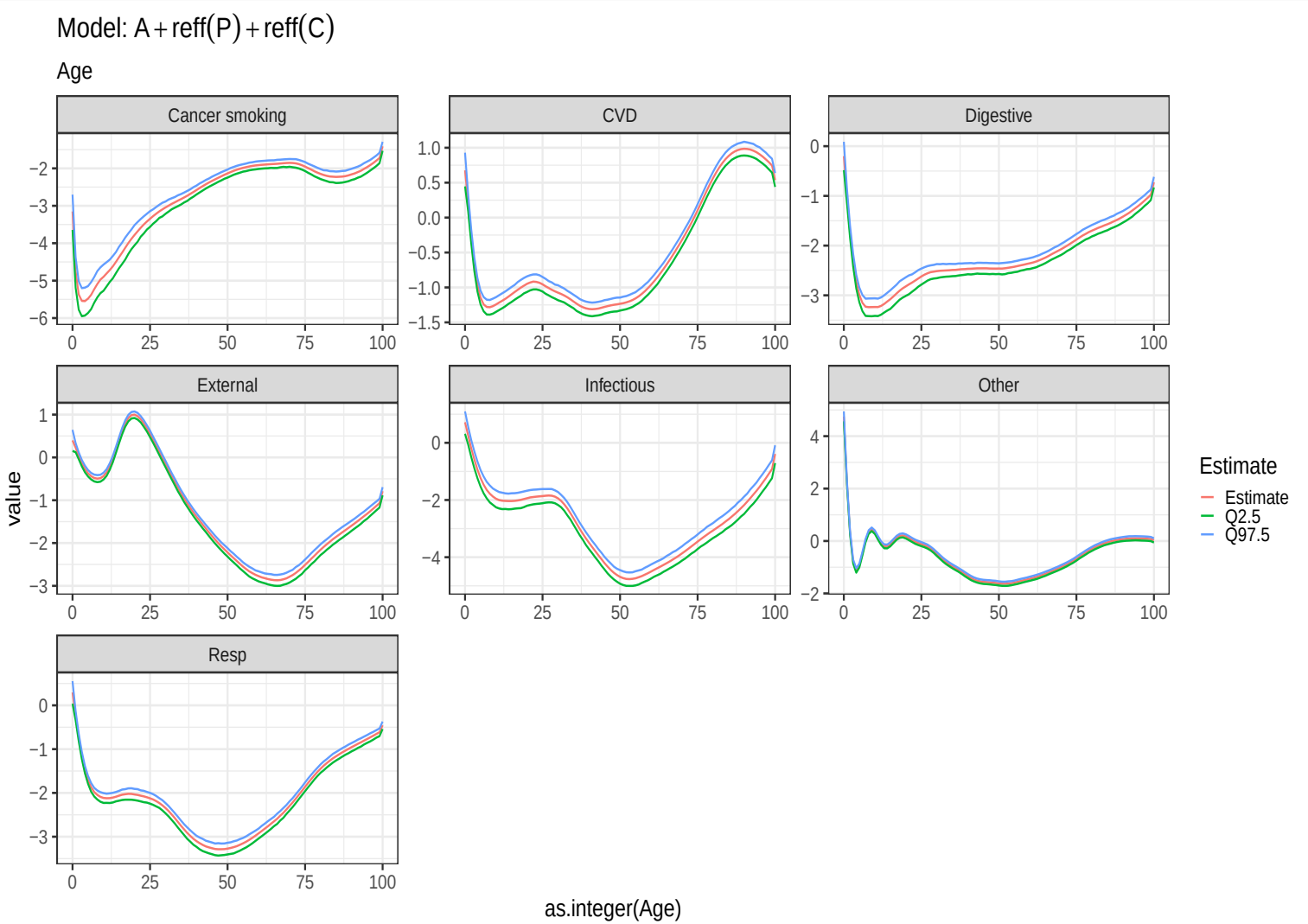
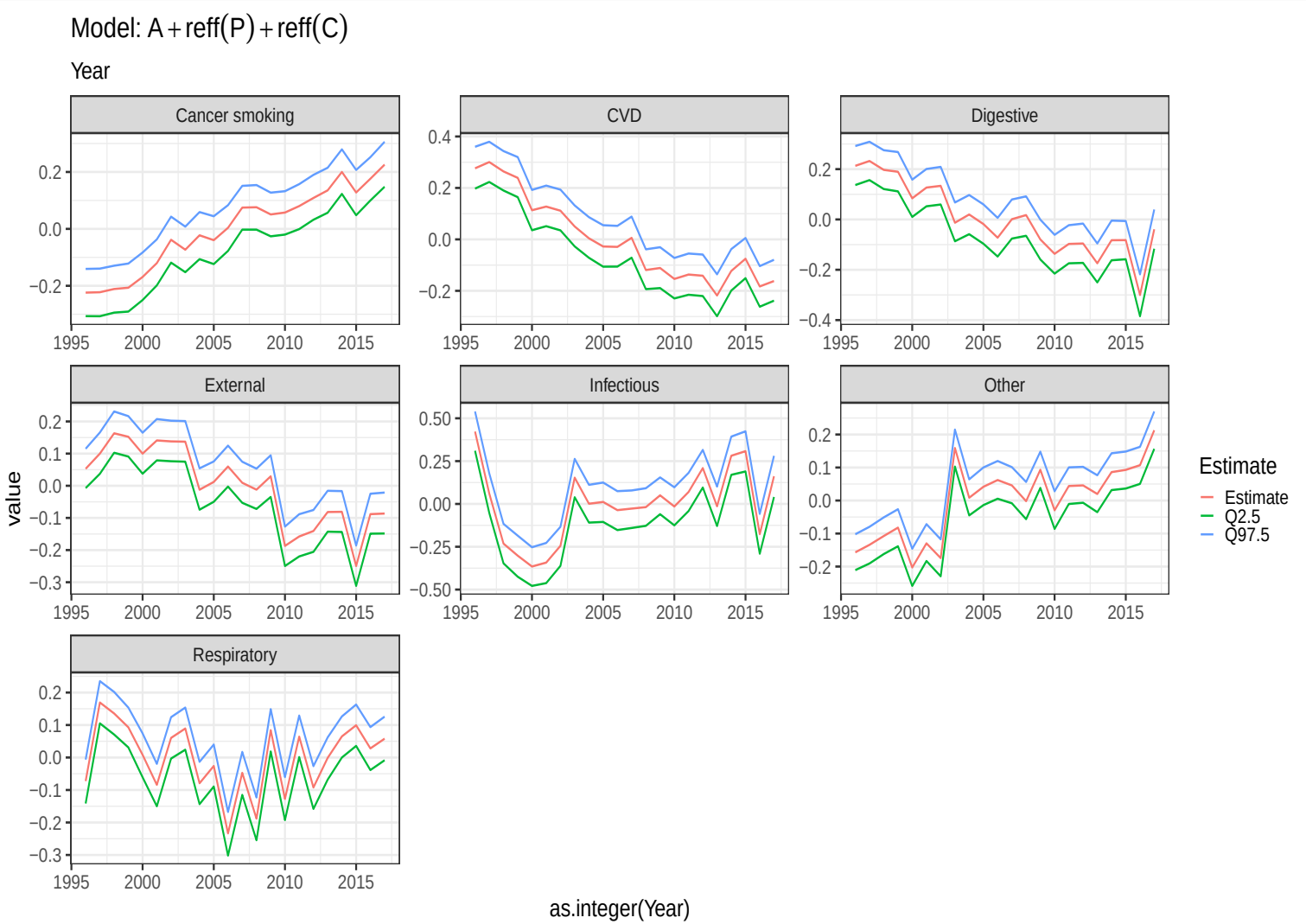
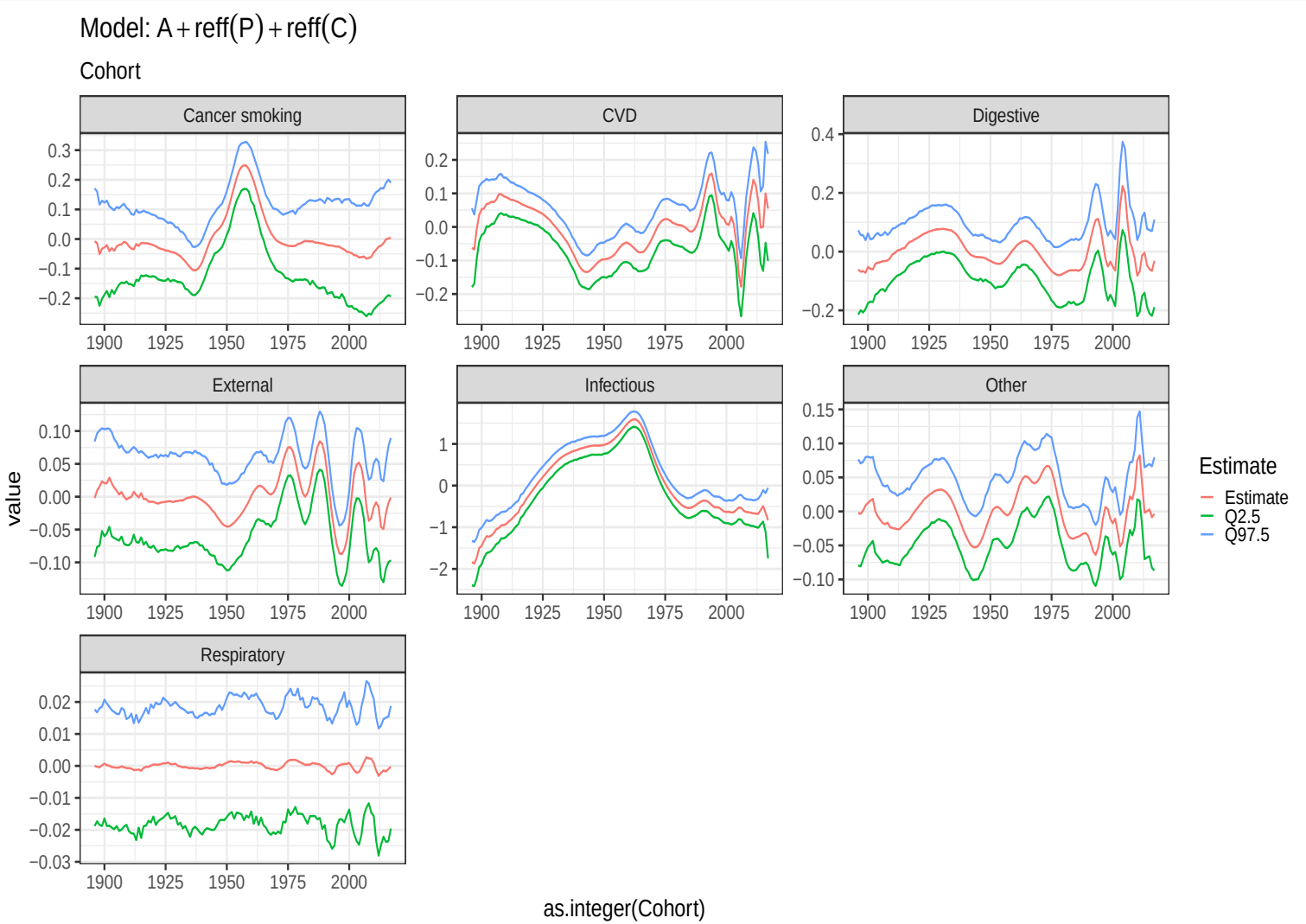


Figure 1: Female population 1996-2017. Age and cause-specific proportions. Estimation using Penalize Composit Link Model (Rizzi et al. 2015) and our ICD classification.

Results


Model: A + reff(P) + reff(C)
Age
Cancer smoking
CVD
Digestive
External
Infectious
Other
Resp
value
as.integer(Age)
Estimate
Estimate
Q2.5
Q97.5


Model: A + reff(P) + reff(C)
Year
Cancer smoking
CVD
Digestive
External
Infectious
Other
Respiratory
value
as.integer(Year)
Estimate
Estimate
Q2.5
Q97.5


Model: A + reff(P) + reff(C)
Cohort
Cancer smoking
CVD
Digestive
External
Infectious
Other
Respiratory
value
as.integer(Cohort)
Estimate
Estimate
Q2.5
Q97.5

Conclusion and further directions

- Cohort effect, one of the big strengths of this study, quite neglected in causes of death modeling.
- Estimating mortality by cause of death can provide valuable information for healthcare and social services planning.

Some observed patterns can be surprising and unexpected but keep in mind:

- We are dealing with proportions instead of mortality rates as usual
- We are modeling causes-specific mortality instead of overall mortality.

Future developments:

- Forecasting, one period and cohort component for each cause, dependence structure in time series model.

Thank you for your attention.

Questions / Suggestions ?