Multivariate response Age-Period-Cohort models

Paul Hewson
Centre for Mathematical Sciences, Plymouth University

Motivation

Age-Period-Cohort models are widely used in epidemiology to project incidence rates into the future. A disease rate is modelled, either as a either standardised morbidity rate or standardised mortality rate depending on whether we measure illness or death, in the form $\frac{y_{ij}}{o_{ij}}$. Here, y_{ij} denotes the observed count of disease or death for all persons of *age* $i = 0, 1, \ldots$ in calendar year (*period*) j; o_{ij} denotes the population of age i in period j at risk of disease or death. It is standard to assume that counts y_{ij} are realisations of a Poisson random variable so that we model:

$$Y_{ij} \sim Poisson(o_{ij}\lambda_{ij})$$

where Y_{ij} is a random variable, o_{ij} is the population at risk and λ_{ij} is the unknown parameter of the Poisson distribution. The Age-Period-Cohort model for λ_{ij} can be fitted as

$$Y_{ij} \sim Poisson(\lambda_{ij})$$

with

$$\log(\lambda_{ij}) = \log(o_{ij}) + \theta_i + \gamma_j + \psi_k$$

where θ_i is a set of parameters capturing the *age* effect, γ_j are a set of parameters capturing the *period* effect and ψ_k has been introduced as a set of parameters capturing a *cohort* effect. These models are known as Age-Period-Cohort (APC) models. Clearly there is a potential "identifiability" problem as that cohort index k is a simple linear function of the age index i and the period index j.

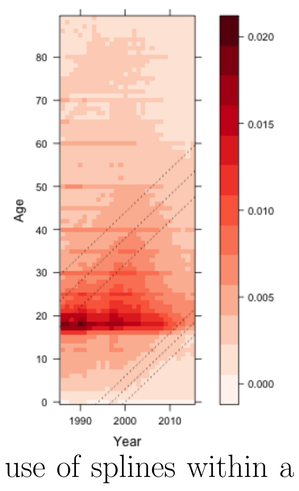
The basic A-P-C model

Our models are developed for a random variable Y_{ijlm} denoting the number of police reported road injuries where l indexes the Highways Authority in which the casualties were injured and m denoting the gender of the casualty. A basic model for such a scenario has been established [3] We have made two extensions to this work. Firstly, rather than the more conventional random walk parameters for age, period and cohort we use penalised splines as previously described [2] and secondly, we consider that various injury counts (cyclists, pedestrians, car occupants) severities (fatally, seriously or slightly injured) will exhibit an informative correlation pattern. Hence we model:

$$\log(\lambda_{ijlm}) = \log(o_{ijlm}) + \alpha_l + s(A) + s(P) + s(C) + \epsilon_{ijm}$$

where α_l are random intercepts for highways authority nested within police authority, ϵ is an over-dispersion random effect. $s(\cdot)$ denotes the respective penalised smooth splines for each of the Age (A), Period (P) and Cohort (C) effects.

Reported road injury in the UK



- By way of illustration, this Lexis diagram represents all reported injury road collisions in the UK between 1985 and 2014, divided by the ONS estimates of population
- Very strong cohort patterns can be seen and this insight itself has considerable implications for practice (which we can describe as "impact")
- There are many unanswered research questions considering models fitted to different types of injury (fatal injury, serious injury, slight injury) as well as different types of road user (car occupant, pedestrian, cyclists, motorcyclists). This work is also potentially useful in enhancing the use of A-P-C models generally.

The use of splines within a Bayesian framework still presents some identifiability problems within A-P-C modelling which we mainly address through the use of penalised splines. We have applied a Gaussian process to the lower level α_l to account for spatial adjacency of highways authorities in the UK but it is also possible that 2 dimensional spline could solve the same problem more efficiently. The judicious use of multivariate outcomes, to provide alternative solutions to the identifiability problem, is potentially very powerful. Some researchers have been able to borrow strength from related populations and to make predictions at the sub-national level for the same disease [3]. Our approach is to borrow strength from related diseases as well as borrowing strength from the same disease at the sub-national level. When modelling the random intercept α_l we have successfully applied a multilevel nesting whereby highways authorities are nested within their respective police authorities and this helps with differential reporting problems in the raw data. Similar work has considered that different cancer registries might have different recording practices [1].

Penalised Splines

For a univariate smooth regression given by $Y_i = s(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ we can use the following low-rank cubic term as the basis:

$$s(x, \theta) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k |x - \kappa_k|^3,$$

with $\boldsymbol{\theta} = (\beta_0, \beta_1, u_1, \ldots, u_K)^T$ as the vector of regression coefficients with $\kappa_1 < \kappa_2 < \cdots < \kappa_K$ as a set of fixed knots. In order to penalise overfitting we minimise:

$$\sum_{i=1}^{n} \big(y_i - s(x_i, \boldsymbol{\theta})\big)^2 + \frac{1}{\lambda}\boldsymbol{\theta}^T \boldsymbol{D}\boldsymbol{\theta},$$

where λ is an unknown smoothing parameter and $\boldsymbol{D}$ is a known positive semi-definite penalty matrix:

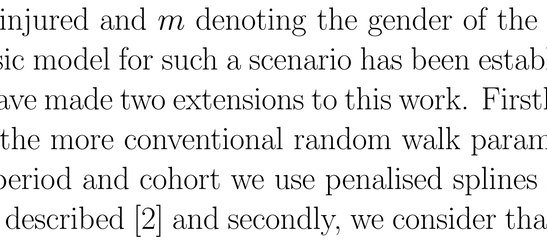
$$\boldsymbol{D} = \begin{pmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \boldsymbol{\Omega}_{K\times K} \end{pmatrix},$$

where the (l, k)th entry of $\boldsymbol{\Omega_K}$ is $|\kappa_l - \kappa_K|$. This lends itself to simple reformulation as a mixed effects model of the form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ denotes the conventional "fixed" effects, $\boldsymbol{b} = \boldsymbol{\Omega^{1/2}}\boldsymbol{u}$ and $\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{\Omega^{1/2}}$ are treated like random effects. In our GLM formulation the ϵ term is absorbed within the random component. These are simple Bayesian models to fit with Hamiltonian Monte Carlo; Gaussian priors are assumed for smooth function regression coefficients and the nested local authority specific intercepts. Structural correlation between different outcomes can be induced in different ways; for example it is possible to assume common age effects across different outcomes or to allow different smooth terms but only allow the intercepts to be correlated. The intercept terms are specifically correlated by Highways Authority and the ϵ terms are correlated across road user type. Standard convergence criteria are checked, overall model fit is assessed using Gneiting's modifications to the Posterior Predictive Score (PPD).

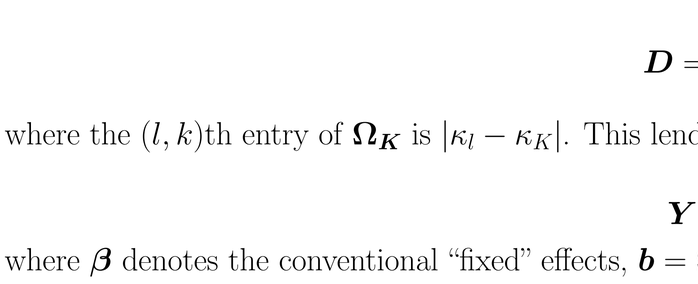
Smooth function for age



Familiar age patterns are seen with road users in their late teens / early twenties at greatest risk of road injury.

Fitting smooth splines to Age-Period-Cohort models

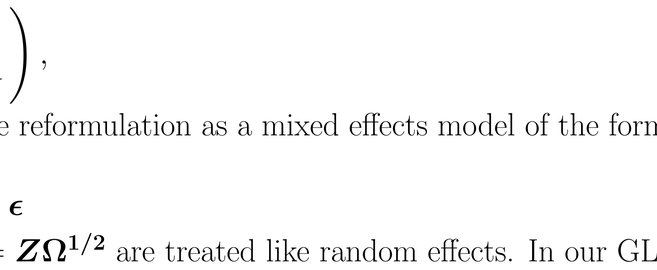
The figure below shows the smooth term for cohort; the underlying linear trend is arbitrary.



The letters on the diagram depict significant events that may have had an effect on driving habits within the UK (for example the introduction of compulsory driving tests) and the arrows point towards the cohort that would first be affected by these events. However, the most interesting feature is the change of shape around the 1970 cohort. Because of identifiability issues this second order change is of more interest than the actual slope.

Illustrative results

The "caterpillar" plot below depicts point estimates (crude *per capita* risk) as well as the smoothed modelled risk for each Highways Authority. The red and green arrows denote how much risk estimates are adjusted up or down respectively through the use of model.



These plots are much more "interesting" (potentially impactful) when distinct road user groups are considered as much greater variation can be seen between the raw and the age adjusted estimates. This work is currently being disseminated through practitioner focussed outputs and the underlying methodological work has been submitted to traditional peer-reviewed academic journals.

References

[1] N Best and J Wakefield (1999) *Accounting for inaccuracies in population counts and case registration in cancer mapping studies* Journal of the Royal Statistical Society: Series A (Statistics in Society) 162.3 (1999): 363-382.

[2] P.J. Hewson (2011) *Analysis of head injuries using a Bayesian Vector Generalized Additive Model* Australian & New Zealand Journal of Statistics 53(2): 233-246.

[3] A. Riebler, L. Held and H. Rue (2012) *Estimation and extrapolation of time trends in registry data - borrowing strength from related populations* The Annals of Applied Statistics 6(1):304-333