Parameter Shrinkage for Age-Period-Cohort Modeling of Opioid Mortality Rates

Gary G. Venter: Columbia University and University of New South Wales

Abstract: Statistical methods that shrink parameters towards zero can produce lower predictive variance than does maximum likelihood. This paper discusses an approach to doing this for age-period-cohort models, and applies it to fitting opioid mortality rates with a generalization of the Lee-Carter model including cohorts. Bayesian parameter shrinkage has some practical advantages over classical versions.

Keywords: MCMC, Lee-Carter, Regularization, Cohorts, Opioid mortality.

Introduction

Here we use "we" in the singular, as is now common for "they," not in the royal sense.

Demography, epidemiology and sociology have used age-period-cohort (APC) models since Greenberg, Wright, and Sheps (1950). Actuaries also use them, but with less interaction with other disciplines. The bilinear AP Lee-Carter model is one form actuaries use, and they and others have extended it with cohort effects. A generalization is the model of Hunt and Blake (2014), which can include multiple trends. These address the issue that the ages with the greatest trend can change over time, for instance as medical research focuses on different diseases. Using multiple trends increases the potential for over-parameterization, however.

Common methods to reduce parameter count include putting the parameters on smoothed curves, like the Gompertz curve for mortality, or on cubic splines. We follow Barnett and Zehnwirth (2000), who use piecewise-linear curves and make the parameters the second differences of the APC parameters. These are slope changes between the line segments of the piecewise-linear curve. Eliminating or shrinking a slope-change continues the previous trend at that point, or at least flattens out the change. If no slope-change parameters are shrunk, the model gives the same fit as using APC level parameters. We refine the approach by applying Bayesian regularization to the parameter shrinkage. The APC level parameters are cumulative sums of the slope changes, and the fitting can be set up as a regression with dummy variables for each slope change. The dummies specify how many times a given slope change is summed in the calculation of the level parameter for a data point. Venter and Şahin (2017) use a similar approach. A related actuarial paper is Gao and Meng (2017), who use Bayesian regularization for cubic spline fitting of an age-cohort model.

Such regularization methods try to reduce the predictive variance by shrinking all parameters except the constant to or towards zero. This usually shrinks fitted and forecasted values towards the mean. It produces biased estimates, but often with a lower estimation variance than that of the minimum variance unbiased MLE estimates.

Actuaries have been shrinking estimates towards the mean to improve predictive accuracy informally since Mowbray (1914), and rigorously since Bühlmann (1967). The latter approach is similar to the James-Stein estimator of Stein (1956), whose very strong result – that some degree of shrinkage always reduces the predictive variance – was known as Stein's Paradox at the time. The famous example of that was improving predicted season ending baseball batting averages by shrinking July 4 averages towards the mean. These methods apply when within and between variances can be estimated for the observed values. The ratio of the variance components controls the degree of shrinkage.

Regularization as discussed here traces to Hoerl and Kennard (1970). Estimation methods like lasso and ridge regression simply shrink regression parameters towards zero. This is controlled by a selected shrinkage scale λ. How much to shrink is informed by cross-validation, that is, by measuring how well the model predicts on holdout samples for various λs. It is unusual for the no-shrinkage limit of MLE to be optimal, but it is generally included.

Bayesian shrinkage is done by giving the parameters mean-zero prior distributions, like normal, double exponential, or Cauchy. The double exponential prior gives the lasso estimates as the posterior mode, and the normal prior does this for ridge regression. The Bayesian versions have a few advantages:

- They come with a goodness-of-fit measure that can be calculated quickly. Traditional measures like AIC are not applicable in that the effective number of parameters is not clear when parameters are shrunk. A fit measure based on leave-one-out cross-validation, looic, can be computed efficiently from the MCMC (Markov Chain Monte Carlo) estimation of the posterior distribution. This measures the relative predictive accuracy of different choices of λ.
- Fischer information assumes MLE. MCMC produces parameter ranges directly, as it generates a sample of the posterior distribution of the parameters.
- It is not necessary to select a single degree of shrinkage. Putting a prior distribution on λ itself generates a posterior sample of the parameters including a range of λ values. Using such "hyper-priors" is considered the fully Bayesian method in that it is not optimizing looic but rather is based purely on the posterior distribution.

Vehtari, Gelman, and Gabry (2017) derive the looic computation. Although their calculations measure the predictive accuracy of the left out points, their derivation is based on adjusting the loglikelihood for sample bias, just as the AIC aims to do. Thus looic evaluates the parameter shrinkage on how well the model fits the data after considering sample bias, which is what AIC does.

APC models of population trends can be used when the data comes in cells, each for a specific age group from a specific year of origin (like year-of-birth cohort) observed at a specific time (the period). Often the estimate of the mean for the cell is the product of an age factor, a period factor, and a cohort factor with an overall constant term. In that case, the logs of the cell observations have an additive APC model.

Of course two of the three age, period and cohort indicators determines the other, usually with period = cohort + age. The parameters too have an identifiability feature – there are offsetting transforms of the parameters that will give the same fitted values for every cell. This was already well understood by the time of Fienberg and Mason (1978), who conclude that APC models require an arbitrary and unsupportable choice of constraints in order to get unique parameters.

Actually, it is not even clear that there are true or meaningful age, period and cohort effects that can be defined independently of each other. Bijlsma et al. (2017) challenge this notion, for example. To some extent outside data, like from surveys of attitude changes across cohorts, can lend insight into the strengths of various constraint choices, but this is not usually specific enough to quantify any absolute effects. One possible constraint is to assume that all the cohorts are identical. Then the age and period parameters would be conditional on this constraint. They would be the same as in a straight AP model, so AP model parameters are the same as would be produced in an APC model under the assumption that there are no cohort differences. That is a very strong assumption and is in fact testable, for example by looking at residuals by cohort.

The aim here is just to estimate age, period and cohort effects relative to each other. The constraints selected attribute the entire time trend to the period parameters, with no overall trend in the cohort direction. This is done by making the cohort parameters the residuals of a time trend through proposed parameters. Doing this does not necessarily force the earliest and latest cohorts factors to be similar, but in the opioid mortality data they turn out to be. Because we fit parameters on piecewise-linear curves, the points where slopes change form a set of constraints themselves, which could be enough to produce unique APC parameters. In fact we find that works for the linear APC model here. This does not make them the true effects however – they are just conditional on those constraints, which are not as easy to interpret as is the constraint of moving any long-term cohort trend into the period parameters. The cohort factors then represent differences among cohorts relative to the assumption that there is no long-term cohort trend.

Methodology

Regularization

The double exponential, or Laplace, prior, has the shape of an exponential distribution for positive values, and this is mirrored for negative values. The density for a parameter b, with the exponential scale λ, is $p(b) = 0.5\lambda e^{-\lambda|b|}$. It pushes the posterior of b towards zero.

MCMC estimation of the set of parameters b for observations y starts with Bayes Theorem: $P(b|y) = P(y|b)P(b)/P(y)$. $P(y)$ is assumed to be an unknown constant, so the posterior probability is proportional to the numerator $P(y|b)P(b)$, which is the likelihood times the prior. That property is enough to generate MCMC samples from the posterior, or even to find the b that maximizes it, i.e., the posterior mode.

For a model with k parameters, the negative log of the numerator is $NLL + k(log2 - log\lambda) + \lambda \sum |b_j|$. The posterior mode is thus produced by minimizing $NLL + \lambda \sum |b_j| - k(log\lambda)$. With a pre-selected value of λ, the final term is a constant that can be dropped. That leaves minimizing $NLL + \lambda \sum |b_j|$, which is what is done in classical lasso. That is why using the Laplace prior is called Bayesian lasso, especially when the posterior mode is taken as the estimator. If λ is itself a parameter to be estimated, the $-k(log\lambda)$ term cannot be dropped, and then the minimization gives a joint estimate for b and λ. In that case, however, a prior would also have to be specified for λ. If a uniform prior is used, that adds a constant that can be dropped, so the joint posterior mode would just minimize $NLL + \lambda \sum |b_j| - k(log\lambda)$.

Doing this same calculation starting with a mean-zero normal prior in σ^2 gets the posterior mode by minimizing $NLL + \lambda \sum b_j^2 - (k/2)log\lambda$, where $1/\lambda = 2\sigma^2$. For a selected value of λ, the last term is a constant, so $NLL + \lambda \sum b_j^2$ is minimized. This is ridge regression.

An increasingly popular shrinkage prior is the Cauchy distribution with $1/p(b) = \pi(\lambda^2 + b^2)/\lambda$ and $-log(p(b)) = -log\lambda + log\pi + log(\lambda^2 + b^2)$. For a fixed λ, the posterior mode minimizes $NLL + \sum log(\lambda^2 + b_j^2)$. This is an alternative to both lasso and ridge regression. The Cauchy prior often yields more parsimonious models with better looic than the normal or Laplace priors give. It has more weight near zero but is also heavier tailed, which allows a few larger parameters when they are called for. We use the Laplace prior below, but checked the final model by refitting it with the Cauchy prior. The fit was very similar in this case.

Bayesians often prefer the posterior mean, as the posterior mode is just a single parameter sample. The parameters that maximize the posterior probability could be doing so by over-fitting some features of the particular sample, whereas the posterior mean considers all parameter sets that provide a plausible explanation of the data. For the Laplace prior, the posterior mode shrinks some parameters exactly to zero, which also happens in classical lasso. This makes lasso a method for variable selection as well as for parameter shrinkage. The posterior mean does not do this, as usually the zero parameters differ among the samples. Then dropping parameters that are near zero, especially if the sample values are equally spread across positive and negative values, is a practical alternative. In fact, doing this often improves the adjusted loglikelihood, looic.

Leave-one-out model evaluation for MCMC traces to Gelfand (1996), who developed an approximation for a point's out-of-sample loglikelihood by using the numerical integration method importance sampling over the parameter samples. He estimated the loo loglikelihood at a data point as the point's weighted average loglikelihood over all the MCMC-generated parameter samples, with the weight for a sample proportional to the reciprocal of the point's likelihood in the sample. This gives greater weight to the samples that fit the point poorly, which presumably would resemble parameters fit without it. The estimate of the loglikelihood of the point from this is the reciprocal of the average over all the samples of the reciprocal of the point's loglikelihood in the sample. With this, the generated posterior distribution of the parameters provides a ready estimate of the out-of-sample NLL.

That gave good but volatile estimates of the loo loglikelihood. Vehtari, Gelman, and Gabry (2017) improved on it by using an approach similar to extreme value theory – they fit a Pareto to the probability reciprocals and use the fitted Pareto values instead of the actual values for the largest 20% of the sample. They call this "Pareto-smoothed importance sampling." It has been extensively tested and has become widely adopted. The penalized likelihood measure is labeled $\widehat{elpd}_{loo}$, standing for "expected log pointwise predictive density."

We use the Stan MCMC software platform here. It provides a loo estimation function that can work on posterior samples from any MCMC package. It calculates $\widehat{elpd}_{loo}$ as well as the implied loglikelihood parameter penalty, and something they call looic – the loo information criterion – which is $-2\widehat{elpd}_{loo}$ in accord to standards of information theory. Since the factor of 2 is not critical, here the term looic is used for $-\widehat{elpd}_{loo}$, which is half of the usual looic but conveniently is the NLL increased by the parameter penalty.

The derivation of looic, like that of AIC, starts by assuming that the data is generated by the model. This is increasingly problematic in some areas, like consumer finance, which tries to optimize interest income by design of credit card terms. The data is generated by a massive collection of neural processes across the population, influenced in part by competitor behavior, and the models do not attempt to represent that process. They can still generate good profits in the short run, but have to be changed often. Areas like this pose a challenge to the idea that the data is generated by the model process. A common response is to use slightly more parsimonious models than the statistical measures suggest, and that could be advisable with the APC models here as well.

The Models

A Poisson model is used here for mortality counts. Theoretically as a sum of Bernoulli processes, these should be binomial, but for low frequency the two are virtually indistinguishable. We tested the negative binomial, as that sometimes captures contagion by other processes, but it closely approximated the Poisson limit. The Poisson mean for the cell at age i, period j is $\mu_{i,j} \times \text{population}_{i,j}$. In the linear APC model with a constant term,

$$log(\mu_{i,j}) = c + p_i + q_j + r_{j-i}, \tag{1}$$

where p, q, r are the age, period, and cohort parameters, respectively. Some constraints are necessary for identifiability. To begin with, we assume that the p, q, r parameters all start at the second value of their subscripts, but other constraints are discussed below. The Lee-Carter model is a bilinear AP model, with another set of age parameters v_i that allow the period trend to vary by age:

$$log(\mu_{i,j}) = c + p_i + q_j v_i \tag{2}$$

Cohorts can be added to this. The Hunt-Blake model does that but also allows more than one period trend so that it can handle changes in the trending by age:

$$log(\mu_{i,j}) = c + p_i + \sum_k q_j^{(k)} v_i^{(k)} + r_{j-i} \tag{3}$$

The age weights on the period trends can capture mortality changing faster for some ages. Including multiple trends can allow the fastest-trending ages to change over time. For identifiability we assume that these weights are all positive and the weight for the age with the greatest trend is 1.0. (This age might vary across the parameter samples in MCMC, however.) With this convention, the weights v_i can start with the earliest age, even though the age terms p_i start at the second age.

We follow Venter and Şahin (2017) in assuming there is no overall trend across the cohorts. This can be forced by fitting a linear trend to the cohort parameters r_{j-i} and subtracting it from the parameters. This is the kind of arbitrary choice needed for identifiability, but at least it allows a consistent interpretation of the parameters:

- The period trend is the trend for the age with the greatest trend, given that there is no overall trend across the cohorts.
- The cohort effect measures the relative differences among the cohorts given that the overall trend is entirely in the period parameters.

This allows the period parameters to also start at the first period.

For the estimation, it can be helpful to string out the array of observations into a single vector. It is also helpful to keep two columns that identify the age and period for each observation. For direct estimation of the age, period, and cohort parameters, each parameter can be given a dummy variable that has value 1 for a cell that it affects, and 0 for the other cells. For the second difference variables, the dummies are a bit more intricate.

The second difference parameters, starting from the first age, period, and cohort, add up cumulatively to be the first differences, and these add up cumulatively to be the APC level parameters. The dummy variable value for such a parameter for a given cell is the number of times that second difference parameter gets added up for that cell. If a_w is the second difference parameter for age w, and a cell is for age i, the a_w dummy variable has value $\max(0, 1 + i - w)$ at that cell. This is the same for the period and cohort second difference dummy variables.

Overdose Mortality Application

The opioid overdose epidemic is of special concern due to the sharply rising period trends. Understanding the age and cohort effects could help target ameliorative resources. The overdose mortality rates incorporate fairly complex trends that this modeling methodology can help identify. The data used is for the US population of non-Hispanic white males ages 17–65 for 1999-2016, from the National Center for Health Statistics (2017), including underlying causes of death unintentional drug poisoning (X40-X44), suicide drug poisoning (X60-X64), assault by drugs (X85), and drug poisoning of undetermined intent (Y10-Y14), as coded in the International Classification of Diseases, 10th Revision.

Here we fit the data by building up from simpler to more complex models, starting with a basic AP model, then a linear APC model, then Lee-Carter with cohorts, and finally the Hunt-Blake model. The fit, as measured by looic, but also by NLL, improves at each step. We do not claim that this fitting process or the final model are optimal.

The slope change variables are negatively correlated, which can give MCMC software problems in finding good starting parameters. Lasso estimation of the AP model provides a starting point for Bayesian lasso. Then MCMC output gives good indications of insignificant variables that can be eliminated. We started with that and then put in the cohort variables, with starting values of zero for these and the AP fit for the AP variables. Eliminating insignificant variables then makes this the starting point for the Lee-Carter with cohorts model, which adds age weights to the period trends. We then address some weaknesses with the Lee-Carter assumptions and add a secondary trend.

Figure 1 shows the evolution of these mortality rates by age over the period. Each year's curve consists of centered weighted moving averages of seven ages, with weights of (1,2,3,4,3,2,1)/16. A few year-of-birth cohorts are highlighted. The mortality for most of the ages is increasing but there seem to be two waves of ages with higher mortality and a dip between them. These waves and the dip are moving to the right as they move up, which could be related to cohorts aging. However they are moving to the right slightly more slowly than the cohorts are aging. The model fit accounts for that as a more rapid increase in the mortality rates for the younger ages with a cohort effect superimposed.

The starting point is a model with the age and period variables in a multiple regression for the log of the OD mortality rate, using the R glmnet lasso package. The R code to run this is fairly simple:

```
library(glmnet)
y = scan('ody.txt')
x = as.matrix(read.table('odx.txt', header = FALSE))
fit1 = glmnet(x, y, standardize = FALSE)
plot(fit1, label=TRUE)
cvfit = cv.glmnet(x, y, standardize = FALSE)
cvfit$lambda.min
```

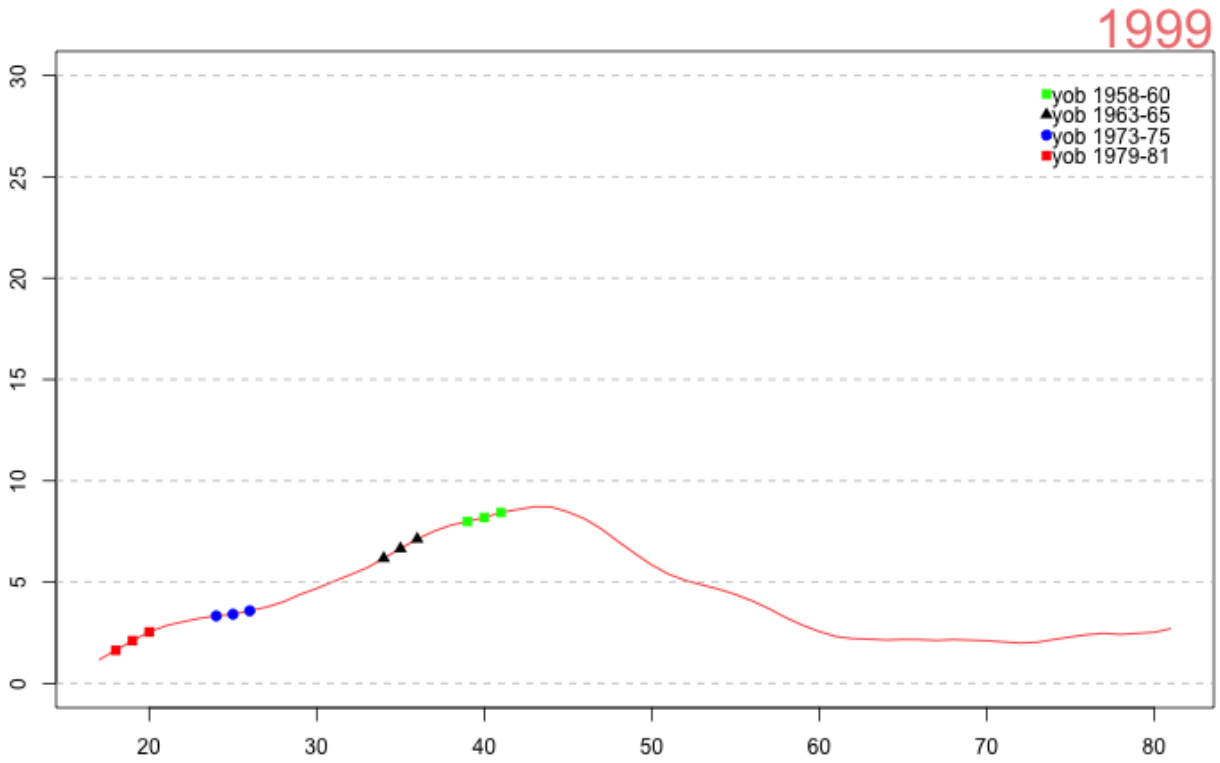

Fig. 1: Mortality Rates by Age and Year with Selected Year-of-Birth Cohorts – Click to Animate, Hold-Click to Pause

This package quickly does the lasso modeling for 100 values of λ that it picks, ranging from a large value that leaves only the constant term down to a small value that produces straight MLE. Typically in lasso the variables are standardized by a linear transform to have mean zero and variance one, so that the scale of the variables does not influence which are eliminated for each λ. We do not do that, as the variables are counts that indicate how many times each coefficient should be summed for an observation.

There are 18 years of data, 17 of which have parameters, and 49 ages with 48 parameters, which gives 65 parameters plus the constant. The standard lasso plot, displayed in Figure 2, shows how the parameters for each variable shrink to zero as λ increases, from right to left. The two X-axis labels are for the L1 Norm, which is the sum of the absolute values of the parameters, and for the number of non-zero coefficients. Both of these increase as λ decreases. The graph here is somewhat unusual, as parameters go to zero then come back as others are eliminated. This is probably due to the negative correlation of the parameters.

The cv.glmnet function uses cross validation as the start of a process to select λ. It produces a suggested range, from lambda.min, which is a small value of λ that does not eliminate very many variables, to lambda.1se, which is higher. We take all the non-zero variables from lambda.min as input to Stan – 57 of them in this case. The graph in Figure 2 is actually for the lasso model rerun using those 57 variables.

In Stan we fit the Poisson model for number of deaths, selecting the posterior mean of the parameters. With the posterior mean, MCMC does not eliminate variables like classical lasso does, but Stan outputs the mean and standard deviation for each variable, and those with means near zero and high standard deviations were dropped from the model. We also used a uniform prior for λ, and this generated a small ranges of λs in the sampling.

The fitting produced an AP model with 14 age variables and 10 period variables. Looic was 5470.5, with NLL of 5348.1 and so a penalty of 122.4. We used this as the starting point for a linear APC model. There are 66 cohorts in this data, 1934–1999, which gives 65 cohort variables, leaving out the first one. We added these 65 variables to the remaining 24 variables from the AP model, then estimated them with Stan. After

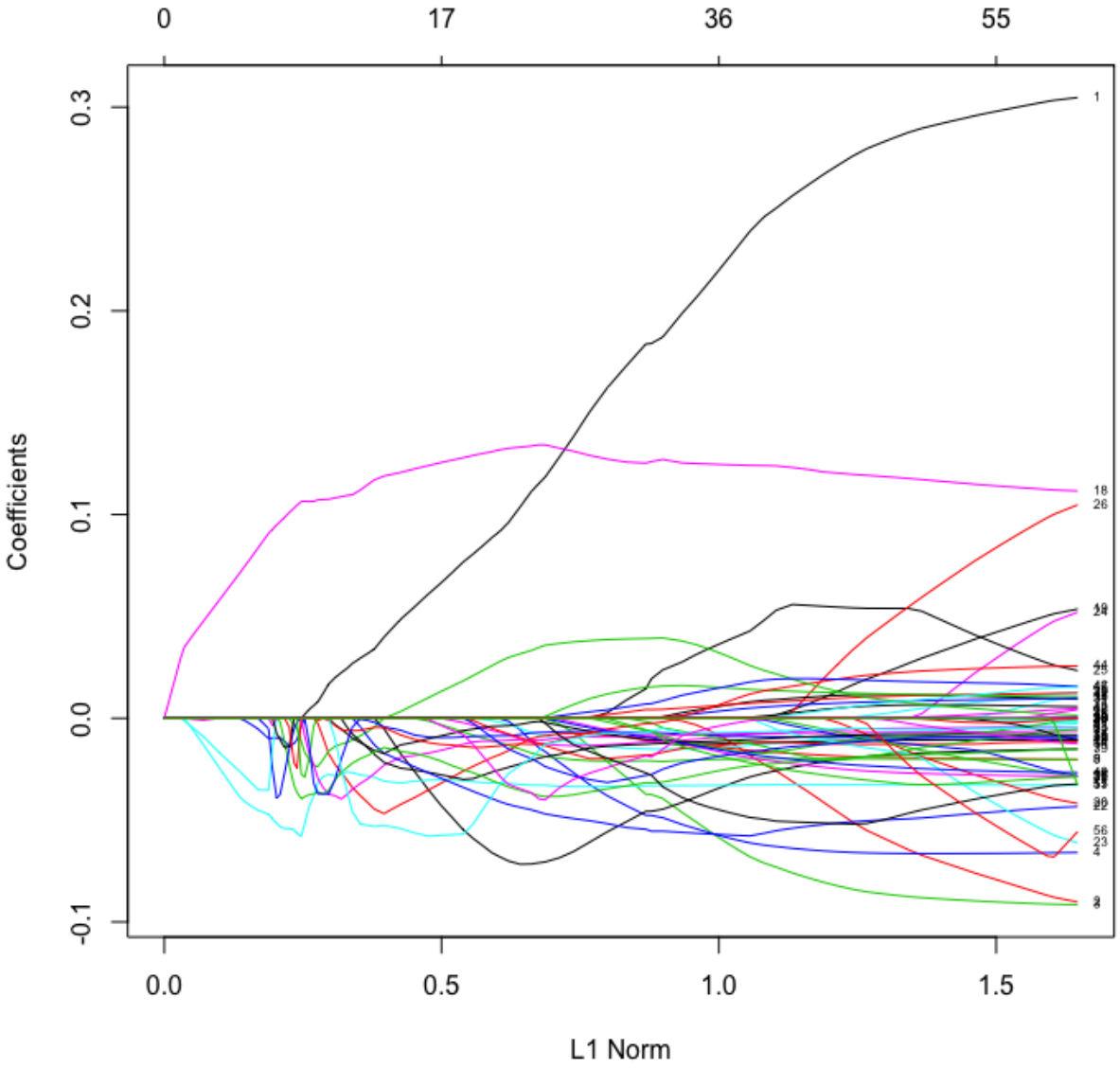

Fig. 2. Parameters for Each Variable as λ Decreases

again taking out the parameters near zero, there were 24 cohort slope-change variables left. Also three more age parameters went to zero, and were dropped as well. That left 11 age variables, 10 period variables, and 24 cohort variables, so 45 in total. This was a much better fitting model, with NLL of 3934.6, looic of 4000.7 and a penalty of only 66.1. With penalty rules like AIC, etc., more parameters would always give a higher penalty. With shrinkage, however, the penalty has to do with how well the left-out points are estimated, and more parameters can sometimes do better at this. The parameters may have been shrunk more towards zero as well.

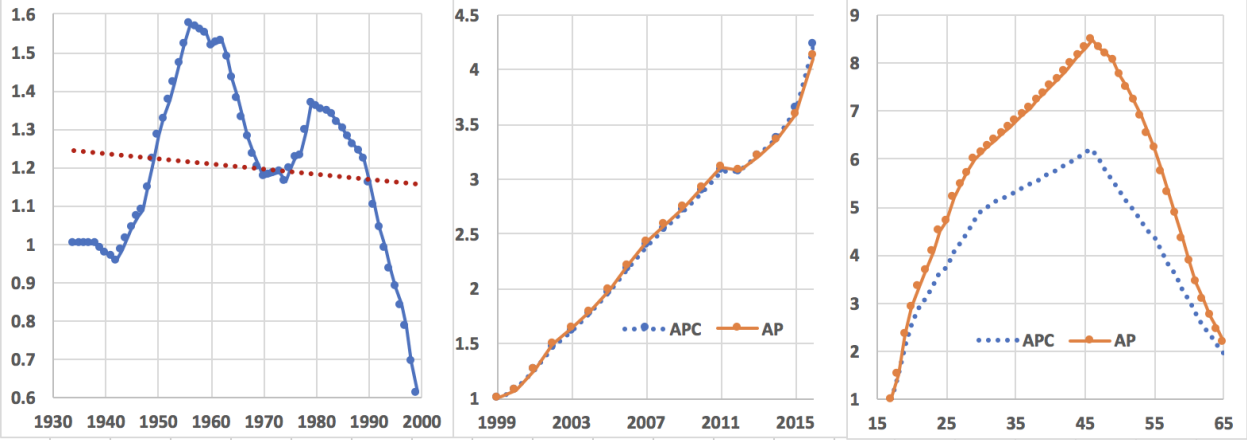

Fig. 3. AP and APC Parameters. Left Panel: Cohorts Factors with Fitted Trend. Center Panel: Period Factors. Right Panel: Age Factors

Figure 3 shows the resulting age, period and cohort parameters for the linear AP and APC models as factor contributions to the Poisson mean. The cohort parameters were not constrained to have zero trend. Keeping the slope-change positions for the age and period parameters provides a degree of constraint for the cohorts. We also constrained the constant to stay in a somewhat narrow range around what it was for the AP model, which forces the average log cohort term to stay near zero. This was enough to get a unique set of parameters in this case, but it is not easy to interpret them. They are conditional on the constraints provided, but those are somewhat complex.

In any case, the years of birth 1956–62 show up as high-risk cohorts, as do several years around 1980. These correspond to the mid-to-late baby boom generation and the early Millennials. There is a dip from about 1968–1977, which is largely Gen X. The pre-boomer and late Millennial groups show relatively lower opioid mortality effects. The fact that the late Millennials are showing as lower than the pre-boomers may be real or may be an artifact arising from not constraining the cohorts enough. The period factors look very similar for both AP and APC models. The age factors in the graph are flatter with the cohorts included, but have a similar shape in both models, peaking at about age 45.

We included all of these variables for the Lee-Carter plus cohorts model, and added age weights for the period parameters for all 49 ages. After doing the initial estimation and eliminating the zero parameters, there were 59 parameters in total, consisting of 10 each for age and period variables, 19 cohort variables, and 20 age weights for the periods. The age weight on the period trend interacts with the cohorts factors, as the cohorts an age is in at the beginning and ending of the data influence the overall trend for the age. Thus it is more critical here to constrain the cohorts to have no overall trend, which we did for this model. The NLL now comes down to 3764.9, with looic at 3823.0 and a penalty of 58.1. The penalty continues to decrease as the fit improves, even with nominally more parameters.

The greatest weight on the trend is for the younger ages, with the weight decreasing by age. The initial mortality curve by age is highest for the oldest ages. This was never the actual experience, but the earliest periods here had the pre-boomers at those ages, so adjusting for the cohort effect could have given that result. Figure 4 shows the evolution of the mortality curve prior to the application of the cohort effect as it rises over the periods. The peak is at age 45 for most of the earlier years, but the curve was pretty flat until recently.

Gradually age 33 takes over as the peak age, which is also happening in the data. The cohort factors now peak at year of birth 1962, and compared to the linear APC model have less of a dip for Gen X and a more minor peak for the early Millennials.

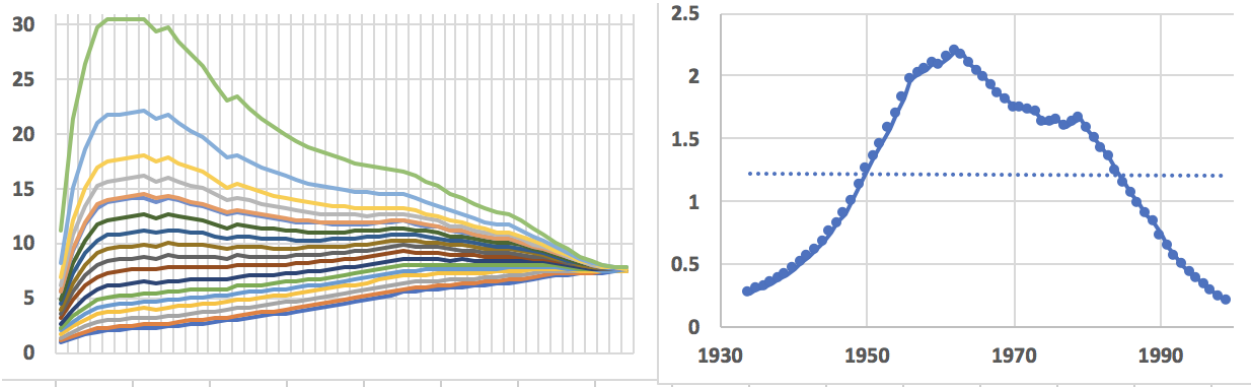

Fig. 4. Lee-Carter Plus Cohorts. Left Panel: Mortality Factors by Age for Each Period before Cohort Factors. Right Panel: Cohort Factors with Trend Line

Comparing the cohort graphs for the linear HPC model in Figure 3 with that for the Lee-Carter plus cohorts model in Figure 4 shows the effects of the degree of shrinkage on the slope changes. Both graphs have slope changes at the same places, but they are generally smaller in Figure 4. Letting trend vary by age apparently picks up some of the patterns that cohorts were accounting for in the linear APC model.

Although varying the trend by age improves the fit, a problem with the Lee-Carter model is that each age gets the same percentage of the annual trend every year. It is quite possible that the trends for different ages are not proportional over time. Figure 5 shows the raw cumulative trend since 1999 for several age groups. The left panel divides the data into age ranges of seven years each. The right panel shows the ages above and below 40.5. In both panels, the younger ages show a steadier upward trend, while the older ages trend more slowly for a while then accelerate near the end. This is a somewhat systematic deviation from the Lee-Carter assumption, although it has cohort influences in it.

The Hunt-Blake model with multiple trends is designed to address violations of the Lee-Carter trend-by-age assumption. It is flexible as to how varying trend can be treated. Here we try a simple version, where all ages stay included in the main trend, adding a secondary trend that affects only ages above 40 and is otherwise constant by age.

This model ends up with 7 age variables, 9 period variables, 19 variables each for cohorts and age weights to trend, and 13 for the secondary trend, and so 67 variables in total. The looic is 3778.4, which is an improvement of 50 over Lee-Carter with cohorts. NLL is 3717.9, so the penalty is 60.5. The extra parameters now have increased the penalty slightly, but looic and NLL are both quite a bit lower. Figure 6 shows the age weight factors for trend for this and the Lee-Carter plus cohorts model, and the secondary trend factors that apply to the older ages only. The latter factors are fairly small – ranging from 0.915 to 1.02. The cohort parameters and the resulting progress of the ex-cohort mortality curve are as in Figure 4.

Fitted Poisson means by age and year are in Figure 7, shown as lines, with the data shown as points. The fitted means combine the curve excluding cohort effects, which moves upward each year, especially for younger ages, with the cohort curve, which moves one year to the right at each step. (It is actually the reflection of the cohort curve across a vertical axis that so moves, as the earlier cohorts represent the older ages.) The fits show seemingly random fluctuation of the points around the lines, but some areas could possibly benefit from still better models. More detailed modeling of the age weights for both trends would be a possible direction. However we stop here, as the example has shown the application of parameter shrinkage of slope changes and the benefits of cohort, trend-weight, and multiple trend variables in capturing the behavior of this data, and the second trend is already fairly subtle.

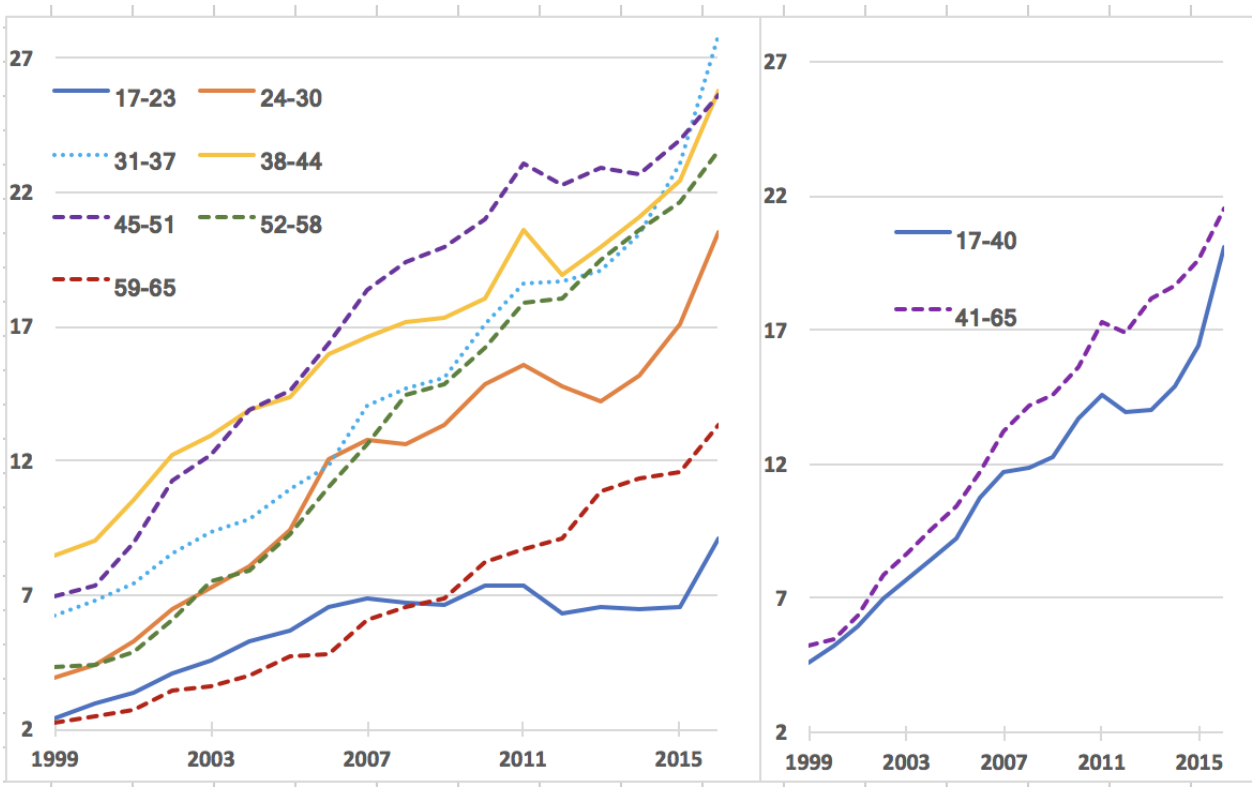

Fig. 5. Cumulative Trend by Age Group. Left Panel:7-Year Groups. Right Panel: Split at Age 41

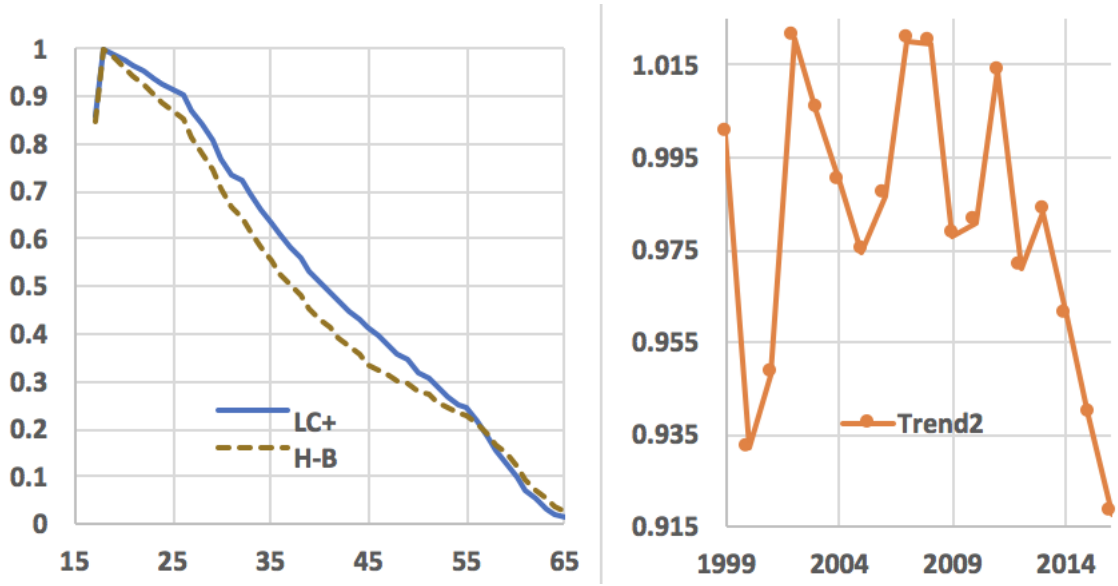

Fig. 6. Hunt-Blake Model Left Panel: Trend Weights by Age for Main Trend Comparison with Lee-Carter Plus Cohorts. Right Panel: Secondary Trend Factor for Ages 41+

Discussion

The bilinear APC models provide better accounts of the data trends, but in some ways the linear model is more intuitive. Its mortality curve by age looks more like the raw curves do, and its cohort effects largely track the two-wave pattern in the data over time. The bilinear versions have the peak age becoming younger throughout the observation period, with a fairly flat age curve for ages 30–50 until after 2010. Then the ages in the 30s start to predominate, although the cohort effects prevent this from showing up in the raw data before 2014, when the raw peak rate starts a rapid transition from the mid-40s to the mid-30s. The bilinear model cohort curve is similar to that for the linear model, but the second peak around 1980 is much less pronounced in it.

In the bilinear models, the wave on the right side of the observed rates in Figure 1 is accounted for as the boomer effect. The peak on the left comes from the higher trend for the younger years. The dip between is somewhat from Gen X cohort rates being less than those for the early Millennials, but is also from the stepping to the right each year of the generally declining post-boomer cohort rates combined with the trend growing more strongly on the left.

The highest age-trend rate is for ages 18-22, which are known to be peak years for risk-taking behavior in males. Mortality in general has a local peak for this group. A bit wider range may be similarly affected. For instance, auto insurance rates tend to gradually decline until age 30. One reason that the peak overdose mortality was previously at older ages may have been increasing availability by age. This seems to no longer be the case. An internet search for "fentanyl online" appears to be all it takes to find sources. The shift to earlier ages could thus be a risk-taking plus availability effect.

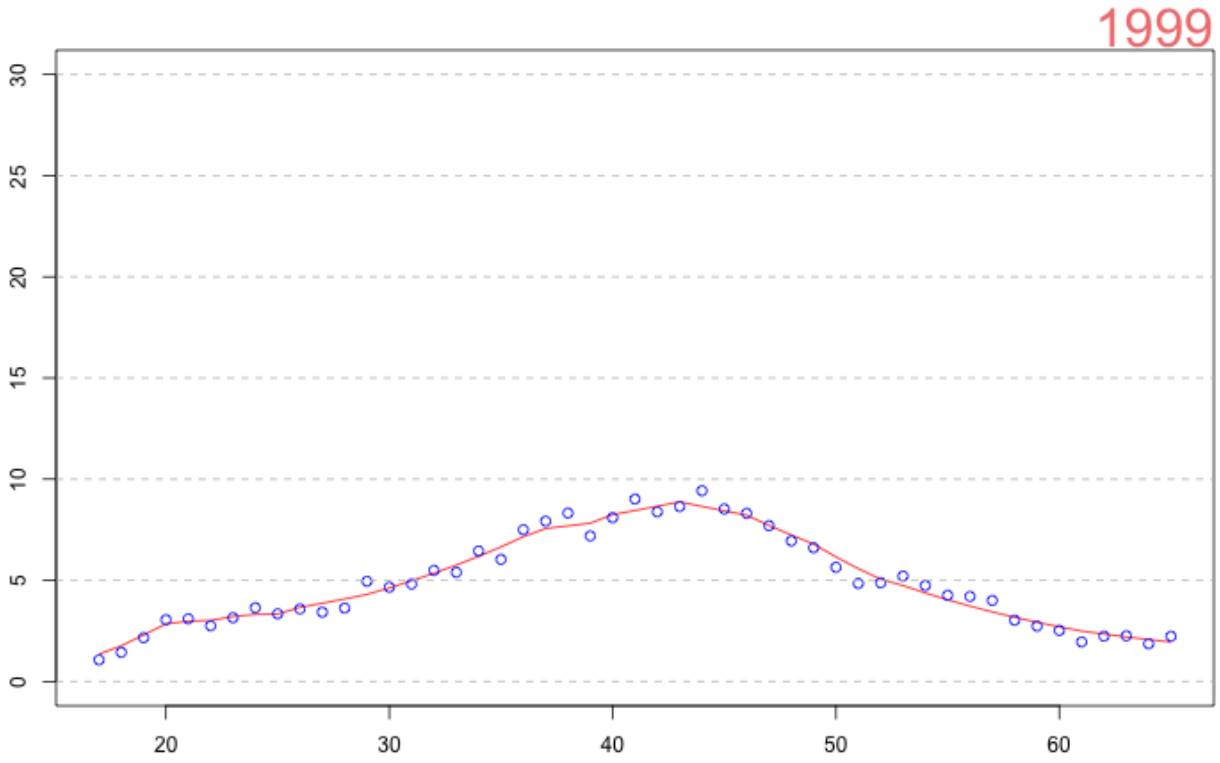

Fig. 7: Data (Circles) and Estimated Poisson Means, with Focus on 1999, 2003, 2007, 2013, and 2016 (click to start; hold click to stop)

Both Miech et al. (2017) and Schulenberg et al. (2017) discuss cohort effects in drug use. They identify relative cohort trends by changes in drug use and attitudes about it that go opposite to the secular trend, or accelerate beyond it. They do not attempt to quantify the trends specifically. One finding from the second volume is similar to some of the results here on mortality: "... in 1996 and 1997 both 10th and 12th graders

actually had higher annual prevalence levels for illicit drug use ... than either college students or all young adults. This changed somewhat after 2001, as the earlier, heavier-using cohorts of adolescents began to comprise the college student and young adult populations, while at the same time use among the incoming secondary school students was declining." Approximating the year of birth by year − grade − 5 gives the higher-use cohorts as 1978–1982, with their predecessors being the Gen X group, and their successors those born five-plus years later. Also this study finds that attitudes towards drug risk parallel usage.

The fact that these cohort changes are so apparent makes the stronger effects in the linear model more appealing, especially for Gen X. The bilinear model explains the Gen X trends as only partly from cohort effects, as it is also driven by the shift in age rates. However the bilinear model better accounts for the fact that the waves and dip between in the raw data move to the right a bit more slowly than the cohorts are aging.

Going forward, the population above 30 is aging away from the highest risk ages, and those younger than that are from cohorts with lower rates. While the trend to higher mortality rates appears to be continuing, the model suggests that the peak age range will be widening.

Causes of the cohort differences are beyond the scope of this paper, but possible research directions on that include:

- The baby boomers had a new outlook on drug use, but since then the cohort effects are similar to the cohorts 20-25 years earlier, which raises the possibility of parental influence. The parents of Gen X were early boomers or pre-boomers, and their children are late Millennials or even later cohorts. The early Millennials parents would have been baby boomers.
- Military service apparently has an impact on overdose mortality. Many Iraq war veterans would be early Millennials.
- Entering the workforce during financial downturns seems to depress long-term earnings, especially at lower income levels. This would create differentials among cohorts in economic fortunes, which could easily be related to drug use.

Bohnert, Ilgen, and Galea (2011) discuss veterans' drug mortality, and Oreopoulos, Wachter, and Heisz (2006) study career implications of economic conditions when entering the workforce.

Conclusions

Adding cohorts and multiple trends to the Lee-Carter model increases the number of parameters, which is high to start with. Emerging statistical methodologies for parameter reduction provide tools for using more complex models like the ones here while holding down the effective number of parameters. Estimating second differences of the usual APC parameters enables parameter reduction for these models, as eliminating these slope changes just continues previous linear trends.

We reviewed the framework of lasso and Bayesian lasso then applied these to fit the Hunt-Blake generalization of the Lee-Carter model to opioid mortality rates. The fitting proceeded by using classical lasso initially to identify parameters that could be eliminated from the linear age-period model, then fitting that with Bayesian lasso. We added cohort effects, then age-sensitive trends, and finally multiple trends. Goodness of fit improved at each step. This sequential approach led to a reasonably-fitting model, but is not guaranteed to find the optimal fit. With 49 ages, each with both an age parameter and an age weight for the trends, 18 periods with two trends, and 65 cohorts, there were nearly 200 APC parameters. Putting these on piecewise linear curves and shrinking the slope changes ended up with a parameter penalty to the NLL of 60.5, and so probably fewer effective parameters than that.

The later stages of the baby-boom generation had the highest cohort parameters, with a smaller peak around 1980. The cohorts before 1950 and after 1990 have the lowest parameters. Absent cohort effects, the ages under 30 end up with the greatest mortality risk, but in the raw data it is currently ages 30–40 that appear to be most vulnerable.

Acknowledgements: Thanks to Qiuli Tang of Columbia University for key contributions to the graphics.

References

Barnett, Glen, and Ben Zehnwirth. 2000. "Best Estimates for Reserves." *Proceedings of the Casualt Actuarial Society* 87: 245–303.

Bijlsma, Maarten J., Rhian M. Daniel, Fanny Janssen, and Bianca L. De Stavola. 2017. "An Assessment and Extension of the Mechanism-Based Approach to the Identification of Age-Period-Cohort Models." *Demography* 54:2: 721–43.

Bohnert, Amy S. B., Mark A. Ilgen, and Sandro Galea. 2011. "Accidental Poisoning Mortality Among Patients in the Department of Veterans Affairs Health System." *Medical Care* 49(4): 393–96.

Bühlmann, Hans. 1967. "Experience Rating and Credibility." *ASTIN Bulletin* 4:3: 199–207.

Fienberg, S. E., and W. M. Mason. 1978. "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data." *Sociological Methodology* 10: 1–67.

Gao, Guangyuan, and S. Meng. 2017. "Stochastic Claims Reserving via a Bayesian Spline Model with Random Loss Ratio Effects." *ASTIN Bulletin*.

Gelfand, A. E. 1996. "Model Determination Using Sampling-Based Methods." *Markov Chain Monte Carlo in Practice, Ed. W. R. Gilks, S. Richardson, D. J. Spiegelhalter* London: Chapman and Hall: 145–62.

Greenberg, B. G., John J. Wright, and Cecil G. Sheps. 1950. "A Technique for Analyzing Some Factors Affecting the Incidence of Syphilis." *Journal of the American Statistical Association* 45:251: 373–99.

Hoerl, A.E., and R. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12: 55–67.

Hunt, Andrew, and David Blake. 2014. "A General Procedure for Constructing Mortality Models." *North American Actuarial Journal* 18 (1): 116–38.

Miech, Richard A., Lloyd D. Johnston, Patrick M. O'Malley, Jerald G. Bachman, John E. Schulenberg, and Megan E. Patrick. 2017. "Monitoring the Future: National Survey Results on Drug Use, 1975–2016 Volume I, Secondary School Students." *The University of Michigan Institute for Social Research*.

Mowbray, Albert H. 1914. "How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium." *Proceedings of the Casualty Actuarial Society* 1: 24–30.

National Center for Health Statistics, CDC. 2017. "CDC: Wide-Ranging Online Data for Epidemiologic Research, Detailed Mortality Files." *(WONDER)*.

Oreopoulos, Philip, Till von Wachter, and Andrew Heisz. 2006. "The Short- and Long-Term Career Effects of Graduating in a Recession: Hysteresis and Heterogeneity in the Market for College Graduates." *NBER Working Papers* 12159.

Schulenberg, John E., Lloyd D. Johnston, Patrick M. O'Malley, Jerald G. Bachman, Richard A. Miech, and Megan E. Patrick. 2017. "Monitoring the Future: National Survey Results on Drug Use, 1975–2016 Volume Ii, College Students & Adults Ages 19–55." *The University of Michigan Institute for Social Research*.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator of the Mean of a Multivariate Normal Distribution." *Proceedings of the Third Berkeley Symposium* 1: 197–206.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic." *Journal of Statistcs and Computing* 27:5: 1413–32.

Venter, Gary, and Şule Şahin. 2017. "Parsimonious Parameterization of Age-Period-Cohort Models by Bayesian Shrinkage." *Astin Bulletin*.