Robust Paradigm Applied to Parameter Reduction in Actuarial Triangle Models

Gary Venter

Abstract The recognition that models are approximations used to illuminate features of more complex processes brings a challenge to standard statistical testing, which assumes the data is generated from the model. Out-of-sample tests are a response. In my view this is a fundamental change in statistics that renders both classical and Bayesian approaches outmoded, and I am calling it the "robust paradigm" to signify this change. In this context, models need to be robust to samples that are never fully representative of the process. Actuarial models of loss development and mortality triangles are often over-parameterized, and formal parameter-reduction methods are applied to them here within the context of the robust paradigm.

Keywords Loss reserving • Mortality • Bayesian shrinkage • MCMC

1 Introduction

Section 2 discusses model testing under the robust paradigm, including out-of-sample tests and counting the effective number of parameters. Section 3 introduces parameter-reduction methods including Bayesian versions. Section 4 reviews actuarial triangle modeling based on discrete parameters by row, column, etc., and how parameter-reduction can be used for them. Section 5 gives a mortality model example, while Sect. 6 illustrates examples in loss reserving. Section 7 concludes.

2 Model Testing Within the Robust Paradigm

Both Bayesian and classical statistics typically assume that the data being used to estimate a model has been generated by the process that the model specifies. In many, perhaps most, financial models this is not the case. The data is known to come

G. Venter (✉)
University of New South Wales, Sydney, NSW, Australia
e-mail: gary.venter@gmail.com

J.A. Londoño et al. (eds.), *Actuarial Sciences and Quantitative Finance*, Springer
Proceedings in Mathematics & Statistics 214, https://doi.org/10.1007/978-3-319-66536-8_1

from a more complex process and the model is a hopefully useful but simplified representation of that process. Goodness-of-fit measures that assume that the data has been generated from the sample are often not so reliable in this situation, and out-of-sample tests of some sort or another are preferred. These can help address how well the model might work on data that was generated from a different aspect of the process. I have coined the term "robust paradigm" to refer to statistical methods useful when the data does not come from the model.

Much statistics today is based on pragmatic approaches that keep the utility of the model for its intended application in mind, and regularly deviate from both pure Bayesian and pure classical paradigms. That in itself does not mean that they are dealing with data that does not come from the models. In fact, even out-of-sample testing may be done purely to address issues of sample bias in the parameters, even assuming that the data did come from the model. But simplified models for complex processes are common and pragmatic approaches are used to test them. This is what is included in the robust paradigm.

When models are simplified descriptions of more complex processes, you can never be confident that new data from the same process will be consistent with the model. In fact with financial data, it is not unusual for new data to show somewhat different patterns from those seen previously. However, if the model is robust to a degree of data change, it may still work fairly well in this context. More parsimonious models often hold up better when data is changing like that. Out-of-sample testing methods are used to test for such robustness.

A typical ad hoc approach is the rotating $\frac{4}{5}$ths method: the data is divided, perhaps randomly, into five subsets, and the model is fit to every group of four of these five. Then the fits are tested on the omitted populations, for example by computing the negative loglikelihood (NLL). Competing models can be compared on how well they do on the omitted values.

A well-regarded out-of-sample test is leave one out, or "loo." This fits the model many times, leaving out one data point at a time. Then the fit is tested at each omitted point to compare alternative models. The drawback is in doing so many fits for each model.

In Bayesian estimation, particularly in Markov Chain Monte Carlo (MCMC), there is a shortcut to loo. The estimation produces many sample parameter sets from the posterior distribution of the parameters. By giving more weight to the parameter sets that fit poorly at a given data observation, an approximation to the parameters that would be obtained without that observation can be made. This idea is not new, but such approximations have been unstable.

A recent advance, called Pareto smoothed importance sampling, appears to have largely solved the instability problem. A package to do this, called `loo`, is available with the `Stan` package for MCMC. It can be used with MCMC estimation not done in `Stan` as well. It allows comparison of the NLL of the omitted points across models. This modestly increases the estimation time, but is a substantial improvement over multiple re-estimation. Having such a tool available makes loo likely to become a standard out-of-sample fitting test.

This is a direct method to test models for too many parameters. Over-fitted models will not perform well out of sample. If the parameters do better out of sample, they are worth it. Classical methods for adjusting for over-parameterization, like penalized likelihood, are more artificial by comparison, and never have become completely standardized. In classical nonlinear models, counting the effective number of parameters is also a bit complex.

2.1 Counting Parameters

In nonlinear models it is not always apparent how many degrees of freedom are being used up by the parameter estimation. One degree of freedom per parameter is not always realistic, as the form of the model may constrict the ability of parameters to pull the fitted values towards the actual values.

A method that seems to work well within this context is the generalized degrees of freedom method of Ye (1998). Key to this is the derivative of a fitted point from a model with respect to the actual point. That is the degree to which the fitted point will change in response to a change in the actual point. Unfortunately this usually has to be estimated numerically for each data point.

The generalized degrees of freedom of a model fit to a data set is then the sum across all the data points of the derivatives of the fitted points with respect to the actual points, done one at a time. In a linear model this is just the number of parameters. It seems to be a reasonable representation of the degrees of freedom used up by a model fit, and so can be used like the number of parameters is used in linear models to adjust goodness-of-fit measures, like NLL. A method of counting the effective number of parameters is also built into the `loo` package.

3 Introduction to Parameter Reduction Methodology

Two currently popular parameter reduction methodologies are:

- Linear mixed models (LMM), or in the GLM environment GLMM
- Lasso—Least Absolute Shrinkage and Selection Operator

3.1 Linear Mixed Models

LMM starts by dividing the explanatory variables from a regression model into two types: fixed effects and random effects. The parameters of the random effects are to be shrunk towards zero, based perhaps on there being some question about whether or not these parameters should be taken at face value. See for example Lindstrom and Bates (1990) for a discussion in a more typical statistical context.

Suppose you are doing a regression to estimate the contribution of various factors to accident frequency of driver/vehicle combinations. You might make color of car a random effect, thinking that probably most colors would not affect accident frequency, but a few might, and even for those you would want the evidence to be fairly strong. Then all the parameters for the car color random effects would be shrunk towards or to zero, in line with this skepticism but with an openness to being convinced.

This could be looked at as an analysis of the residuals. Suppose you have done the regression without car color but suspect some colors might be important. You could divide the residuals into groups by car color. Many of these groups of residuals might average to zero, but a few could have positive or negative mean—some of those by chance, however. In LMM you give color i parameter b_i and specify that b_i is normal with mean zero and variance $d_i\sigma^2$, where σ^2 is the regression variance and d_i is a variance parameter for color i. LMM packages like in SAS, Matlab, R, etc. generally allow a wide choice of covariance matrices for these variances, but we will mainly describe the base case, where all of them are independent.

The d_i's are also parameters to be estimated. A color with consistently high residuals is believably a real effect, and it would be estimated with a fairly high d_i to allow b_i to be away from zero. The b_i's are usually assumed to be independent of the residuals. LMM simultaneously maximizes the probability of the b_i's, $P(b)$, and the conditional probability of the observations given b, $P(y|b)$, by maximizing the joint likelihood $P(y, b) = P(y|b)P(b)$.

For a b_i parameter to get further from zero, it has to improve the likelihood of the data by more than it hurts the density of the b's. This is more likely if the previous residuals for that color are grouped more tightly around their mean, in the color example. Then the parameter would help the fit for all those observations. That clustering is not what is measured by d_i, however. It instead determines how much b_i could differ from zero, and its estimate increases to accommodate a useful b_i.

3.2 Lasso

Lasso is a regression approach that constrains the sum of the absolute values of the parameters. It is related to ridge regression, which limits the sum of squares of the parameters. In practice with a lot of variables included, Lasso actually shrinks a fair number of the parameters to zero, eliminating them from the model, whereas ridge regression ends up with many small parameters near zero. Lasso is preferred by most modelers for this reason, and is also preferable to stepwise regression.

In its standard application, all the parameters except the constant term are shrunk, although there is no reason some parameters could not be treated like fixed effects and not shrunk. See Osbourne et al. (2000) for an introduction. Also Pereira et al. (2016) gives examples more general than standard regression.

To make the competition among the independent variables fair, all of them are standardized to have mean zero and variance one by subtracting a constant and dividing by a constant. The additive transform gets built into the constant term of the regression, and the multiplicative one scales the parameter of that variable.

Then what is minimized is the NLL plus a selected factor times the sum of the absolute values of the parameters. The selection of the factor can be subjective—several are tried with the resulting models evaluated by expert judgment. Using loo to see how well the models with different factors do on the omitted points is more highly regarded, but in a classical setting requires a lot of re-estimation, depending on the sample size.

3.3 Problem with LMM: All Those Variances

Counting parameters is an issue with classical Lasso and LMM. For both, fewer degrees of freedom are used than the apparent number of parameters, due to the constraints. For LMM there is a partial shortcut to counting parameters.

In a regression, the so-called hat matrix is an $N \times N$ matrix, where N is the sample size, which can be calculated from the matrix of independent variables—the design matrix. Multiplying the hat matrix on the right by the vector of observations gives the vector of fitted values. The diagonal of the hat matrix thus gives the response of a fitted value to its observation, and in fact is the derivative of the fitted value with respect to the actual value.

The sum of the diagonal of the hat matrix is thus the generalized degrees of freedom. This holds in LMM as well, but only conditional on the estimated variances. Thus the degrees of freedom used up in estimating the variances do not show up in the hat matrix.

Different LMM estimation platforms can give slightly different parameters—but usually with fits of comparable quality. One triangle model we fit, similar to those discussed below, nominally had 70 parameters, not including the variances. We fit it with two methods. Using the diagonal of the hat matrix indicated that 17.3 degrees of freedom were used by one fitting method, and 19.9 by the other. The second one had a slightly lower NLL, and the penalized likelihoods, by any methods, were comparable.

Since these parameter counts are conditional on the estimated variances d_i, we then did a grind-out generalized-degrees-of-freedom calculation by re-estimating the model changing each observation slightly, one at a time. That got the variances into the parameter counts. The same two methods as before yielded 45.1 and 50.7 degrees of freedom used, respectively. That means that 27.8 and 30.8 degrees of freedom, respectively, were used up in estimating the variances.

In essence, the fitted values responded much more to changes in the actual values than you would have thought from the hat matrix. The parameter reduction from the apparent 70 original parameters was much less than it at-first appeared to be. For the models we were fitting we concluded that base LMM with variances estimated for each parameter was not as effective at parameter reduction as we had thought. This lends more support to using Lasso, or perhaps LMM with fewer, perhaps just a single, variance to estimate. That is, you could assume the d_i are all the same.

3.4 Bayesian Parameter Reduction

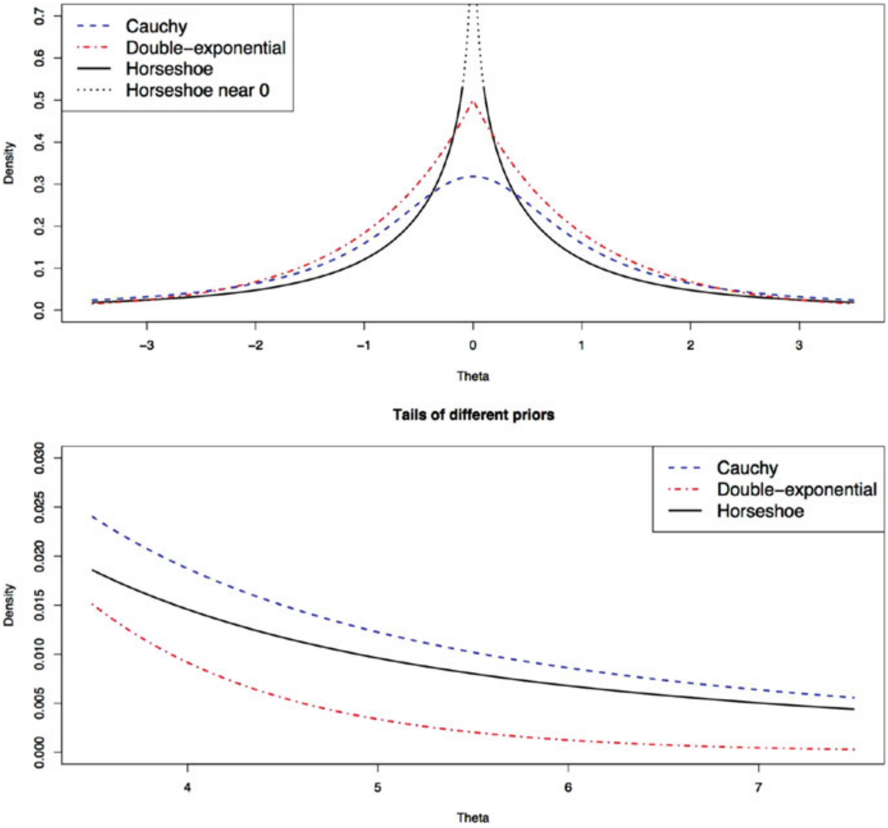
A way to shrink parameters towards zero in Bayesian models is to use shrinkage priors. These are priors with mean zero and fairly low variances, so tend to prioritize smaller values of the parameters. An example is the Laplace, or double exponential, distribution, which is exponential in x for $x > 0$ and in $-x$ for $x < 0$:

$$x > 0 : f(x) = e^{-x/b}/2b \tag{1}$$

$$x < 0 : f(x) = e^{x/b}/2b \tag{2}$$

This has heavier tails and more weight near zero than the normal has. Even more so is the horseshoe distribution, which is a normal with σ^2 mixed by a Cauchy.

Typically shrinkage priors are used in MCMC estimation (Fig. 1). There is a lot of flexibility available in the choice of the variances. They can all be the same,



Tails of different priors

Fig. 1 Shrinkage priors

which is Lasso-like, or vary for different parameters. Some or all of the parameters can have shrinkage priors. Thus the distinctions between LMM and Lasso are not so meaningful in MCMC. There is a wide variety of approaches that can be used.

One fairly viable approach is to use the same variance in the shrinkage priors for all the parameters, and then use loo to see approximately what this variance should be to get the best out-of-sample performance.

3.5 Non-informative Priors

For parameters you do not want to shrink, if you have information or beliefs about a reasonable range for the parameter, that can be coded into the prior distribution. A convenient alternative is non-informative priors. For instance in `Stan`, for a parameter that could be positive or negative, if a prior is not specified the prior is assumed to be uniform on the real line.

This prior density is infinitesimal everywhere and in fact is just specified as being proportional to 1. In `Stan` it is typical to omit constants of proportionality, even if they are not real numbers. This prior, however, viewed as a prior belief, is patently absurd. Most of the probability would lay outside of any finite interval, so it is like saying the parameter probably has a very high absolute value, but we don't know if it is positive or negative.

Nonetheless using it as a prior tends to work out well. Posterior variances from it are often quite similar to what classical statistics would give for estimation variances. Thus the results seem familiar and reasonable. In essence, the prior ceases to be an opinion about the parameter, and instead is chosen because it tends to work out well. This is further evidence that we are no longer in the realm of either classical or Bayesian statistics—it is a pragmatic focus more than a theoretical one.

Things get more awkward when a parameter has to be positive. Assuming uniformity on the positive reals is problematic. While the uniform on the real line has infinite pulls both up and down, on the positive reals the infinite side is only an upward pull. There is thus a tendency for this prior to give a higher estimate than classical statistics would give.

An alternative is to use a prior proportional to $1/x$. This diverges at zero and infinity, so pulls infinitely in both directions. It tends to produce estimates similar to classical unbiased estimates. It is equivalent to giving the log of the parameter a uniform distribution on the reals, which is the easiest way to set it up in `Stan`.

People who do not like non-informative priors sometimes use very diffuse proper priors. One example can be written Gamma(0.001, 0.001). It has mean one and standard deviation about $31\frac{5}{8}$. It is, however, a quite strange prior. Even though the mean is one, the median is in the vicinity of 10^{-300}. The 99th percentile is about 0.025, while the 99.9th is 264 and the 99.99th is 1502. Thus it strongly favors very low values, with occasional very high values showing up. It usually works out alright in the end but can cause difficulty in the estimation along the way.

4 Actuarial Triangle Models with Time Variables

Data for the evolution of insurance liabilities and for mortality can be arranged in two-dimensional arrays, for example with rows for year of origin, and columns for lag to extinction. Actually the time periods are not always years—they could be quarters, months, or even days—but here we will call them years for simplicity. For liabilities, year of origin is often the year the event happened, and lag is the time it takes to close the case and make final payments. For mortality, year of origin is year of birth and lag is the number of years lived. For mortality, the rows are sometimes taken as the calendar years that the extinctions occur in, which is just a different arrangement of the same data—the diagonals are rotated to become the rows, and vice versa.

A common arrangement within the array has the data all above the SW—NE diagonal, giving the term triangle, but various shapes are possible. Mortality triangles for a population usually contain the ratio of deaths in the year to the number alive at the start of the year. Liability triangle cells could contain incremental or cumulative claims payments or claims-department estimates of eventual payments. Here we will assume they are incremental paid losses and are positive or blank.

A popular class of models the log of each entry as the sum of a row effect and a column effect—so there is a dummy variable and a parameter for each row and each column. It is also not unusual to have a parameter for each calendar year, which is the year of origin plus the lag (assuming beginning at lag zero). The calendar-year effects are trends—perhaps inflation for liabilities and increased longevity over time for mortality. It is fairly common in mortality modeling to allow for different ages to feel the longevity improvement more strongly, so an additional parameter might be added for each age as a multiplier to the trend to reflect how strongly that age benefits from the trend.

In doing this modeling actuaries have found that trends in longevity sometimes affect different ages differently, so a single pattern of age-responsiveness does not always hold. To account for this, models now allow a few calendar-year trends, each with its own impact by age. Some models also allow for interaction of age with the year-of-birth cohort parameters, but this effect does not seem to be consistent across cohorts and is less common in recent models. Even in the liability models there could be changes in the lag effects over time, which could be modeled by interactions of lag with year of origin or calendar year.

Letting $p[n]$ be the year-of-origin parameter for year n, $q[u]$ be the age parameter for age u, r refer to a calendar year trend, and s be a set of age weights, the model for the logged value in the n, u cell can be expressed as:

$$y[n,u] = p[n] + q[u] + \sum_i r_i[n+u]s_i[u] + \varepsilon_{n,u} \tag{3}$$

The sum is over the various trends. With a single trend and no age interaction with trend, this would be a typical liability emergence model. There it is not unusual to even leave off the trend entirely—for instance if the trend is constant it will project onto the other two directions.

4.1 Parameter Reduction

The model as stated so far is over-parameterized. One approach to parameter reduction is to require that nearby years or lags have similar parameters. Life insurance actuaries have tried using cubic splines for this. General insurance actuaries have independently been using linear splines. That is, differences between adjacent parameters (i.e., slopes) are constant for a while, with occasional changes in slope. The slope changes are thus the second differences across the parameters.

As the second differences change only occasionally, they are good candidates for parameter-reduction methods. That is the approach explored here. The slope changes are the parameters modeled with specified priors, and these accumulate to the slopes and those to levels, which are the p, q, r, s in the model equation. This can apply to long or short trend periods so can be used for both the life and the general insurance models.

The fitting was done with the `Stan` package, taking double exponential priors for the slope changes. A single variance was specified for all these priors, which in the end was determined by loo in the mortality example. Judgment was used for this in the liability example, but that is not a finished model.

5 Mortality Model Example

US mortality data before 1970 is considered of poor quality, so we use mortality rates in years 1970–2013. Cohorts 1890–1989 were modeled for ages 15–89. A model using Eq. (3) with two trends r_1 and r_2 was selected (i takes on two values: 1 and 2). The first trend is for all the years and the second is zero except for the years 1985–1996, which had increased mortality at younger ages, primarily associated with HIV, but also drug wars. The latter trend was strongest for ages 27–48, so weights were estimated for those years. Here n is the year of birth and u is the age at death, so $n + u$ is the year of death.

The model was calibrated using the MCMC package `Stan` with the second differences of the p, q, r, and s parameters given double exponential priors. Then the parameters in (3) are cumulative sums of the second differences. A lot of the second differences shrink towards zero due to this prior, so the parameters come out looking like they fall on fairly smooth curves—which are actually linear splines.

It is possible to get fairly different parameter sets with quite similar fits, so a fair number of constraints are needed for the sake of specificity. For symmetry, a constant term was added to the model, and then a base mortality parameter q, a trend parameter r, and a cohort parameter p were set to zero. The HIV trend was forced to be upward (positive), and all the trend weights were forced to be in [0,1].

It is a bit awkward in `Stan` to force these parameters to be positive. They are sums of the underlying slope parameters, which in turn are sums of slope changes. Any of those could be negative, as slopes could go up or down. Simple constraints, like using the minimum of the parameter and zero, are problematic in Stan because you then lose derivatives needed for the internal calculations. Squaring the value is awkward as well, as then different paths for the slope changes can get to the same level parameter, which makes it look like the slope changes did not converge. In the end, however, this choice is easier to deal with and was taken. Modeling the logs of the levels as piecewise linear is an alternative worth exploring.

The weights were made to stay in [0,1] by dividing them all by the highest of them after squaring. This may make finding parameters more difficult as well, and it seems to slow down the estimation considerably, but it looks like the best way to get specificity.

Cohort levels are regarded as the year-of-birth effects left after everything else is modeled, so were forced to have zero trend—just by making them the residuals around any trend that did appear.

Another problem with cohorts is that the most recent ones are seen only at young ages, which creates a possible offset with the trend change weights. In fact, giving the most recent cohorts high mortality and simultaneously giving the youngest ages high trend weights gave fairly good fits, but does not seem to be actually occurring.

In the end we forced all cohorts from 1973 to 1989 to have the same parameter—which in fact was made zero to avoid overlap with the constant term. For similar reasons, cohorts 1890–1894 all got the same parameter.

Stan simulates parameter sets from the posterior distribution of the parameters in several parallel chains—typically four of them. One check of convergence is to compare the means of each parameter across the chains, and the within and between variances. With these constraints, even though estimation was slow, all the chains had very comparable mean values for every level parameter. The slopes and slope changes from different chains sometimes look like mirror images, however, even though they have the same squares.

The parameters graphed here include all four chains as separate lines mainly to show how well they have converged, as the four lines are all very close.

The main trend is fairly steady improvement, but with a slowdown in the 1990s that is not fully accounted for by the HIV trend, and another slowdown in the last 3–4 years. The trend take-up factors by age range from 65% to 100%, and are lowest in the early 30s and the late 80s (Fig. 2).

The HIV trend is highest in the mid-1990s just before treatments became available (Fig. 3). The ages most affected are the 30s (Fig. 4).

The cohort parameters show a fair degree of variation over time (Fig. 5). Relative to trend, etc. the most longevity is seen in those born in the 1940s and before 1910, with a dip around 1970 as well. While thorough modeling of these patterns is a future project, some clues are available in demographic, macroeconomic, and public health events (Fig. 6).

Those born in 1900 were 70 by the start of this data. The portion of this group that got that old seems to have been particularly hardy. In fact they displayed as much chance to get from age 70 to 90 as those born decades later. The cohort parameters would reflect only the ages in the dataset, so are not necessarily indicative of the cohort mortality for earlier ages.

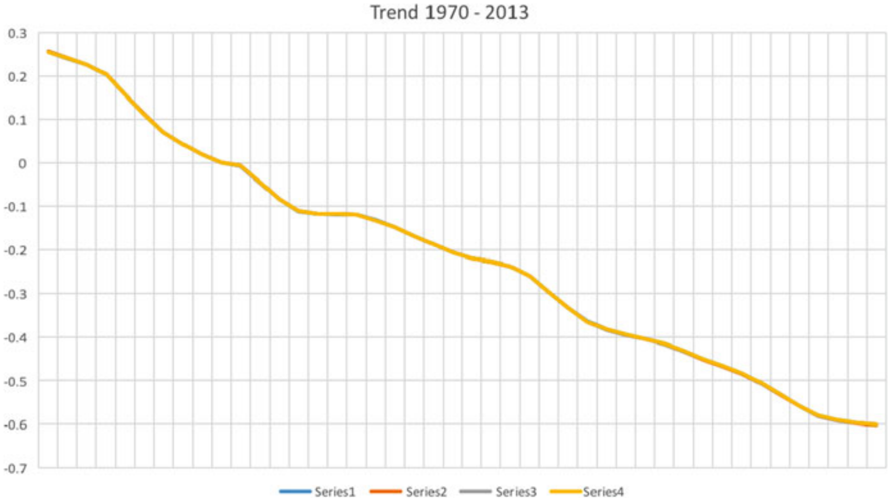


Fig. 2 Time trend 1970–2013

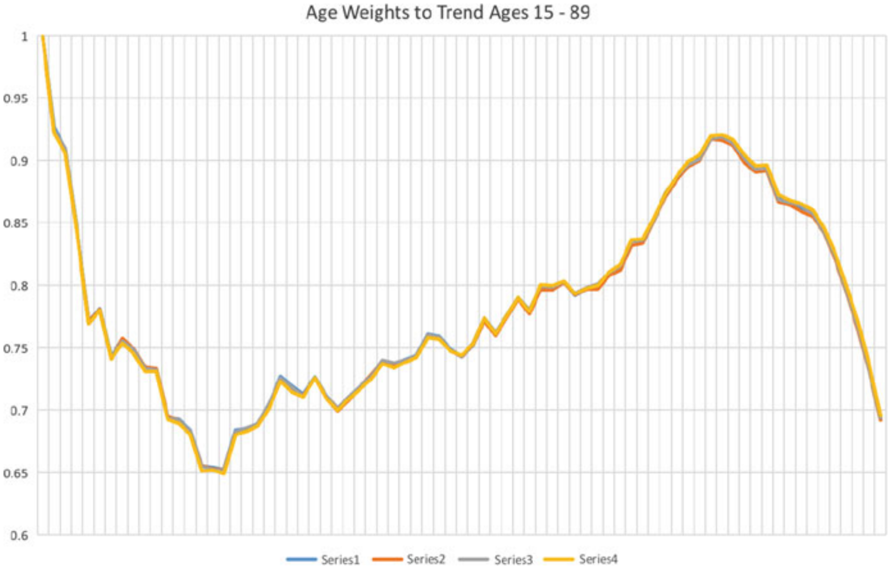


Fig. 3 Age weights to trend ages 15–89

The group born in the 1930s and early 1940s is called the silent generation, or sometimes the fortunate few, and is a unique population. They have had by far the highest real income and net worth of any American generation. This is often attributed to demographics—it was a relatively small population and had little

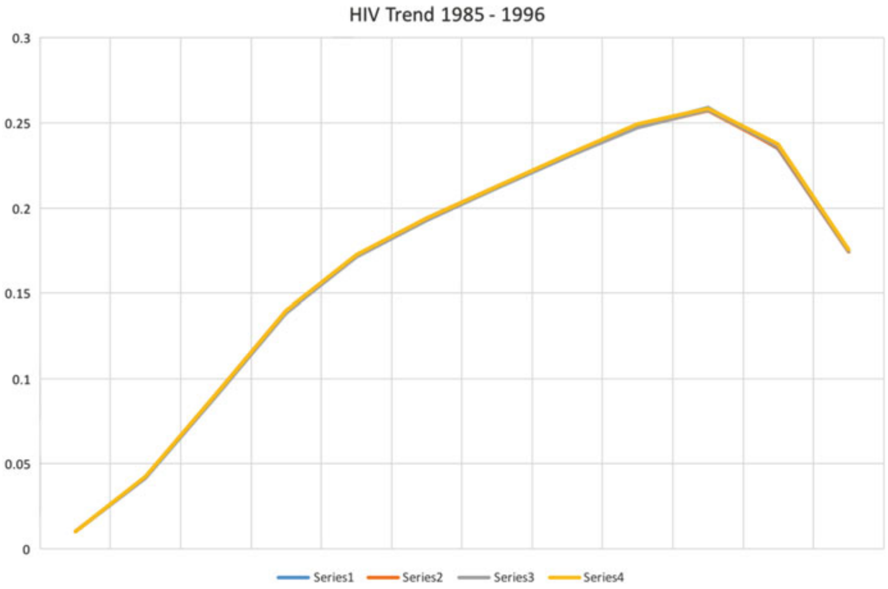


Fig. 4 HIV trend 1985–1996

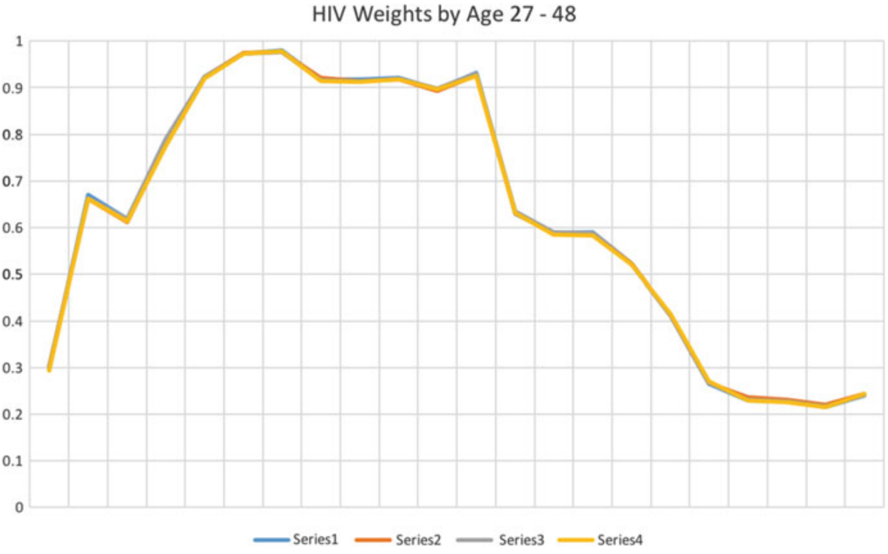


Fig. 5 Age weights to HIV trend ages 27–48

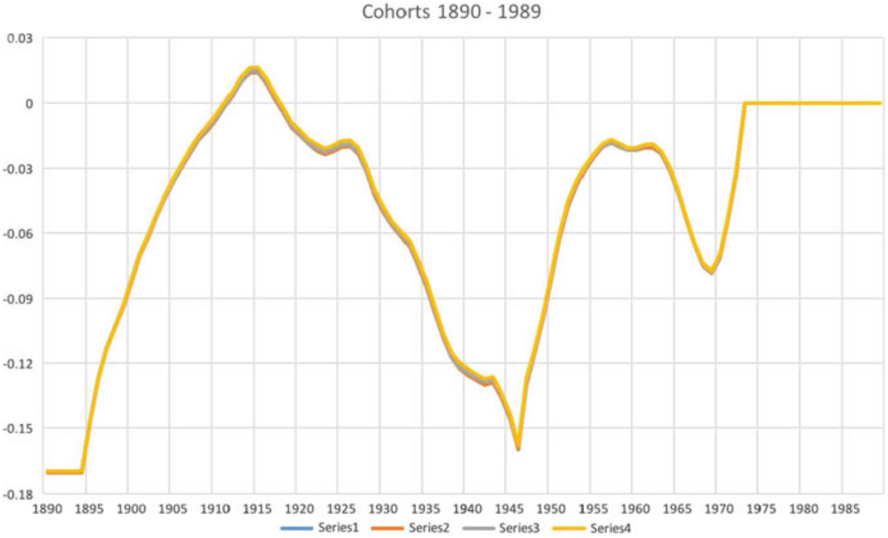


Fig. 6 Cohort level parameters for years of birth 1890–1989

workplace competition from earlier generations. Wealth is linked to longevity and if that were the entire story, this set of cohorts would have had the lowest mortality rates.

However, this was also a generation of heavy smokers. The early boomers, born in the late 1940s, probably smoked less, and had some of the demographic advantages of the fortunate few. The early-boomer cohort may also have been a bit less exposed to obesity than the next group.

Having a small or shrinking population five or so years older seems to be good for career opportunities. Being from the mid-1940s group, I can say that many in my cohort stepped into easily available leadership roles, and hung in there for 30–40 years. The mid-50s cohorts were always back there one level lower—although individual exceptions abound.

The cohorts around 1970 were part of a slowing of population growth that probably also lead to ample career opportunities. Another determinant of career wealth accumulation and so average mortality is the state of the economy upon entering the workforce. That would be another factor to include in this study (Fig. 7).

Looking at the raw mortality rates by age (across) and cohort (down) shows how the age pattern of mortality has been evolving. The width of that graph at an age shows how much mortality improvement that age has experienced from 1970 to 2013.

One thing that stands out is the clumping of lines at the upper right. For most of this period there was little change in the mortality rates at older ages. Then in the last 10 or 11 years, mortality in this group started reducing considerably. This looks

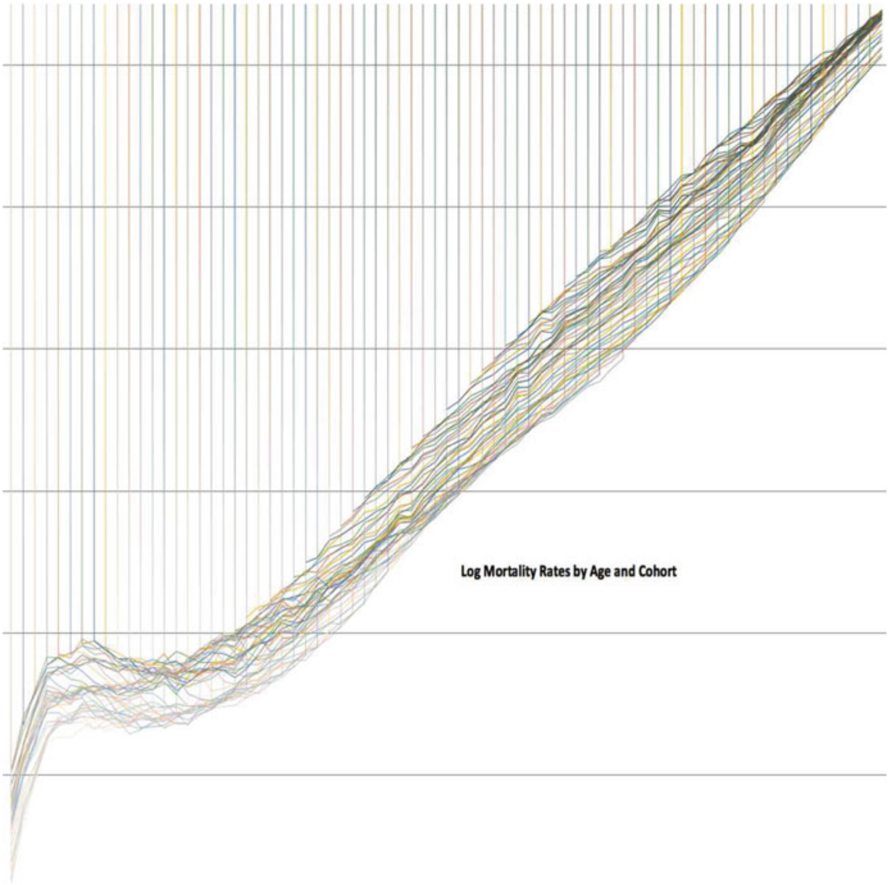


Fig. 7 Log mortality rates by age (increasing from left to right) and cohort (individual lines, most recent generally lower)

like another candidate for a separate trend. Probably the way to do this is to have a separate upward trend in mortality for ages 75+ before 2002, and then give this group the overall trend after that.

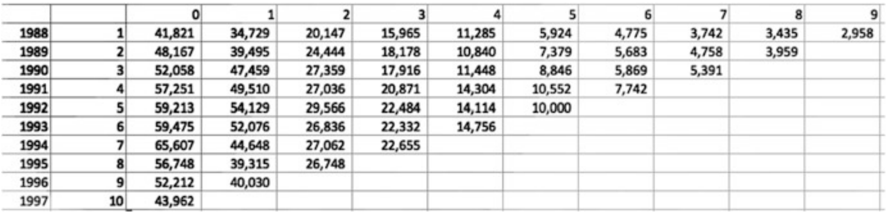
Another new trend since 2000 or so is to find little or no improvement in mortality rates for ages in the late 40s through early 60s. This shows up as a clumping of lines at the bottom of the graph above the word "Log." This actually is producing higher mortality for some parts of the population, as has been reported widely in the press. (Our data does not have subpopulation breakouts.) It is again a candidate for its own trend. However, this is also the mid-to-late boomer cohort, which shows up having higher mortality rates anyway, and was also impacted by HIV, so there could be a combination of effects here. Nonetheless, the cohort effect is supposed to be after all other trends have been accounted for, so it seems appropriate to put in a trend here and see what it does to the cohorts.

6 Reserve Modeling Example

Loss reserving has much smaller triangles than mortality does—usually—and simpler models—only one trend and no trend weights by lag typically.

$$y[n,u] = p[n] + q[u] + r[n+u] + \varepsilon_{n,u} \tag{4}$$

We explore here a bit broader model, but will start off with the above. Below is a worker's compensation loss paid triangle for a New Jersey insurer from Taylor and McGuire (2016). The cells are incremental payments.

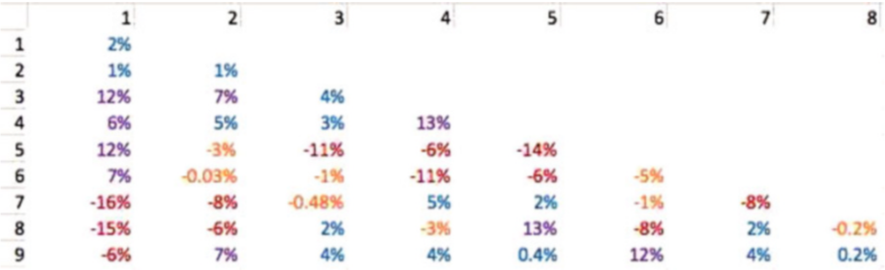


		0	1	2	3	4	5	6	7	8	9
1988	1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
1989	2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
1990	3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		
1991	4	57,251	49,510	27,036	20,871	14,304	10,552	7,742			
1992	5	59,213	54,129	29,566	22,484	14,114	10,000				
1993	6	59,475	52,076	26,836	22,332	14,756					
1994	7	65,607	44,648	27,062	22,655						
1995	8	56,748	39,315	26,748							
1996	9	52,212	40,030								
1997	10	43,962									

6.1 Exploratory Analysis

Looking at residuals from standard development factor analysis can provide information about possible changes in trend and payout patterns. The first test is to calculate the incremental/previous cumulative development factors for each cell, then subtract the column averages from the cell values.

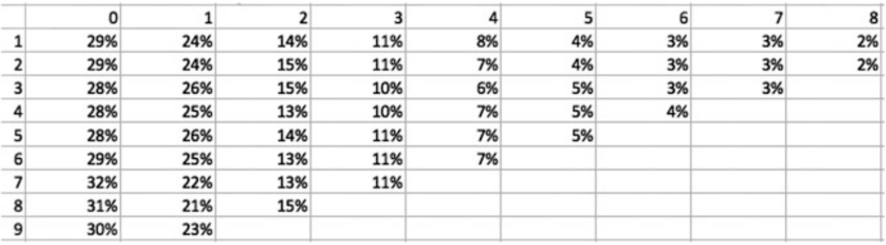
Looking at the results by diagonal can show calendar-year differences. Consistently high or low differences of individual trend factors from column averages along a given diagonal would suggest a possible cost difference for that diagonal compared to the triangle as a whole. It is easier to see such patterns by rotating the triangle so that the diagonals become rows. That was done below with some color coding, and decimals expressed as percents.



	1	2	3	4	5	6	7	8
1	2%							
2	1%	1%						
3	12%	7%	4%					
4	6%	5%	3%	13%				
5	12%	-3%	-11%	-6%	-14%			
6	7%	-0.03%	-1%	-11%	-6%	-5%		
7	-16%	-8%	-0.48%	5%	2%	-1%	-8%	
8	-15%	-6%	2%	-3%	13%	-8%	2%	-0.2%
9	-6%	7%	4%	4%	0.4%	12%	4%	0.2%

It is apparent that the first four diagonals are all positive and the next four mostly negative, with the last again positive. This is suggestive of a calendar-year trend change. The first column seems to be on its own path, however, and may be a payout-change indicator.

A look at payout patterns can be taken by developing each row to ultimate by development factors, then taking the ratio of paid in column to ultimate paid in row for each cell. This can be done for lag zero as well. This test can show changes in payout pattern, but changes in the later columns would be included in averages below that, obfuscating some of the impact.



	0	1	2	3	4	5	6	7	8
1	29%	24%	14%	11%	8%	4%	3%	3%	2%
2	29%	24%	15%	11%	7%	4%	3%	3%	2%
3	28%	26%	15%	10%	6%	5%	3%	3%	
4	28%	25%	13%	10%	7%	5%	4%		
5	28%	26%	14%	11%	7%	5%			
6	29%	25%	13%	11%	7%				
7	32%	22%	13%	11%					
8	31%	21%	15%						
9	30%	23%							

Starting with row 5, there is an increasing trend in payouts at lag 0, offset by a decreasing trend at lag 1. These might reverse slightly in row 9, but that could be due to calendar-year trend.

6.2 Modeling

The model without interaction terms does not include any provision for payout pattern changes. We start with that, however, to see what it says about calendar-year trends, and to see if those could account for the apparent payout shift. Again the double exponential distribution was used for the changes in slope, here with a fairly high variance to make sure that shrinkage was not obscuring any real effects. The development year and accident year parameters came out fairly smooth anyway (Fig. 8).

The main effect seen in the calendar-year trend is a substantial downward jump in 1993. There are two inflation drivers in workers comp. Wage replacement is driven by wage inflation, but is mostly fixed at the wages at time of injury, so shows up in the accident-year, i.e., row, parameters. Medical payments are made at the cost levels at time of payment, on the other hand, so are calendar-year effects (Fig. 9).

Many state laws specify that payments are to be made at the medical providers' standard rates. At some point providers and medical insurers agreed that the providers would increase their rates substantially but those insurers would get a

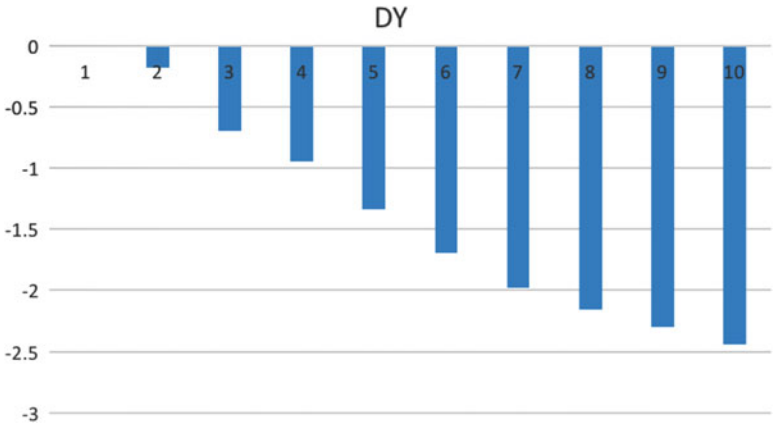


Fig. 8 Log payout level by lag

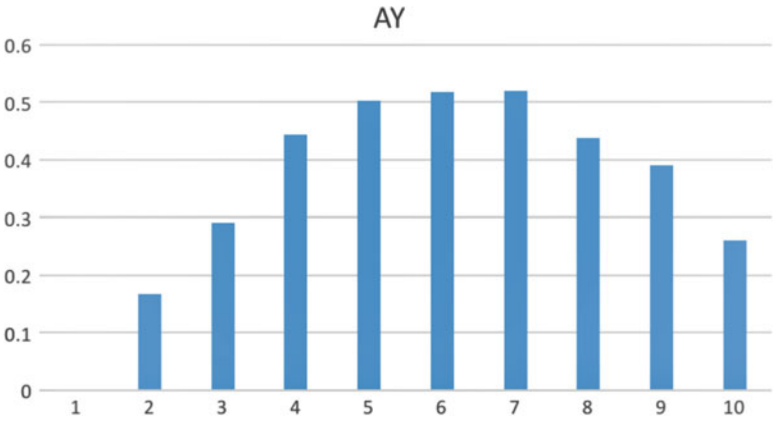


Fig. 9 Accident year levels

discounted rate. That left comp insurers as the only ones paying those artificial standard rates. At some point states started to realize this and basically get the comp insurers inside the game—perhaps through medical fee schedules for comp or other approaches. The comp insurers did not have the political clout to accomplish this, but they pass costs on to employers, who often do. Still, however, some states have higher medical payments for workers comp compared to other insurers (Fig. 10).

The downward jump in costs on the 1993 diagonal could well have come from this kind of reform. By 1997 it appears to be eroding a bit, however.

In any case, this model does not resolve the payout pattern issue. Lag 0 and lag 1 residuals show an inverse relationship starting with row 5 (Fig. 11).

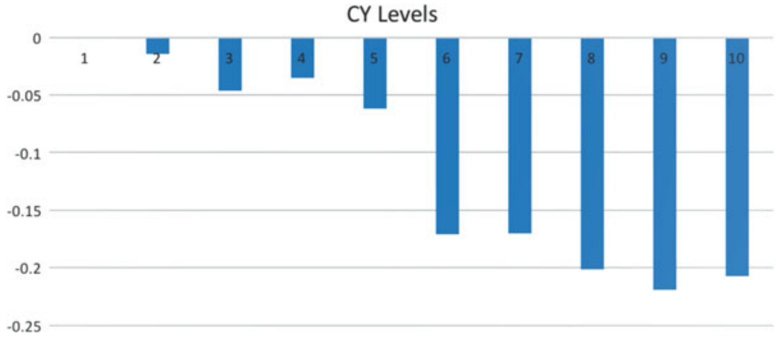


Fig. 10 Calendar year levels

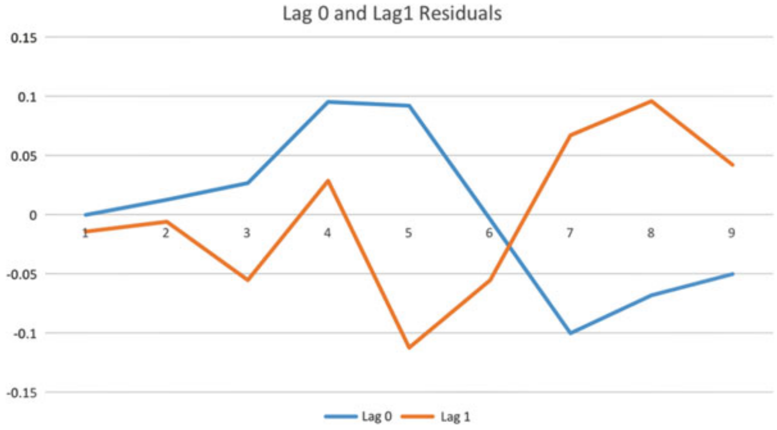


Fig. 11 Lag 0 and Lag 1 residuals by accident year

6.3 Model Extensions

Probably the most typical actuarial response to changes in payout pattern is to just use the data after the change. Meyers (2015) introduces modeling of changing payout patterns. With $y[n, u]$ = log of incremental claims for year n and lag u, one of his models can be written as:

$$y[n, u] = p[n] + q[u]z^{n-1} + \varepsilon_{n,u} \tag{5}$$

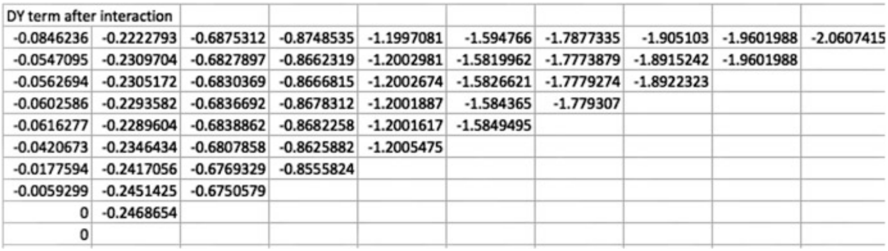
If $z = 1$, the payout pattern is constant, but if it is a bit above or below 1, the payout is speeding up or slowing down. This model does not include changes in trend, however, nor parameter reduction. One possible way to incorporate all of these effects is to add an interaction term between lag and accident year:

$$y[n, u] = p[n] + q[u] + w[n]x[u] + r[n + u] + \varepsilon_{n,u} \tag{6}$$

The slope changes for $u[n]$ and $x[u]$ in the interaction term were modeled as starting at the bottom and right, and built up going across the triangle right to left. The linear combination $q[u] + w[n]x[u]$ for the changing payout pattern is shown by cell.

The zeros at the bottom left are for identifiability and are the largest numbers in the triangle. A payout shift is seen from lag one, mostly to lag zero, but slightly to lag two as well. With the payout change modeled, the calendar-year levels below seem to be moving more uniformly. However, there is still a bigger change showing up in 1993 (Fig. 12).

At this point this model with interaction is still exploratory, but it does suggest such interactions may have a place in reserve triangle modeling (Fig. 13).



DY term after interaction									
-0.0846236	-0.2222793	-0.6875312	-0.8748535	-1.1997081	-1.594766	-1.7877335	-1.905103	-1.9601988	-2.0607415
-0.0547095	-0.2309704	-0.6827897	-0.8662319	-1.2002981	-1.5819962	-1.7773879	-1.8915242	-1.9601988	
-0.0562694	-0.2305172	-0.6830369	-0.8666815	-1.2002674	-1.5826621	-1.7779274	-1.8922323		
-0.0602586	-0.2293582	-0.6836692	-0.8678312	-1.2001887	-1.584365	-1.779307			
-0.0616277	-0.2289604	-0.6838862	-0.8682258	-1.2001617	-1.5849495				
-0.0420673	-0.2346434	-0.6807858	-0.8625882	-1.2005475					
-0.0177594	-0.2417056	-0.6769329	-0.8555824						
-0.0059299	-0.2451425	-0.6750579							
0	-0.2468654								
0									

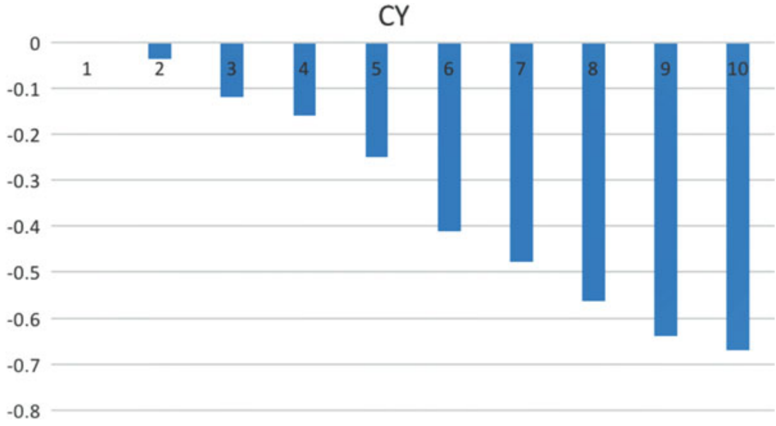


Fig. 12 Calendar year levels in model with changing payout patterns

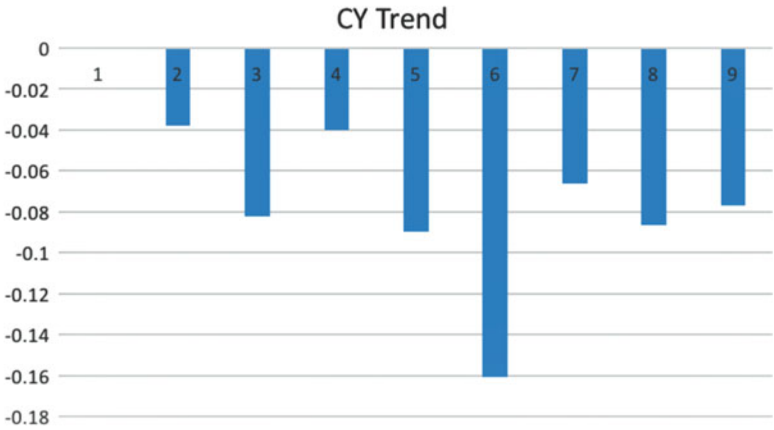


Fig. 13 Calendar year trends (slopes) in model with changing payout patterns

7 Conclusion

I like the maxim: All statements that begin with “All” are aimed at dramatic effect.

Still the idea that models may be only approximations but nonetheless can be useful is a key element of the shift towards pragmatism taking place in statistics. I am calling this the Robust Paradigm because of the notion that models need to be robust to effects that do not show up in the data at hand. This is broader than what usually is called robust statistics.

Assuming that the data is generated by the model process produces statistical tests that are mainly suggestive in this context. Out-of-sample testing is the requirement now. The availability of fast loo allows this to be standardized to a degree. Overfitting and so penalizing for too many parameters is no longer an issue when model performance out of sample is the focus.

But this is not traditional Bayesian either. Prior and posterior distributions are not statements of opinion. They are pieces of the story the model is telling us, and are as real as any other mathematical objects, such as quantum fields in the standard model of physics. And they are first and foremost pragmatic—helping to build a coherent narrative that provides insight into a process.

Parameter reduction now has classical and Bayesian modes. In the end the Bayesian approaches look more flexible and so more useful, particularly because of efficient loo.

The actuarial model with time variables is over-parameterized and so is a natural place for parameter reduction. This appears promising both for mortality and loss development applications. The more complex versions with interactions seem applicable to reserves, especially with payout pattern changes. Fairly extensive constraints are needed to get the parameters to do what they are meant to, however. There are a lot of possible overlaps and tradeoffs among parameters that need to be recognized explicitly if the models are going to perform as intended.

References

Lindstrom, M.J., Bates, D.M.: Nonlinear mixed effects models for repeated measures data. Biometrics **46**(3) (1990). ftp://www.biostat.wisc.edu/pub/lindstrom/papers/biometrics.1990.pdf

Meyers, G.: Stochastic loss reserving using Bayesian MCMC models. CAS Monograph Series, No. 1 (2015). http://www.casact.org/pubs/monographs/papers/01--Meyers.PDF

Osbourne, M.R., Presnell, B., Turlach, B.A.: On the lasso and its dual. J. Comput. Graph. Stat. **9**(2), 319–337 (2000). http://www.stat.washington.edu/courses/stat527/s13/readings/osborneetal00.pdf

Pereira, J.M., Basto, M., Ferreira da Silva, A.: The logistic lasso and ridge regression in predicting corporate failure. Proc. Econ. Financ. **39**, 634–641 (2016). http://www.sciencedirect.com/science/article/pii/S2212567116303100

Taylor, G., McGuire, G.: Stochastic loss reserving using generalized linear models. CAS Monograph Series, No. 3 (2016). http://www.casact.org/pubs/monographs/papers/03--Taylor.pdf

Ye, J.: On measuring and correcting the effects of data mining and model selection. J. Am. Stat. Assoc. **93**, 120–131 (1998)