Triangles in Life and Casualty

Gary G. Venter

gary.venter@gmail.com

Abstract The loss development triangles in casualty insurance have a similar setup to the mortality datasets used to identify and project mortality trends for life insurance. Some of the techniques used in casualty insurance triangle modeling are reviewed to see if they may be applicable to mortality modeling. Some turn out not to be while others are promising. The mortality models may have some applicability in casualty loss development as well.

Triangles in Life and Casualty

The loss development triangles in casualty insurance are similar to the mortality triangles used to study trends in mortality. Both arrays are triangles, or triangles with some part missing, with the bottom of the triangle to be filled in by modeling. We will look at applying some of the techniques used in casualty insurance, particularly parameter reduction, to see if it helps with the life insurance problem.

An example is worked through on French mortality data from the Human Mortality Database. As the interest is in longevity, ages at death of 50 and on were used. This was stopped at age 99, as the data got a bit dicey after that. Calendar years used started at 1947, as since then there has been stability in national boundaries, with the data ending with 2004. This resulted in using year-of-birth cohorts 1848 to 1954.

In casualty insurance the rows of the triangles are typically year of accident occurrence, and upward-right sloping diagonals are all in the same year of payment. The columns are called payment lags, but as any loss occurring in say 2007 and paid in 2008 is identified as lag 1, lag 1 can be anywhere from 0 to 2 years of time elapsed between occurrence and payment. In the life case the rows are typically age at latest birthday before death, and the columns year of death, so the similar diagonal is year of birth. This can be a two-year range, however, similar to lag in the casualty case.

Some of these choices are arbitrary, so to make the life triangle a little more like the casualty triangle, we take the rows to be year at birth and the columns age at death. Then the diagonals become calendar year of death. The diagonals could be defined precisely as in the casualty case, but we choose to define the rows and columns precisely so the implied diagonals are two-year ranges. We will be sloppy about how we refer to this, usually talking like it is a single year.

One reasonably good model of mortality is Lee-Carter plus cohorts, from Renshaw-Habermann (2006) (RH). Using casualty-type notation for this, let d represent the age in a column and w the birth-year for a row. Then w+d is the year of death, so is constant on the diagonal. The mortality rate for a cell is $m_{w,d} = D_{w,d}/E_{w,d}$, where D and E are the deaths in the cell and the number alive from year of birth w at age d. This could be at the beginning of the year or the average for the year. Suppressing the error term, the RH model in this notation can be expressed as:

$$\log m_{w,d} = a_d + b_d h_{w+d} + c_d u_w$$

Since a_d is usually close to the average of the log $m_{w,d}$, defining $x_{w,d} = \log m_{w,d} - a_d$ gives the model $x_{w,d} = b_d h_{w+d} + c_d u_w$. In casualty insurance it would be more usual to use the model $x_{w,d} = b_d h_{w+d} u_w$, so this is tested as an alternative and referred to as the TV model, as Taylor (1977) proposed it for x = losses, following Verbeek (1972) who proposed it for x = claim counts. However we estimate a_d by MLE instead of using the average, so TV here is $\log m_{w,d} = a_d + b_d h_{w+d} u_w$. The RH model may be interesting for casualty insurance, taking x as losses in $x_{w,d} = b_d h_{w+d} + c_d u_w$.

To do MLE, E is considered deterministic and D stochastic. A typical assumption is that $D_{w,d}$ is Poisson in $m_{w,d}E_{w,d}$. The loglikelihood is then:

$$L = \sum_{w,d}\{D_{w,d}\log[m_{w,d}E_{w,d}] - m_{w,d}E_{w,d} - \log[D_{w,d}!]\}$$

The RH and TV formulas for $m_{w,d}$ can be substituted here. To do the MLE, start with setting the partial derivatives of L with respect to (wrt) each parameter to zero.

For a particular j, differentiating the RH formula wrt a_j gives:

$$\exp(a_j) = \sum_w D_{w,j} \,/\, \sum_w E_{w,j}\exp(b_j h_{w+j} + c_j u_w)$$

Denote the modeled point as $\mu_{w,d} = E_{w,d}\exp(a_d + b_d h_{w+d} + c_d u_w)$ and the residuals as $R_{w,d} = D_{w,d} - \mu_{w,d}$. The first and second derivatives of L wrt b_j are:

$$\sum_w R_{w,j} h_{w+j} \,; \qquad -\sum_w h_{w+j}^2 \mu_{w,j}$$

Wrt c_j are:

$$\sum_w R_{w,j} u_w \,; \qquad -\sum_w u_w^2 \mu_{w,j}$$

Wrt u_i are:

$$\sum_d R_{i,d} c_d \,; \qquad -\sum_d c_d^2 \mu_{i,d}$$

And wrt h_k are:

$$\sum_{w+d=k} R_{w,d} b_d \,; \qquad -\sum_{w+d=k} b_d^2 \mu_{w,d}$$

For the b, c, u and h parameters it is not possible to solve explicitly for the parameter in terms of the others and the data. If it were, you could iterate for the parameters in the fashion of Bailey and Simon (1960), i.e., solve for one set in terms of the others, going through each set in turn, and repeating until convergence. However Goodman (1979) provided a work-around in this case: do a

Newton-Raphson iteration at each step instead of solving for the parameters exactly. Since you want the derivatives to be zero, the iteration for parameter θ is $\theta_{i+1} = \theta_i - L'/L''$, where the derivatives are wrt θ.

Differentiating the TV formula wrt a_j gives:

$\exp(a_j) = \sum_w D_{w,j} / \sum_w E_{w,j} \exp(b_j h_{w+j} u_w)$ or $a_j = a_j + \log(\sum_w D_{w,j} / \sum_w \mu_{w,j})$ for iteration.

The first and second derivatives wrt b_j are:

$\sum_w R_{w,j} h_{w+j} u_w$; $\qquad -\sum_w h_{w+j}^2 u_w^2 \mu_{w,j}$

Wrt u_i are:

$\sum_d R_{i,d} b_d h_{i+d}$; $\qquad -\sum_d b_d^2 h_{i+d}^2 \mu_{i,d}$

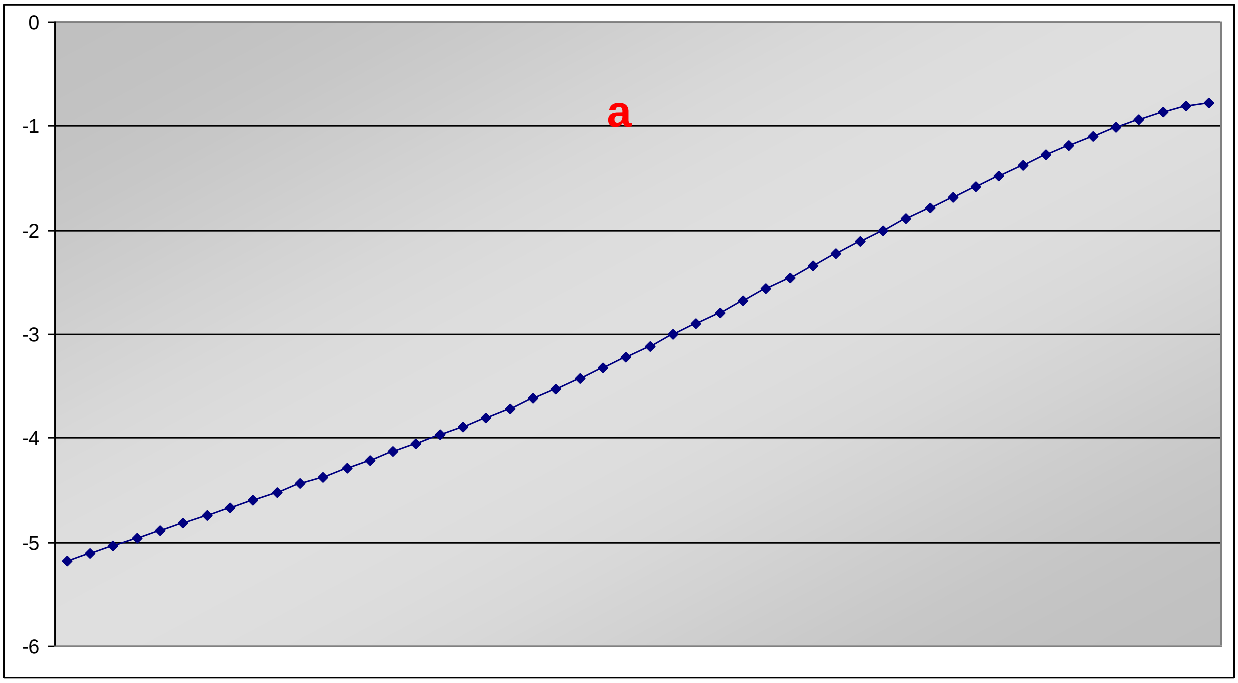
And wrt h_k are:

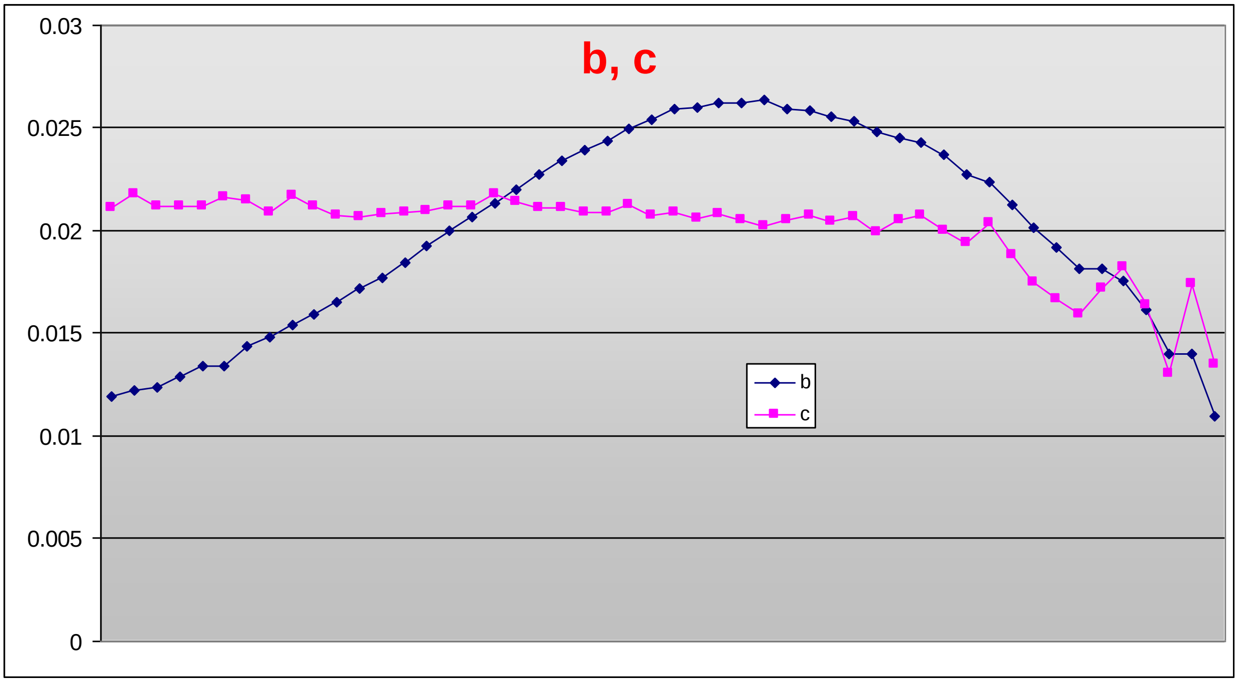
$\sum_{w+d=k} R_{w,d} b_d u_w$; $\qquad -\sum_{w+d=k} b_d^2 u_w^2 \mu_{w,d}$.

For RH, Cairns etal. (2007) suggest the following parameter constraints for identifiability reasons: $\sum_d b_d = \sum_d c_d = 1$; $\sum_w u_w = \sum_k h_k = 0$. For starting parameters in the iteration, a_d is taken as the average over w of log $m_{w,d}$, each b_d and c_d is 1/n, and u_w and h_k start at zero. The latest estimate of the μs is used at each step. Since they are initially zero, the h and u parameters are iterated first, then b and c, then a, which has the best starting values. The constraints are applied to adjust the iterated parameters at each iteration using the formulas: $b_d^* = b_d/\Sigma b_d$; $c_d^* = c_d/\Sigma c_d$; $u_w^* = [u_w - \text{average}(u_w)]\Sigma c_d$; $h_k^* = [h_k - \text{average}(h_k)]\Sigma b_d$; $a_d^* = a_d + c_d \text{average}(u_w) + b_d \text{average}(h_k)$. Then $a_d^* + b_d^* h_{w+d}^* + c_d^* u_w^* = a_d + b_d h_{w+d} + c_d u_w$ and the constraints are satisfied.

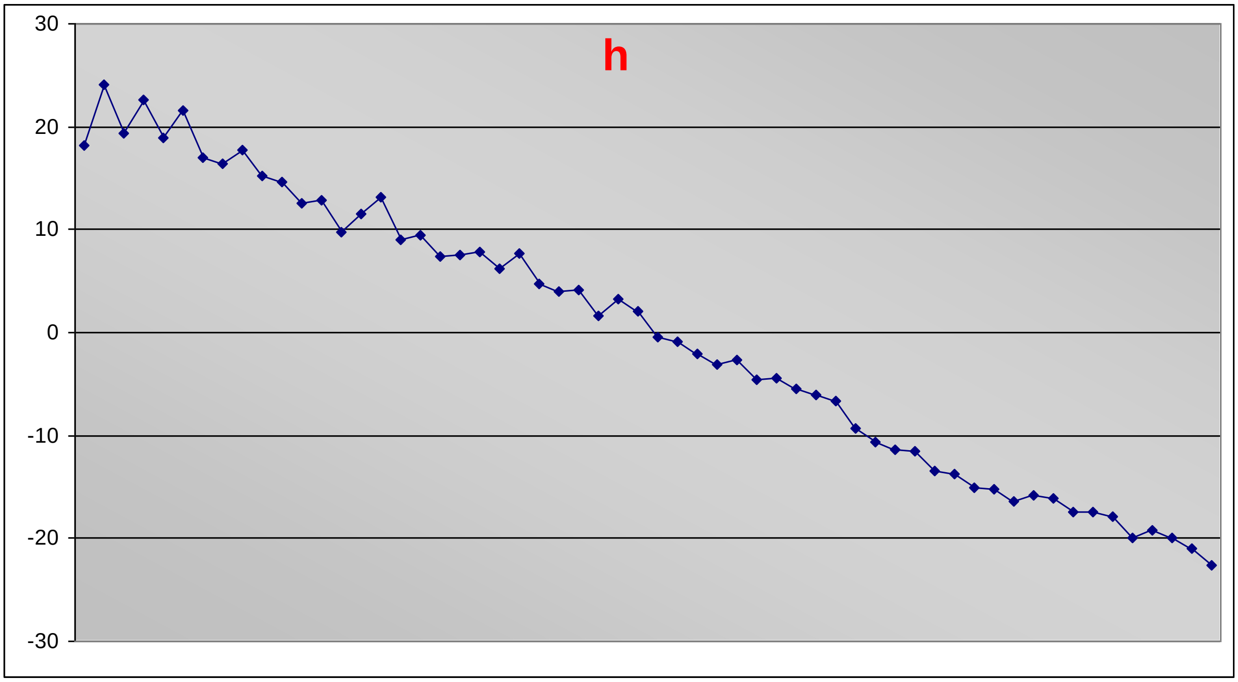
The estimated parameters for the RH model under these constraints are graphed below. The loglikelihood is −19005.8. The trend in the a's just shows the degree of increasing mortality by age after adjusting for the other bits. The u parameters are the age-of-birth, or cohort, effects, which generally increase until about 1900 and have been declining gradually since, but with some sharp exceptions, particularly in the 1915 – 1920 range, which was perhaps influenced by wars and epidemics. The c parameters show how the cohort effect varies by age. As these are gradually declining, then fall off more sharply at higher ages, the cohort effect decreases as the cohort attains older ages. The trend parameters h are declining over time, showing continuing mortality improvement by this measure. The effect of the mortality trend on different ages is shown by the b parameters, which rise until age 79, then decline.

RH model a, b, c parameters for ages 50 to 99

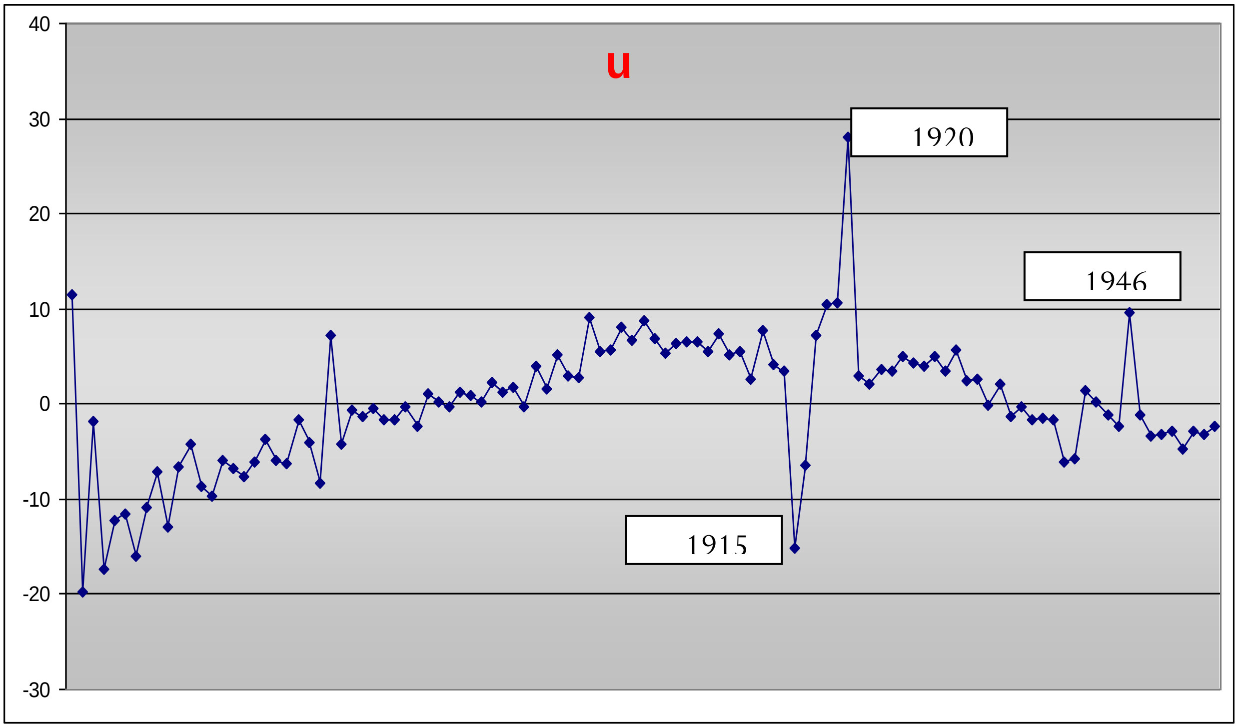




RH model h parameters for years of death 1947 to 2004



RH model u parameters for years of birth 1848 to 1954



For the TV model we take as constraints that the h and u parameters average 1, which will be their starting values. In the iterations, their iterated values are divided by their averages, and each b parameter is multiplied by the product of those averages. The starting values for b are taken as 0, and a starts at the average of log m as before. The first iterated is b, to move it off zero, then h and u and finally a. The result is a loglikelihood of −38,130.9. The reduction of 49 parameters is not worth this decrease in loglikelihood by any information criterion. The RH model seems to be a much better formulation for this data. Having calendar-year and birth-year effects that are adjusted by age seems to make sense and works well. Perhaps it is worth considering this type of formulation in casualty triangles. The way calendar-year and accident-year effects interact and vary by lag in casualty loss development is not always well handled by the TV model, so the RH may help.

Parameter Reduction

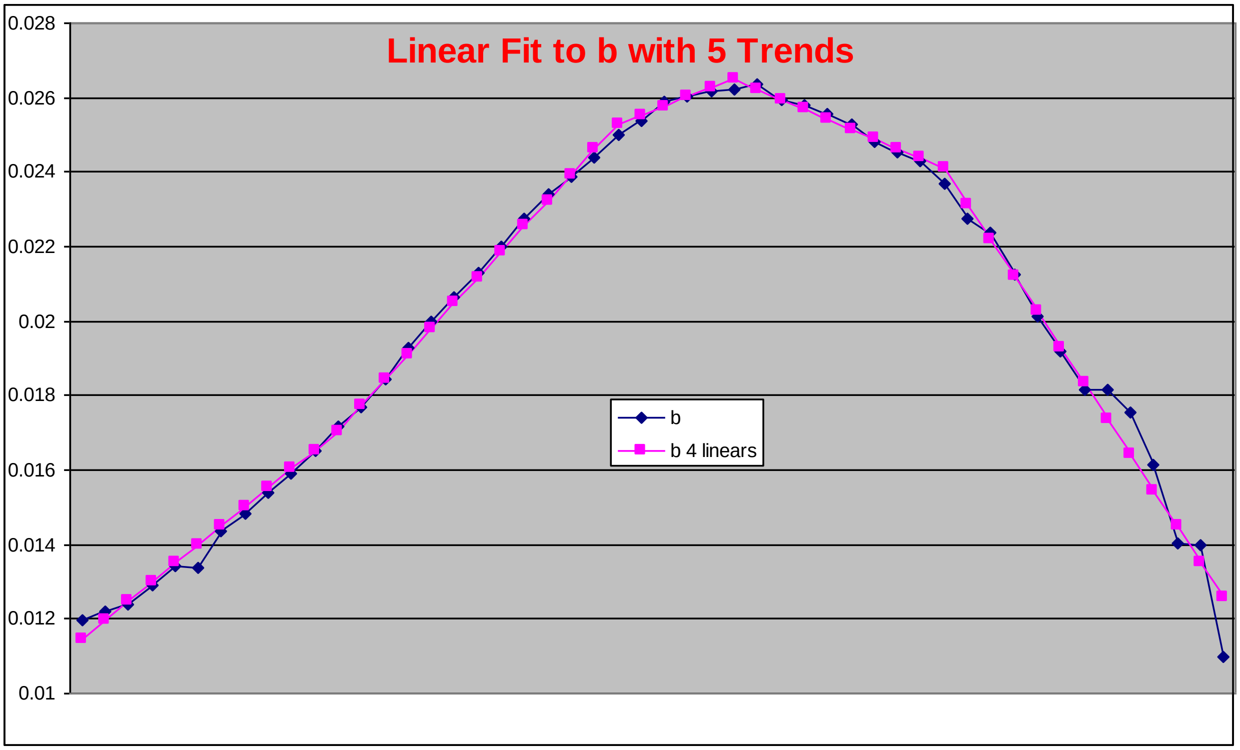
In the RH model for this data there are 150 age parameters a, b, c, 107 cohort parameters u, and 58 calendar-year parameters h, with 4 constraints, giving a total of 311 parameter equivalents. A better way of counting parameters is to use Ye's (1998) generalized degrees of freedom, which add up the derivatives of the fitted points wrt the corresponding data points. This could be somewhat computationally extensive, however, since an iteration would be needed to solve for the derivative at each of the 2900 data points, and so is not pursued for now.

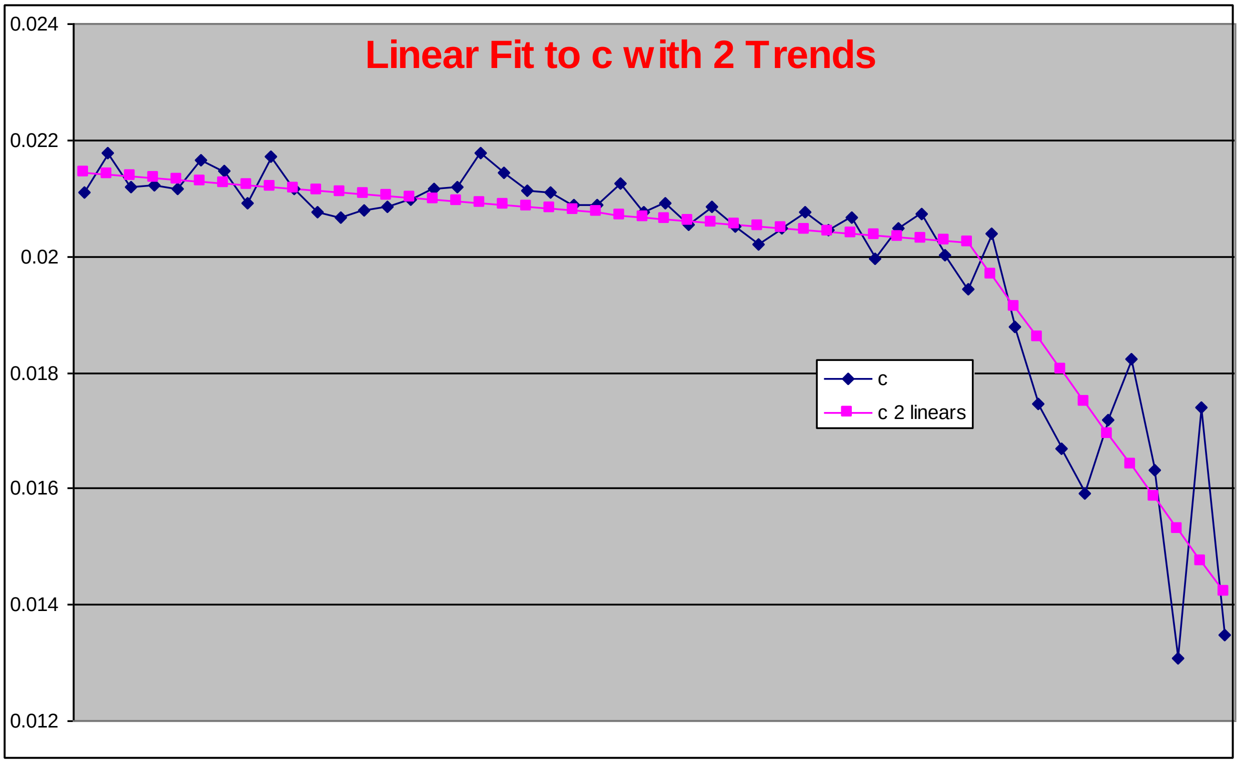
In casualty development, relationships among parameters, like trends, are used to reduce the number of parameters used. For instance, Venter (2007) shows that for the classical Taylor-Ashe triangle, Mack chain ladder gives narrower prediction intervals than does the ODP model with the same estimates. The ODP model fits the data better but has more parameters and more parameter uncertainty. However this better fit is retained with a judicial reduction in the number of ODP parameters, which gives an ODP model with narrower ranges than the Mack model's.

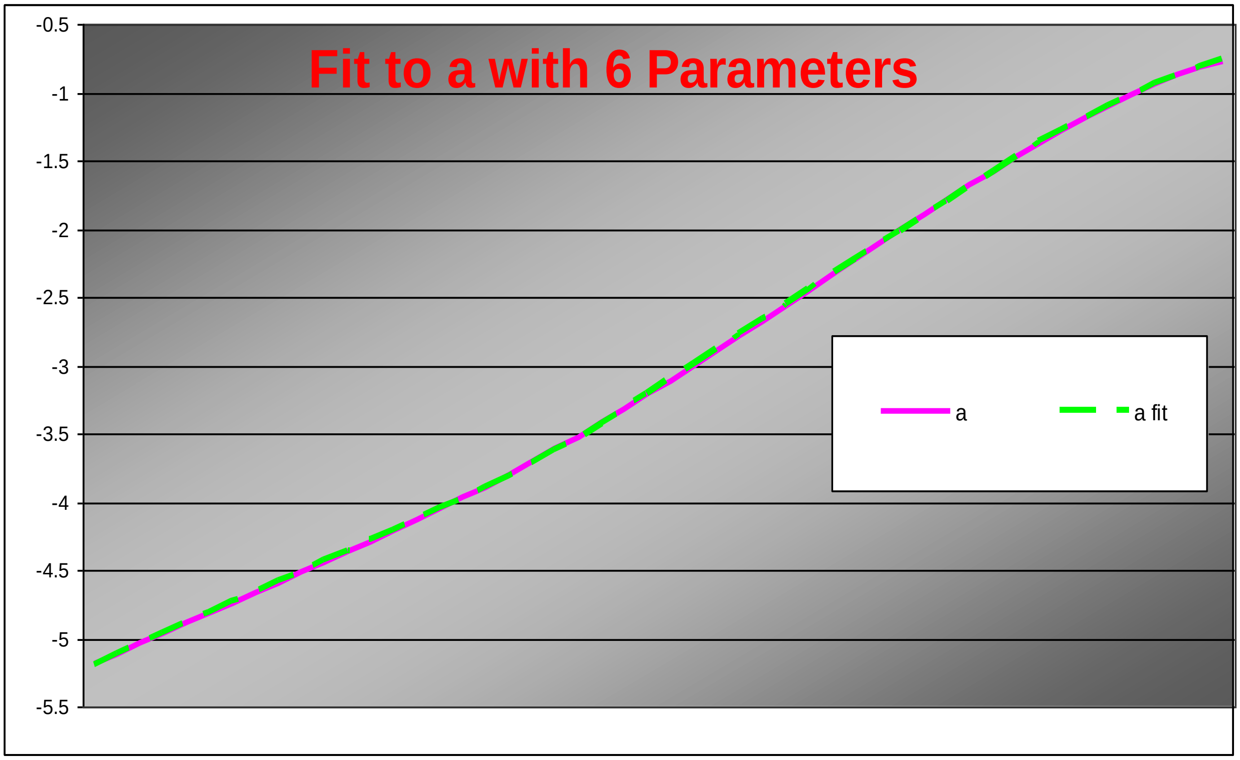
Barnett and Zehnwirth (2000) provide a systematic way to do parameter reduction through linear trends. For instance, they might express a_d as $\Sigma_{i=1:d}y_i$. Then making several sequential y parameters equal replaces the corresponding a parameters with a linear trend over that period, thereby reducing the number of parameters. Basically this is fitting first differences of the parameters.

An alternative is smoothing with splines (e.g., see Currie etal. 2004), which may have a similar effect. Here we use information criteria to look at the tradeoff of goodness-of-fit and number of parameters. In the classic AIC, the loglikelihood has to improve by 1 to make an additional parameter worthwhile. For a sample size of N, the more stringent SB criterion requires an improvement of ½ log(N) = log($N^{½}$). There is increasing sentiment that this is too harsh on additional parameters. A compromise is the Hannan-Quinn information criterion, HQIC, which requires an improvement of log(log(N)). Here with a sample size of 2900, the per-parameter penalties to the loglikelihood are 1, 2.08, and 3.99 for the AIC, HQIC and SB, respectively, which basically doubles the penalty at each step.

The Barnett-Zehnwirth method was tried on each of the five sets of parameters a, b, c, u, h. The b's were fit fairly well with 5 trends, or 6 parameters, and the c's by 2 trends, or 3 parameters (y_1 is a parameter, then each trend requires one more parameter). Using such trends for b and c constrained to average to 1 reduces the loglikelihood to −19117.9, or by 112.1. After constraints this is a reduction of 91 parameters. Thus the parameter reduction is easily justified according to the HQIC and SB criteria, but however not by the AIC. This result is suggestive that parameter reduction can be fruitful. However similar attempts at this aggressive degree of parameter reduction were not justified for the a, h, or u parameters. Milder parameter reduction may be helpful there.

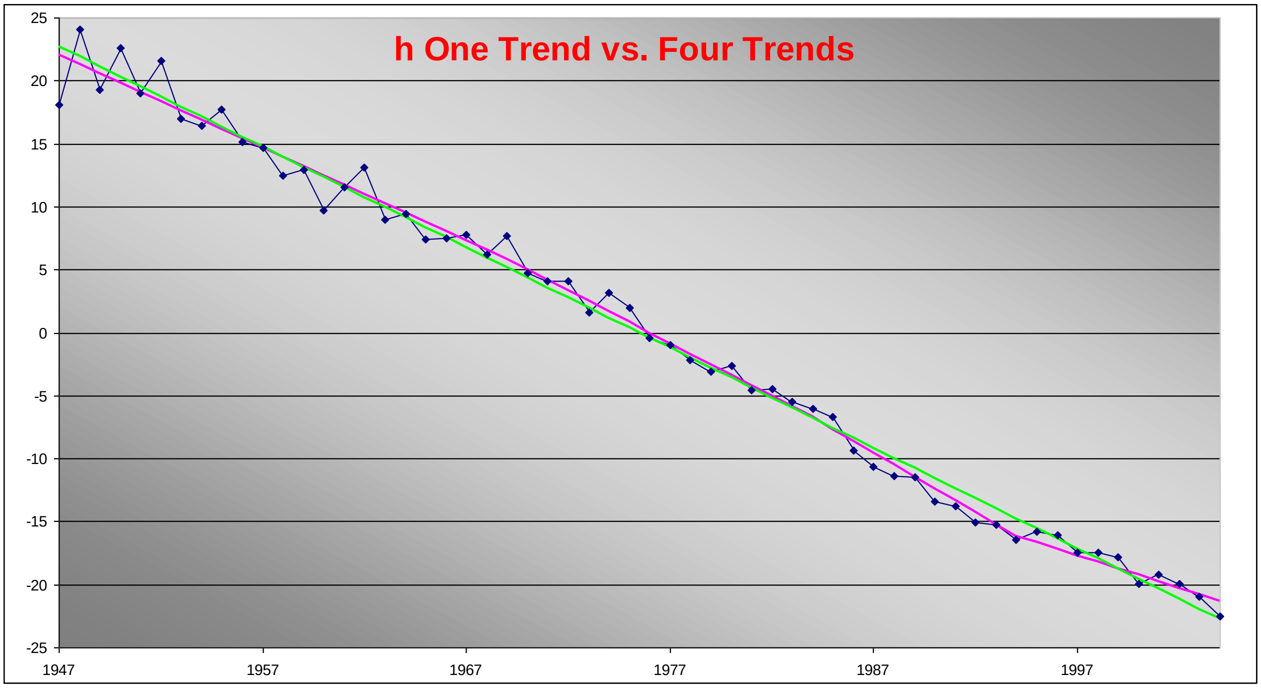






A fit to the a's with 4 linear trends and a power curve tail is hard to tell from the a's by eyeball, but the likelihood function was not impressed.

Looking at the linear trends gives other insights. The calendar-year trend h was fit reasonably well by four linear trends, as below:



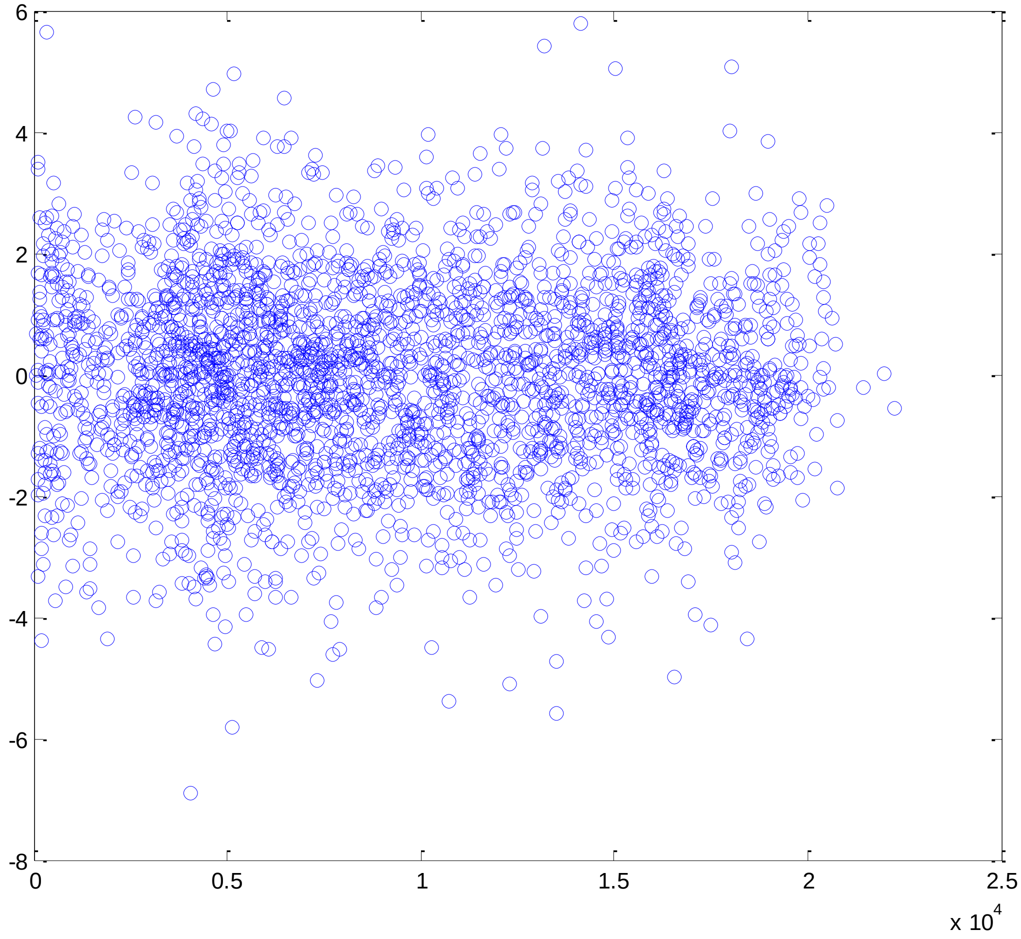
1947	to 1969	to 1984	to 1994	to 2004
22.094	-0.738	-0.838	-0.943	-0.517
22.739	-0.798	-0.798	-0.798	-0.798

A slowing of the trend in the last 10 years is suggested, and may have implications for the forecasted trend. Trend is addressed in Appendix 1. The apparent reduction in trend does not hold up, however, as the first differences in h are fairly volatile. Moreover this is complicated by a possible downward trend in the cohort parameters u in the recent years, as seen earlier.

Poissonness

A natural question is whether or not the Poisson distribution is realistic. Assuming deaths in a cell are from independent Bernoulli trials implies a binomial distribution, which has a lower variance than the Poisson. However there could be contagion effects, etc. that could make the variance higher. One test frequently done (e.g., see Barnett-Zehnwith) on casualty development triangle residuals is to plot standardized residuals against fitted values. Under the Poisson assumption, the standardized residuals are the residuals divided by the square root of the fitted values. These are graphed below.

<u>Standardized residuals against fitted values</u>



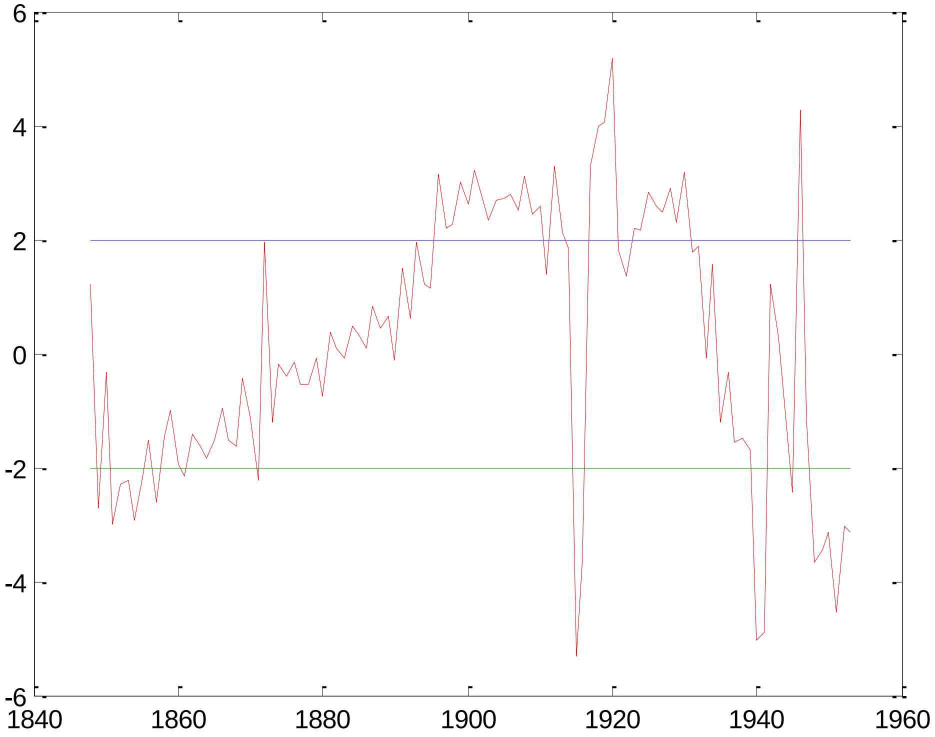
There is no clear spreading of the residuals moving to the right or the left, suggesting that the standard deviation of the residuals is in fact close to being proportional to the square-root of the means. The ranges of the residuals seem high for the Poisson, however. The negative binomial can be parameterized so that the variance is a multiple greater than one of the mean (although this version is not in the exponential family). It might be more realistic in this case.

Standard Errors of Parameters

For casualty triangles, one approach to getting standard errors of the estimated parameters is to estimate the Fischer information matrix using the inverse of the matrix of the negatives of the mixed 2nd partial derivatives of the loglikelihood function. E.g., see Clark (2003). Appendix 2 shows the formulas for the mixed partials for the RH model. The diagonal of the resulting covariance matrix is the variance vector of the parameters. Due to the constraints, the last b, c, h, and u parameter were fixed and were not used in the matrix, so it is a 311x311 matrix.

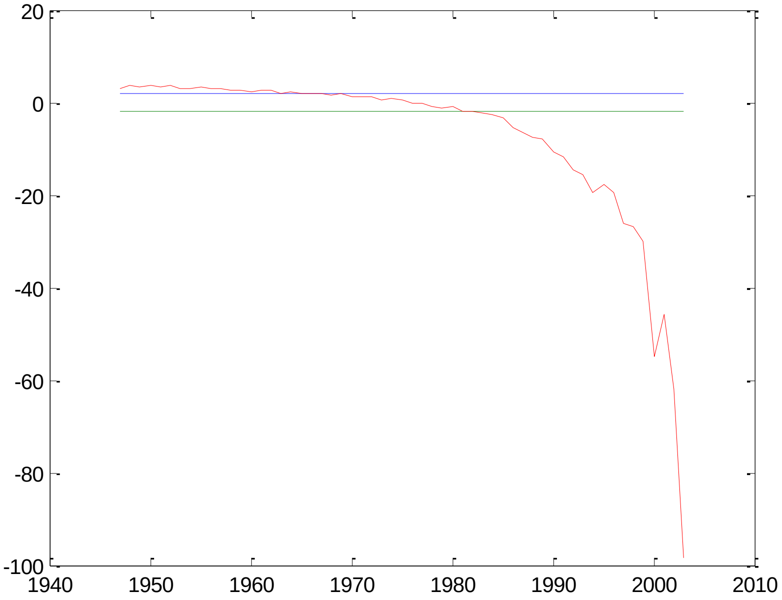
The a, b, and c parameters were all quite significant compared to their standard errors, with t statistics of 5 or more in absolute value. The h parameters (for calendar year of death) became more significant in the later years as the data became more stable, possibly due to larger populations. The u (cohort) parameters were not nearly as significant, however. Plotted below are the t statistics for these parameters.

t statistics of u parameters with bars at 2 and -2

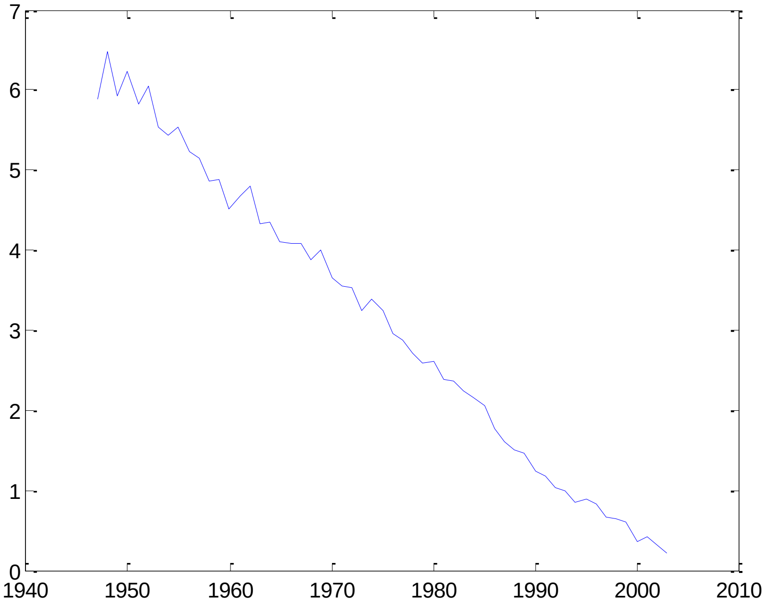


The significance is often low, which suggests the possibility that some parameter reduction may be useful for the cohorts. A similar graph for the h parameters shows that the more recent years have very significant parameters, but this is not so for the older years. The standard errors themselves decline steadily over time as well, which is in line with the greater stability of the more recent data, suggesting that some parameter reduction might be useful for the older years. The se's of the cohort parameters displayed a similar pattern (not shown).

t statistics of h parameters with bars at 2 and -2



Standard errors of h parameters



Summary

It turns out that life and casualty actuaries have both been filling in the bottoms of triangles. Who knew? A model from life insurance and one from casualty were fit to French life data. The life model fits much better and may be applicable in casualty loss development as well. Most but not all of the parameters are statistically significant. Fairly extreme parameter reduction methods were tried, and this was fruitful for two of the parameter types. Less aggressive parameter reduction may help with the other types. Whether the more parsimonious models produce narrower runoff ranges on actual business is not clear, however, as the uncertainty in the projected mortality trend is a large part of the overall uncertainty, and this might overwhelm the other sources. An AR1,1 model provided a reasonable fit to the mortality trend. The variance of residuals is close to being proportional to the mean, but the distribution is more heavy-tailed than Poisson. Negative binomial may be a good alternative.

References

[1] Bailey, Robert A. and Leroy J. Simon, 1960 "Two Studies In Automobile Insurance Rate Making," *ASTIN Bulletin* 1, 192-217.

[2] Barnett, Glen and Ben Zehnwirth, 2000 "Best Estimates for Reserves," PCAS LXXXVII, 245-321.

[3] Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and Balevich, I.(2007) "A quantitative comparison of stochastic mortality models using data from England & Wales and the United States," Working paper, Heriot-Watt University, and Pensions Institute Discussion Paper PI-0701.

[4] Clark, David R., 2003 "LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach," *CAS Forum*, Fall, 41-92.

[5] Currie, I.D., Durban, M. and Eilers, P.H.C., 2004 "Smoothing and Forecasting Mortality Rates," Statistical Modelling, 4: 279-298.

[6] Goodman, L.A., 1979 "Simple models for the analysis of association in cross-classifications having ordered categories," Journal of the American Statistical Association 74, 537–552.

[7] Hannan, E. and B. Quinn 1979 "The determination of the order of an autoregression," Journal of the Royal Statistical Society B(41), 190–195.

[8] Lee, R.D., and Carter, L.R. 1992 "Modeling and forecasting U.S. mortality," Journal of the American Statistical Association, 87: 659-675.

[9] Renshaw, A.E., and Haberman, S. 2006 "A cohort-based extension to the Lee-Carter model for mortality reduction factors," Insurance: Mathematics and Economics, 38: 556-570.

[10] Taylor, Greg C., 1977 "Separation of Inflation and Other Effects From the Distribution of Non-Life Insurance Claim Delays," *ASTIN Bulletin* 9, No. 1-2, 219-230.

[11] Venter, Gary G., 2007 "Refining Reserve Runoff Ranges," *CAS e-Forum*

[12] Verbeek, Harry G., 1972 "An approach to the analysis of claims experience in motor liability excess of loss reassurance," *ASTIN Bulletin* 6, No. 3, 195-202.

[13] Ye, Jianming. 1998 "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, March 1998, 93(441): 120-131.

Appendix 1 – Mortality Trend

It is fairly typical for life actuaries to fit ARIMA or Kalman filter models to the mortality trends h. Casualty actuaries use similar models for the inflationary trend in the calendar-year direction and even for the accident-year trend, which corresponds to cohorts in the mortality data. Cohorts are sometimes trended similarly by life actuaries.

The Kalman filter can pick up the tendency for the trend to change, so probably is a good starting point for fitting the mortality trend. If z_k is used to describe the first differences in the calendar year parameters, i.e., $z_k = h_k - h_{k-1}$, a simple formulation of the Kalman filter can be applied as follows.

The true first difference process is denoted X_k and its estimate is x_k. $P_{k|k}$ is the estimated expected squared estimation error after the k^{th} observation. Start with $P_{0|0}$ large and x_0 fairly arbitrary. These wash out early. X_k is assumed to change each period by a random amount distributed $N(0, Q_k)$ – i.e., with variance Q_k. The observation z_k is X_k plus a random number distributed $N(0, R_k)$. The residual is $y_k = z_k - x_{k-1}$. The new estimate is $x_k = x_{k-1} + K_k y_k$, where K_k is calculated by:

$P_{k|k-1} = P_{k-1|k-1} + Q_k$

$S_k = P_{k|k-1} + R_k$. S_k is then the variance of y_k.

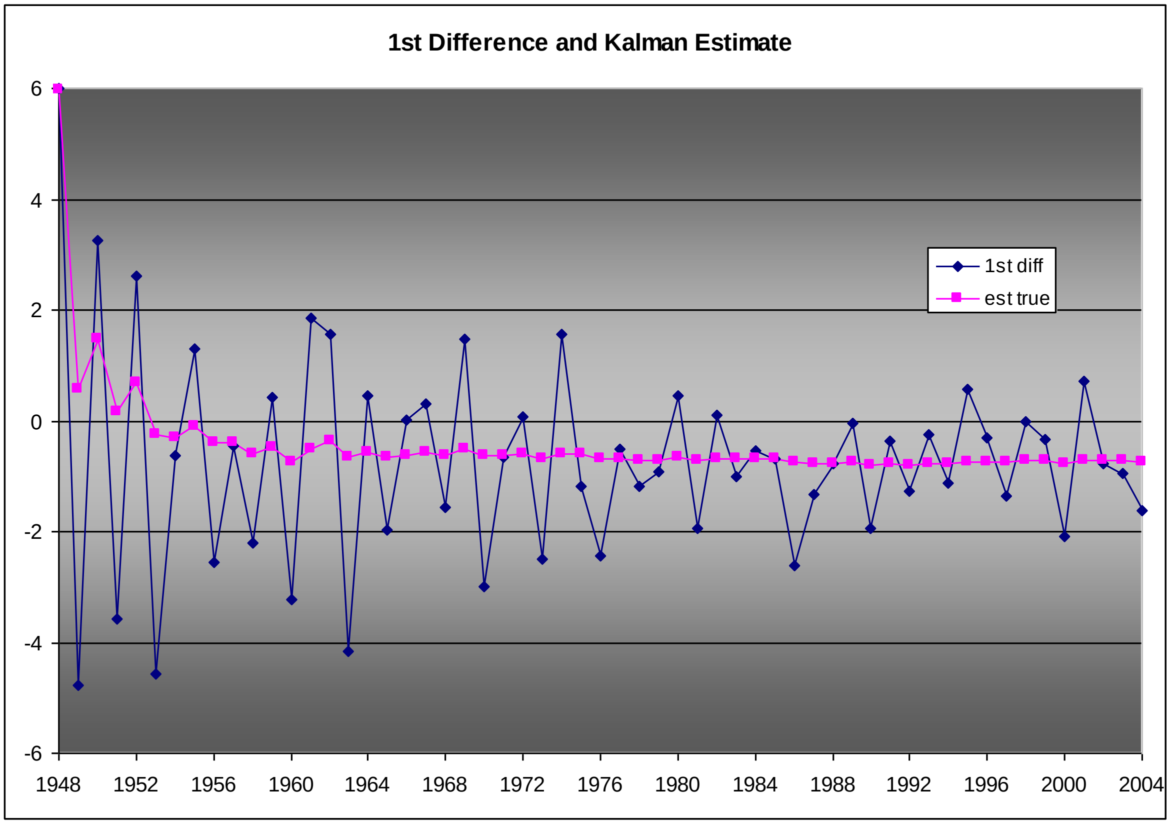
$K_k = P_{k|k-1} / S_k$.

The updated $P_{k|k}$ is $(1 - K_k)P_{k|k-1}$.

To apply this to the h parameter first differences, we assume that for some constant R, $R_k = (R/\text{exposed population at } k)^2$. The problem then is to estimate R and Q_k, which is assumed to be a constant Q.

The only relevant data available is the set of observed residuals y_k. These should be as small as possible and have expected squared errors close to S_k. A somewhat ad hoc way to get a reasonable estimate of R and Q is thus to minimize $\Sigma y_k^2 + \Sigma(S_k - y_k^2)^2$ over some period excluding the early start-off years. In practice we looked at the periods 1964 – 2004 and 1975 – 2004 and compromised between them.

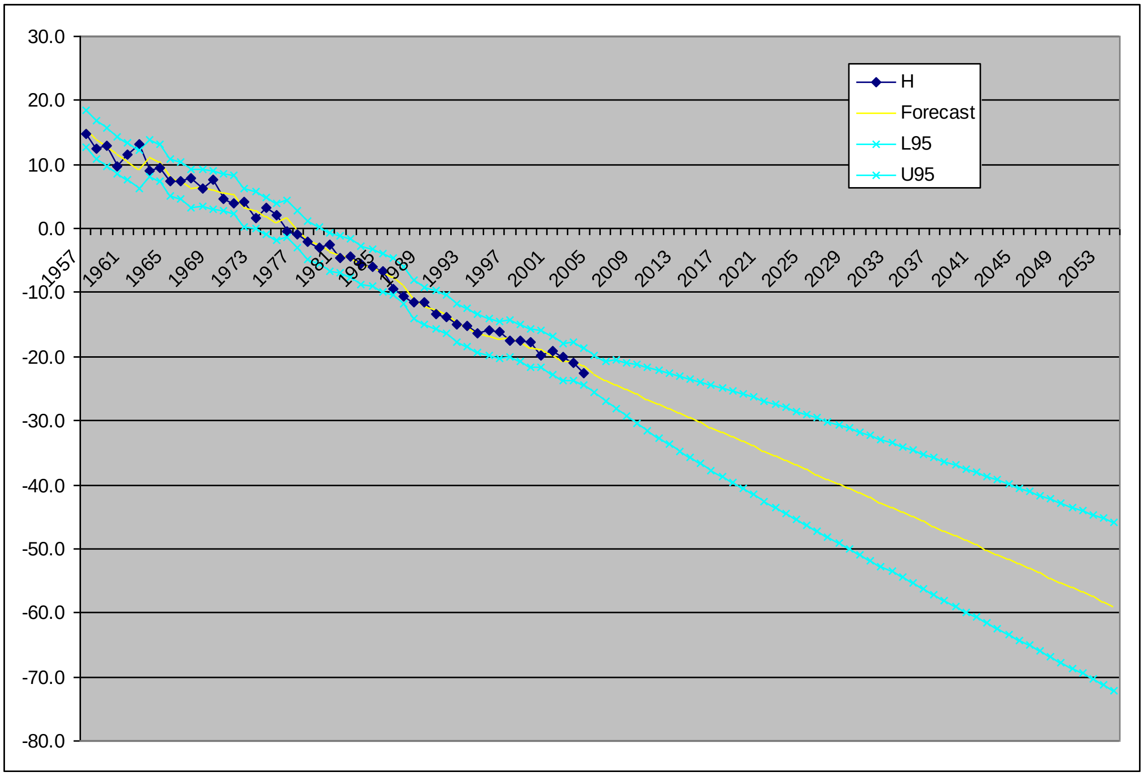
$P_{0|0}$ and x_0 were taken as 100 and 5, respectively. The population at ages 50 to 99 ranged from about 11million to 20million over the period. R was estimated at 15million. The quantity minimized was optimized with Q very close to zero. This says that there is no extra variance due to changes in X needed to explain the observed residual variability. It can all be attributed to the sampling variance R_k plus the estimation error at k – 1. The values of K_k were around 3% near the end. These parameters imply that the true X does not change, but its estimate changes a bit with the data. The graph below shows the z's and the resulting x's.



With no change in trend, a time-series model can pick up any autoregressive structure. An AR1 model[1] was significant for the z's, with $z_k = -0.735 - 0.613(z_{k-1}+0.735)+N(0, 2.25)$. This is consistent with mortality trends in other studies. It produces a linear trend with parabolic errors, projected for fifty years below. However the residuals were not homoscedastic, and adjusting the variance for the square of population growth was not enough to make them so. For some reason the mortality data in France is growing more stable than can be explained by population growth alone. Since the variance trend is downward, the forecast confidence intervals are probably conservative.

---

[1] Fit by John Major



Appendix 2 – Mixed 2nd Partials of RH

The derivative of the RH loglikelihood function wrt a_j is $\sum_w R_{w,j}$. The derivatives of that wrt any a_i, b_i or c_i with $i \neq j$ is zero. Its derivatives wrt a_j, b_j and c_j are $-\sum_w \mu_{w,j}$, $-\sum_w h_{w+j}\mu_{w,j}$, and $-\sum_w u_w \mu_{w,j}$, respectively, so these are the corresponding 2nd partials for a_j. For u_i the partial is $-c_j\mu_{i,j}$ and for h_k it is $-b_j\mu_{k-j,j}$, so these are the corresponding 2nd partials with a_j.

For b_j, the derivative of the loglikelihood is $\sum_w R_{w,j}h_{w+j}$. Thus the mixed 2nd partial wrt b_j and c_j is $-\sum_w h_{w+j}u_w\mu_{w,j}$. For b_j and u_i the 2nd partial is $-h_{i+j}c_j\mu_{i,j}$. For b_j and h_k it is $R_{k-j,j} - b_jh_k\mu_{k-j,j}$.

For c_j, the derivative of the loglikelihood is $\sum_w R_{w,j}u_w$. For c_j and h_k the 2nd partial is $-b_ju_{k-j}\mu_{k-j,j}$. For c_j and u_i it is $R_{i,j} - c_ju_i\mu_{i,j}$.

For u_i, the derivative of the loglikelihood is $\sum_d R_{i,d}c_d$. For u_i and h_k the 2nd partial is $-b_{k-i}c_{k-i}\mu_{i,k-i}$ but this is only for k up to i+50.