# Mortality Trend Models

Gary G. Venter

**Abstract**

Every 50 years or so a study of workers compensation mortality patterns is done, generally finding that after medical stabilization – 10 or more years after injury – mortality for seriously injured workers is comparable to that of the overall population. It has been about 25 years since the latest study, so we might be half way to the next one. But in the meanwhile there are trends in population mortality, and these impact loss reserve risk.

Mortality data over time can be arranged in triangles, and models fit to such data are similar to those used in casualty loss development – particularly those that model trends in the three dimensions of calendar year of finalization, age at finalization, and origin year. We fit such models to U.S. population male and female mortality data for death (finalization) ages 55 to 89, with several distributions of residuals. The information matrix is used to estimate parameter standard deviations.

Although there is an extensive literature on fitting these models, most of the papers do not address parameter significance through t statistics, etc. and doing so finds problems with the standard models. One problem is overparameterization, and a conclusion here is that parameter reduction methods such as smoothing should be used. Other authors have tried this, but a sticky issue is finding parameter reduction methods that actually produce improvements in goodness of fit, as measured by AIC, etc. This is an open problem as far as we know and a direction for future research.

Typically the starting point for the distribution of model residuals is Poisson, but several authors have found that negative binomial fits better. Unfortunately, some of these have misinterpreted the derivation of the negative binomial as a gamma-mixed Poisson to conclude that the negative binomial arises because there are different subpopulations each with different Poisson distributions. But a sum of subpopulations each Poisson distributed is itself Poisson distributed. The mixture becomes interesting when you are drawing at random from a subpopulation whose parameter you do not know. Probably the negative binomial arises from other contagion effects, like weather, disease outbreaks, etc. Unfortunately, these also make residuals across cells not independent, and this effect has been found in other studies as well.

A few alternative ways of parameterizing negative binomial residuals are discussed, and these are also applied to the Poisson-Inverse Gaussian distribution and its generalization, the Sichel. For females the negative binomial fits best but the male data is a bit more skewed than the negative binomial. However the Poisson inverse-Gaussian appears to be too skewed for this data. The Sichel is more flexible, with one more parameter, and fits best.

Further insight into the shifts in mortality over time is provided by fitting Makeham-like curves to each year of death. One implication from this exercise is that male mortality trends at the older ages had a shift in 1988, possibly data related. Probably data older than that is not reliable, or at minimum comes from a different process. The overall conclusion is that more work is needed to come up with reasonable models for mortality trend, with parameter reduction a leading candidate.

For trending, ARIMA models have often been fit to the calendar-year parameters after first differencing for stability. But since the parameters are estimated with error, differencing induces an autocorrelation, so the ARIMA models could be mostly fitting this artifact. Alternatives are discussed.

**Keywords:** Mortality Risk; Lee-Carter Model; Cohort Effects; Parameter Risk; Model Risk

## MORTALITY TREND MODELS

The general categories of process, parameter and model risk are applicable to mortality projection. Model risk is particularly problematic, as it turns out that the better fitting models have aspects that make them questionable for projection purposes. Lee-Carter models with and without cohort

effects with a few distributions of residuals are fit to the population mortality data from the Human Mortality Database (HMD) and are compared based on penalized maximum likelihood.

The models were fit to years of death starting with 1971. Preliminary analysis found different trends for ages of death below 55, due perhaps to reproductive health issues and the impact of HIV during some of this period. Female mortality below age 55 improved dramatically in the 1970s and has changed little since, whereas for males there was a sharp increase in mortality in the 1990s that has since recovered. The oldest age used is 89, as older ages had quite unusual mortality patterns before 1990—mortality reducing with age, etc. These could be data issues. The data available for this study ends with year of death 2006. This resulted in using year-of-birth cohorts 1882 to 1951. The cohort is year of death minus latest attained age at death, so is close to year of birth.

The fits with cohort parameters turn out to be problematic in part because the oldest cohorts have only a few observations, which makes their parameters very responsive to just a few data points, and this in turn creates distortions in other parameters. Adding the data for all years of death 55 – 89 for cohorts 1882 and later, reduces this problem. Another problem with the fits is that in the case of female death rates, the correlations among parameter estimates is high, which reduces the significance of the parameters and leads to questionable values.

Section 1 discusses the models used; Section 2 looks at the fits; Section 3 tries to interpret the parameters; Section 4 address adding more years of death; Section 5 looks at Makeham-like fits; and Section 6 gets to projection risk.

## 1. MODELS

HMD data comes in the form of number of deaths and number of living, who are considered the exposures to death. These are in cells by year of death and age at death. Subtracting age from year gives the cohort, which is approximately the year of birth, but can be slightly different depending on the time of year that birth and death occurred. Data is also available by actual year of birth but in most models that is considered less important, and cohort is used instead.

Here arrays are taken to have rows for year of death and columns for age at death. The years are 1971 to 2006, and the ages 55 to 89, so the arrays are 36x35, with 1260 elements. The years are indexed by t and the ages by d. The cohort is t – d and is constant along the NW-SE diagonals of the arrays.

The starting point for recent models of mortality is the LC model from Lee and Carter (1992). It models the mortality ratio m, which is deaths divided by exposures, in log form the mean is:

$$\log m_{t,d} = a_d + b_d h_t. \qquad (1.1)$$

Here $a_d$ is the base mortality for age d, $h_t$ is the trend level at year t, which generally goes down over time as mortality decreases, and $b_d$ allows different ages to have trend rates that are factors times the overall trend. This is useful in the case of male mortality, for example, where mortality rates for ages 55 to 60 have improved at a greater rate than those for 85 to 89. However, this is where the LC model can run into differences from actual data, as some ages might trend faster or slower for a while but not always.

A popular extension of LC is LC plus cohorts, from Renshaw-Habermann (2006) (RH):

$$\log m_{t,d} = a_d + b_d h_t + c_d u_{t-d}. \tag{1.2}$$

The cohort term u allows for mortality to also vary by year of birth, independently of year of death and age. It is not always clear why it should, but allowing it to seems to substantially improve the goodness of fit of the models. The c factor allows the cohort effect to vary by age; e.g., it might wash out at older ages, or it might be stronger at older ages.

There are some identifiability problems with these models. For instance, increasing every b by a factor and reducing every h by the same factor does not change the fitted values. This is similar for c and u. Here the constraints used for this are to set $b_{1955} = c_{1955} = 1$ and $h_{1971} = u_{1917} = 0$. The cohort 1917 was chosen as it is the last cohort that includes all calendar years. It is also one of the highest mortality cohorts for both males and females. The result of these constraints is that in the LC model $a_d$ is the fitted mortality for t = 1971 and $h_t$ is the trend level for age 55. All the other parameters are relative to these. In the RH model, every u is the cohort effect at age 55, where the cohort values are relative to cohort 1917. Traditionally sums of parameters have been constrained as a way to address the identifiability problems, but the approach here eliminates a few parameters, which is necessary to make the information matrix non-singular.

Fitting is done by maximum likelihood estimation (MLE). Denote the exposures in the t,d cell by $E_{t,d}$ and the deaths by $D_{t,d}$. The Poisson model is that $D_{t,d}$ is Poisson in $m_{t,d}E_{t,d}$, where $m_{t,d}$ could come from either the LC or RH model. With mean $\mu$, the log of the Poisson probability at k is $k\log(\mu) - \mu - \log(k!)$. The loglikelihood is then:

$$\sum_{t,d}\{D_{t,d}\log[m_{t,d}E_{t,d}] - m_{t,d}E_{t,d} - \log[D_{t,d}!]\}. \tag{1.3}$$

Two forms of the negative binomial distribution are also fit. The negative binomial has two parameters r and $\beta$, with mean $r\beta$ and variance $r\beta(1+\beta)$. But in modeling a whole array of negative binomial variates it is customary to make the mean a parameter and model it with the covariates. In this case the mean would still be $\mu_{t,d} = m_{t,d}E_{t,d}$, as in the Poisson case.

To make the mean a parameter, set $\mu = r\beta$. The two forms arise by either eliminating r by setting $r = \mu/\beta$, or eliminating $\beta$ by setting $\beta = \mu/r$. Here these are called NB1 and NB2, respectively. Both

have mean $\mu$, but NB1 has variance $\mu(1+\beta)$ and NB2 has variance $\mu(1+\mu/r)$, which are linear and quadratic in $\mu$, respectively. Denoting the log of the gamma function by lgamma, the log of the probability at k for the negative binomial in r and $\beta$ is:

$$\text{lgamma}(r+k) + k\log(\beta) - \text{lgamma}(r) - \text{lgamma}(1+k) - (r+k)\log(1+\beta). \quad (1.4)$$

The loglikelihoods for NB1 and NB2 can be obtained by substituting $\mu/\beta$ for r or $\mu/r$ for $\beta$, then $D_{t,d}$ for k and $m_{t,d}E_{t,d}$ for $\mu$, and summing over the observations.

## 2. FITS

Goodness of fit of different models can be compared using penalized likelihood. The traditional comparison is to start with the negative loglikelihood (NLL) and add a penalty. Here the traditional criteria divided by 2 are used, as these are more directly related to the NLL, but the standard names are retained. Thus the Akaike Information Criterion (AIC) uses a penalty of 1 for each parameter. If N is the sample size (number of observed cells), the Bayesian Information Criterion (BIC) uses a penalty of ½ log N for each parameter. There is some feeling among information theorists that the AIC is too lenient on extra parameters, but the BIC is too punitive. The Hannan-Quinn Information Criterion (HQIC) is intermediate. It gives a penalty of log log N for each parameter. It turns out that most of the conclusions are the same for each criterion, so until a difference arises, only the BIC will be used, but HQIC will be the fallback if there is a difference. For N = 1260, the penalty is about 3.57 per parameter. Thus an extra parameter has to improve the NLL by that much to be justified.

LC and RH Poisson models were fit to male and female mortality. For both datasets, the RH model fit quite a bit better than LC. The RH NB1 and NB2 models were then fit. Table 1 shows the NLL for each model and the improvement in NLL required to meet the BIC requirement for the extra parameters from the model above it. After the parameter constraints there are 35 a parameters, 34 b parameters and 35 h parameters, so the LC model has 104 parameters. In the RH model there are 34 c parameters and 69 u parameters, for cohorts 1882 to 1951, ex 1946. Thus it has 207 parameters. The negative binomial versions have yet one more parameter.

**Table 1. Fit Comparisons**

| | NLL | | | | NLL Improvement | |
|---|---|---|---|---|---|---|
| **Model** | **Female** | **Male** | **Parameters Added** | **BIC Needed NLL Improvement** | **Female** | **Male** |
| LC Pois | 13670 | 14868 | | | | |
| RH Pois | 9598 | 10081 | 103 | 368 | 4072 | 4787 |
| RH NB1 | 8798 | 8996 | 1 | 3.6 | 800 | 1085 |
| RH NB2 | 8748 | 8972 | 0 | 0 | 50 | 24 |

The LC model fits considerably worse for males, with an NLL 1198 higher than for females. The RH model fits much better for both, with an improvement in NLL of 4072 for females and 4787 for males, compared to an improvement of 368 required to justify the extra parameters according to BIC. The difference between males and females is narrowed to 483, so RH is an even more substantial improvement for males. Adding the extra parameter for NB1 also significantly improves both fits, and NB2 is a bit better yet. The β parameter for NB1 is about 2.5 for females, and 3 for males, so the variance for each cell is 3.5 to 4 times the cell mean, compared to equal to the mean for Poisson. That is a substantial difference, and with variances that big it is no wonder the Poisson fit is not as good. The r for NB2 is about 8600 for females and 7500 for males, which for this data translates to variances of 2 to 7 times the mean, with the higher ratios going to the larger cells.

The best NB NLLs for LC were 9444 for females and 9652 for males, so Poissoness is the bigger culprit for LC – Poisson than the lack of cohort parameters. Still the NB RH model is better than NB LC for females by 696 and for males by 680, which are still well above the BIC need of 368, although nowhere near the NLL improvements of 4000+ for the Poisson models.

To give a visual impression of the fits, the empirical and modeled values of log m are graphed for a few years of death by age at death for the Poisson models. The graphs do not look much different for the negative binomial models, and in fact the parameters are not that different either. The advantage of the negative binomial models is more in the error distributions than in the fitted means. Essentially the cells with higher variance are not penalized as much in the likelihood functions for being different from their means, so the fit gets better for the smaller cells. This is not enough to be very noticeable in the graphs, however.

Figures 1 and 2 show the female data and fits. The mortality rates increase by age and this is close to a linear function for the log rates. In a graph of the rates for several calendar years of death, most

of the vertical range is taken up by this increasing trend, which makes it difficult to see the differences among the calendar years. To look at goodness of fit, the linear trend by age is not so critical, so in the graphs this is eliminated by subtracting age at death / 11 from each log mortality rate, essentially rotating the graph to make the lines roughly horizontal. This makes all the vertical range available to compare the actual and fitted rates for the various calendar years. A constant of 10.5 has been added to make the resulting numbers start near zero on the vertical axis. Rates have been declining over time, so the most recent calendar year is at the bottom of the graph. The dotted lines are the data, and the solid lines are the model.

Figures 3 and 4 are similar for males, but more years are able to be shown as the trends are greater for males, which separates the years a bit. Also, since the male mortality rates are higher, the rotated rates start around 0.45 instead of zero.

**Figure 1. Rotated Graph of LC Female Mortality Rates**

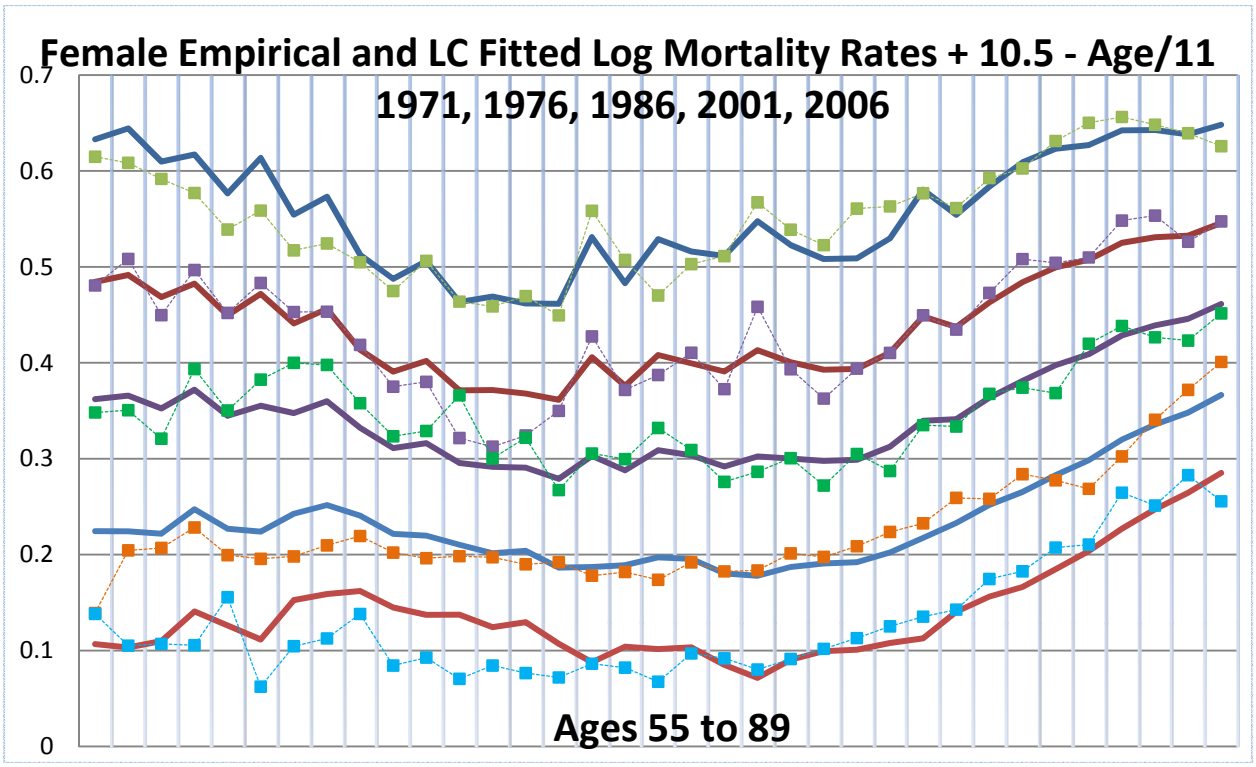


**Figure 2. Rotated Graph of RH Female Mortality Rates**

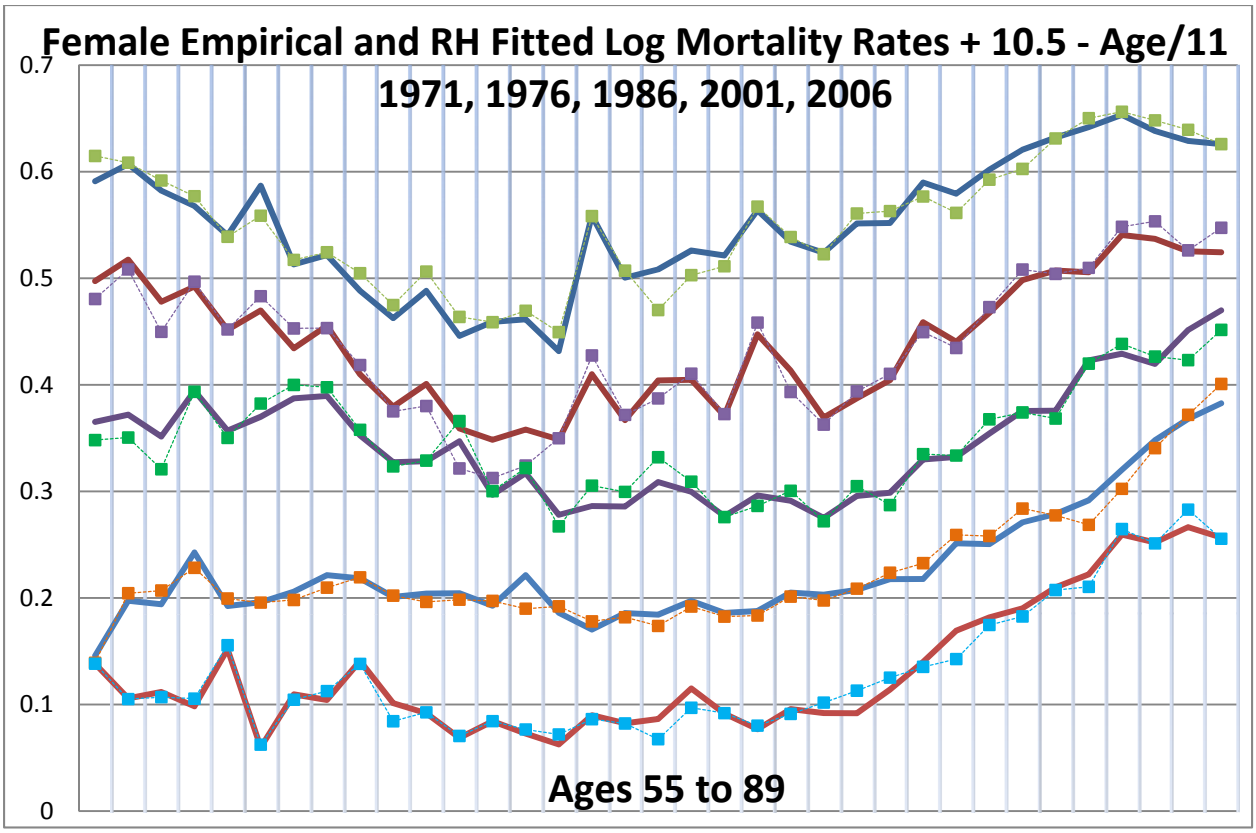


**Figure 3. Rotated Graph of LC Male Mortality Rates**

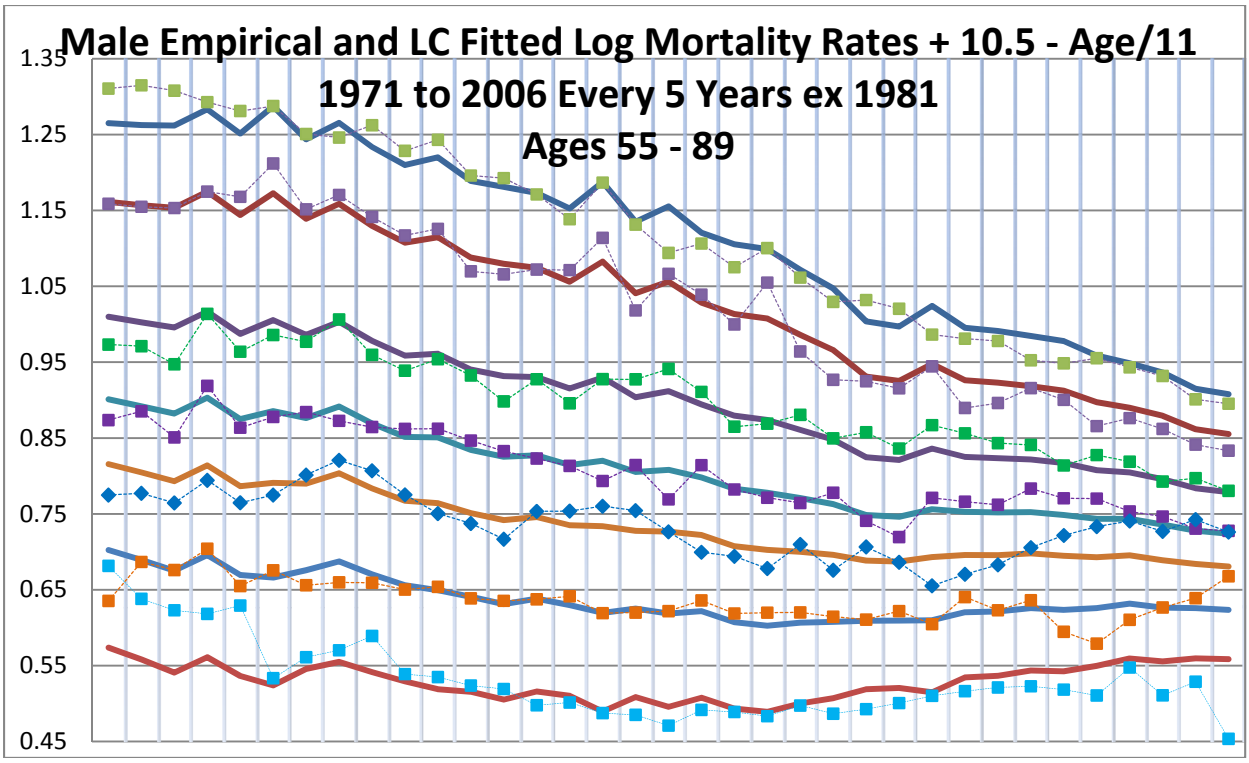
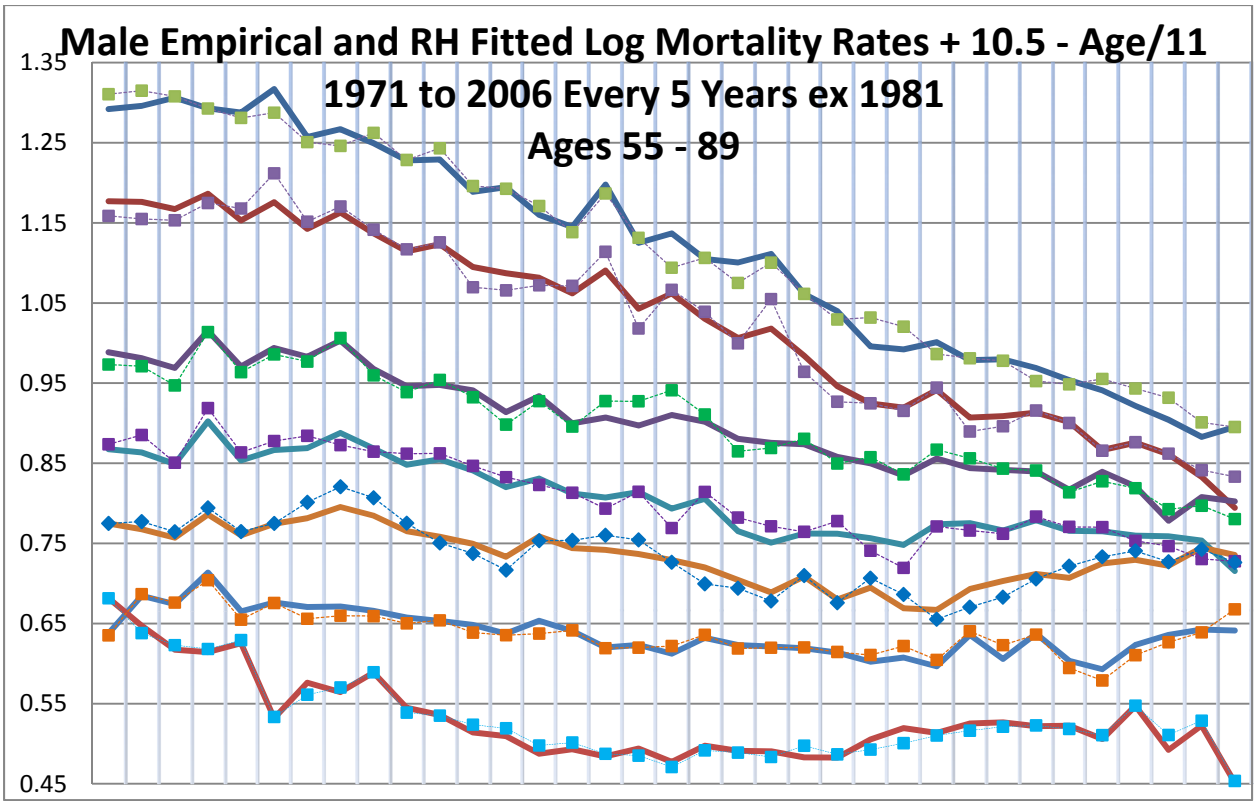


**Figure 4. Rotated Graph of RH Male Mortality Rates**




For males, the downward trend from year to year is less at the older ages. The LC model can handle this by having lower b factors for the older ages. When the LC model misses, it seems to be mostly for the youngest and oldest age groups. By adding in cohort parameters, the RH model can account for many of these effects. However, some of this is suspicious, as some of the cohorts have few observations in the data. Hence the modeled mortality at age 55 is higher in 2006 than in 2001, which follows the data, but that in itself does not establish 1951 as a high-mortality cohort.

For females, the shape of the graph is somewhat different than for males, and there is not so much diminishing trend at older ages. The LC model does not have enough flexibility to capture the changes in shape, and some of the years have long strings of significant errors of the same sign. The RH cohort-effects are able to adjust for a lot of this, but again this is sometimes because of cohorts with only a few observations, such as age 55 for 2001 and 2006.

The RH model provides better fits both in graphical tests and penalized MLE. Although the fits are worse for males, the difference narrows considerably for the better-fitting models. The negative binomial versions are much better than the Poisson, with the NB2 a bit better-fitting than the NB1. There are some plausibility problems with the RH model, however.

## 3. INTERPRETING THE PARAMETERS

The best-fitting NB2 parameterizations are used in this section. Does the fact that NB2 fits better than NB1 have any implications? NB2 is the form that comes from mixing a Poisson by a gamma distribution. This arises in experience rating, for instance, if each policy is Poisson-distributed, but there is a gamma distribution of Poisson means across the population. Taking a policy at random, its claims are conditionally independent given its Poisson mean, but unconditionally correlated due to the common Poisson mean. This is a way of modeling non-independent claims, or contagion.

It is tempting then to argue that the population as a whole is a mixture of groups with different mortality, due to different lifestyles, access to medical care, etc., and that is the source of the contagion observed. However that is a different kind of mixture. The population as a whole consists of all the groups taken together, not one drawn at random. The sum of independent Poisson distributions is itself Poisson, so the mixture argument does not explain contagion at the level of the entire population. Moreover, the number of deaths is the sum of Bernoulli processes and would be binomial, not Poisson, if there were not already some source of contagion to begin with.

There are factors affecting mortality rates for the population as a whole, such as weather, flu outbreaks, etc., that make deaths not independent. This could be the principal source of contagion at the population level. The NB1 model makes the variance about 4 times the mean for each cell in the data, whereas for the NB2 model it ranges from about 2 to 7 times the mean, with the factor larger for the larger cells. The fact that NB2 fits better suggests that the contagion events hit the larger cells harder. That is, the ages with the greatest number of deaths also have the greatest increases in deaths when adverse conditions arise.

Figure 5 graphs the $a_d$ parameters, which represent the base log mortality rate by age, before application of trends and cohort effects, for males and females. Male mortality is higher than female at all ages, but that does not show with these parameters. The calendar-year parameters and cohort parameters interact with these so in themselves they are not that meaningful.

**Figure 5. $a_d$ Parameters**

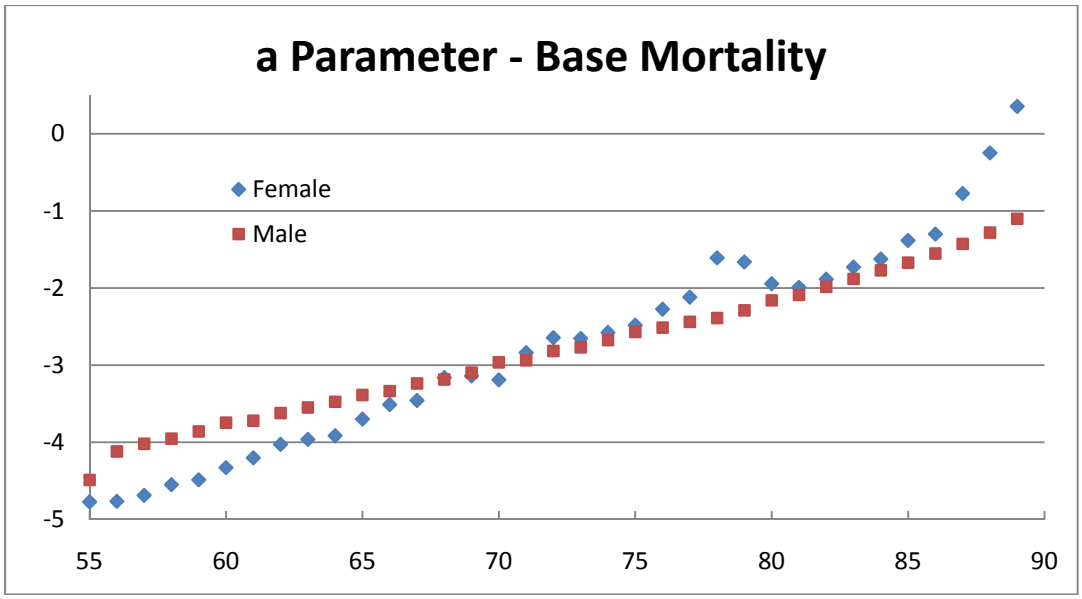
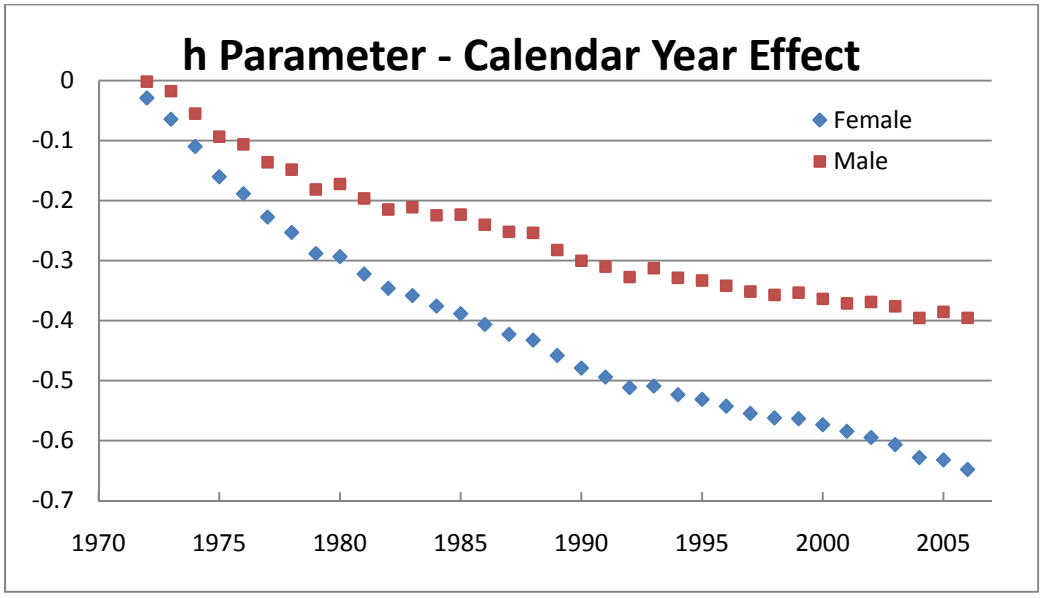



Figure 6 shows the calendar-year trends as reflected in the $h_t$ parameters. These were forced to be zero at year 1971. The female levels are trending faster, even though at the younger ages in the range the male rates are dropping faster. This is a distortion of the calendar-year parameters due to interaction with other parameters.

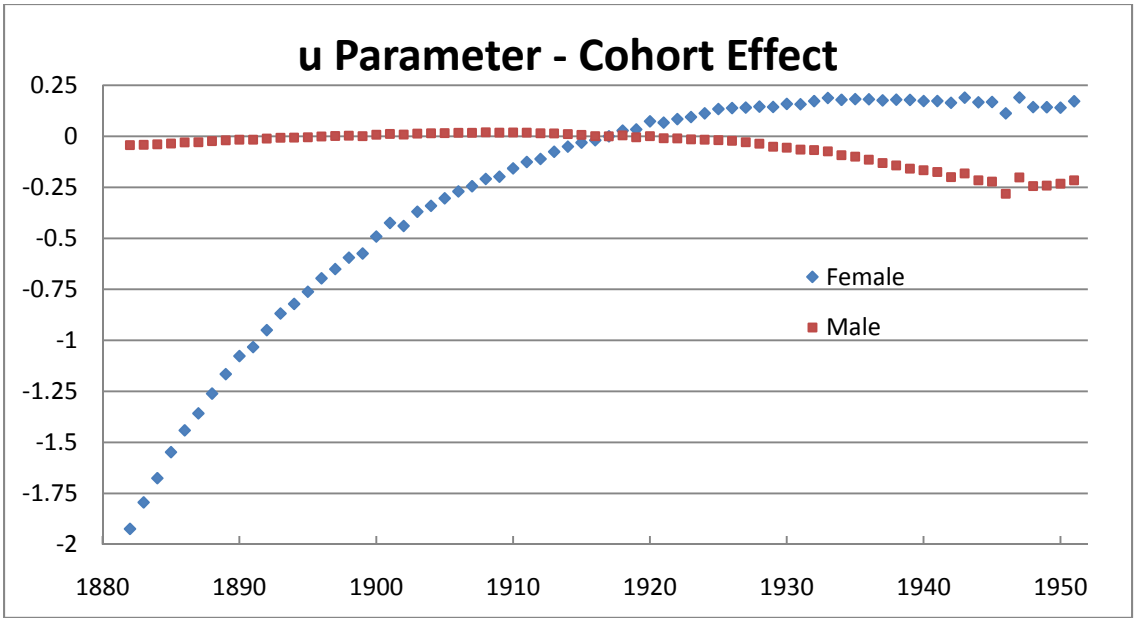
**Figure 6. $h_t$ Parameters**




The $b_d$ parameters, which modify the overall trend to produce an age-specific trend, are shown in

Figure 7. The surprise here is the accentuation of the trend effect for male octogenarians, whose trend has actually been less than for other ages. In the model the cohort effects offset this effect to match the data. The last cohort that affects ages 87, 88 and 89 is 1899, and cohorts prior to 1900 do not get into this dataset at ages less than 72, leaving room for the parameters to adjust themselves to produce the best possible fit at older ages without affecting younger ages. This raises questions, however, about the applicability of the parameters beyond this data range.

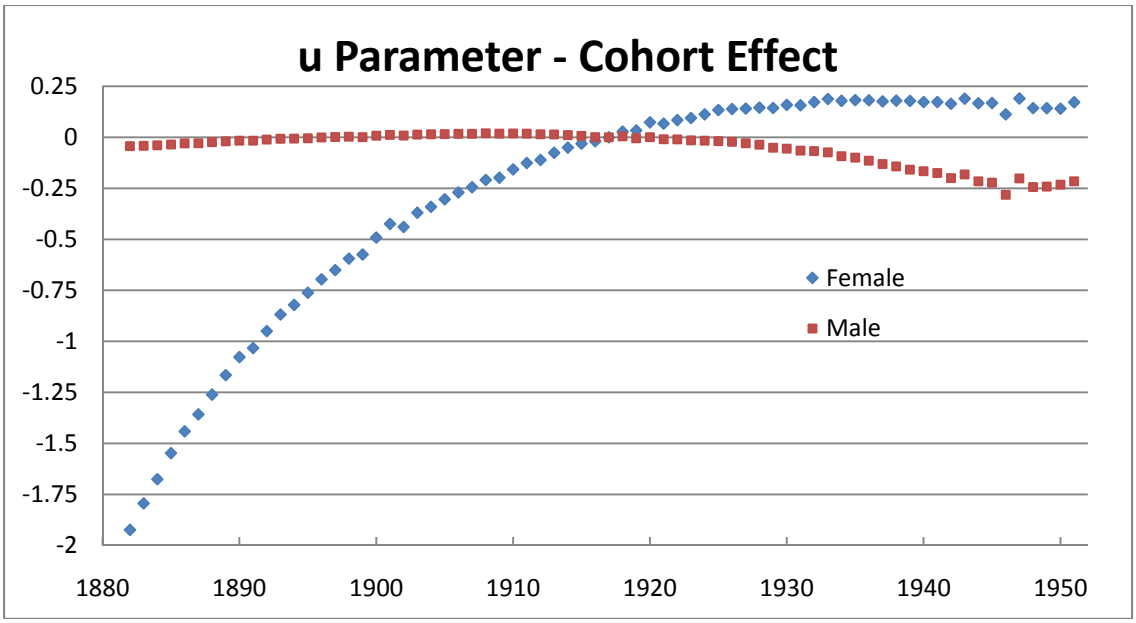
**Figure 7. Trend Age Modifiers $b_d$**




The cohort parameters $u_{t-d}$ are graphed in Figure 8. This includes the 1917 cohort that is forced to zero. The sharply lower mortality for females in the earlier decades is not realistic. It produces an increasing trend across cohorts, which partially offsets the trend across calendar years.

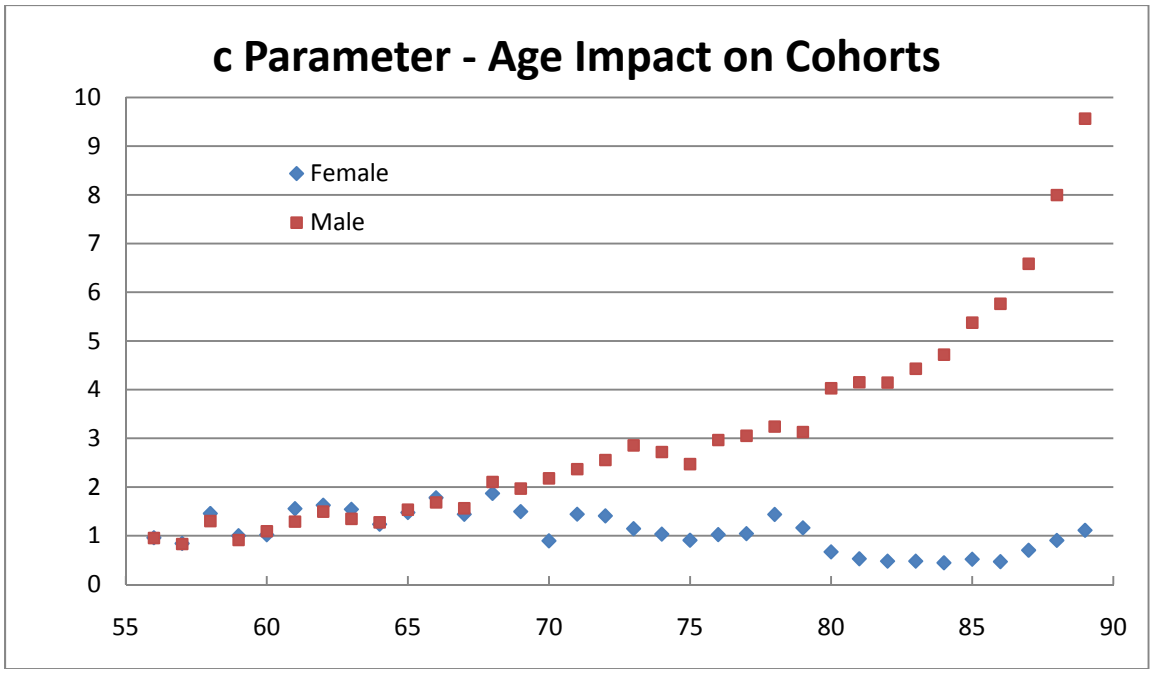
Although it is much milder, an increase in mortality is seen in the first 30 or so cohorts for males. A possible explanation is that these reflect a selection effect. It was more unusual for men and women born in the 1880s to reach higher ages, so those who did were a select group more likely to live even longer. The cohort parameters are intended to represent a mortality differential for the cohort that would apply at all ages, but because not all ages are in the data, the most it could represent is a conditional differential, conditional on having lived long enough to get into the data in the first place. Such an effect would be further enhanced by the fact that the oldest cohorts only appear in the oldest ages in this data.

**Figure 8. Cohort Parameters $u_{t-d}$**




Finally the $c_d$ parameters, which are by-age modifiers to the cohort effects, are in Figure 9. These are forced to unity at age 55, and stay near there for a year or two before moving to the vicinity of 2 for ages in the late 60s. From there the female parameters get generally lower for the older ages, indicating a reduced impact of the cohort effects at later ages. This partially offsets the much greater trend in cohorts for females. For the males, the parameter accelerates in the 80s, ending near 10. This appears to be part of the fine-tuning of the fit for higher ages made possible by the cohort parameters with few supporting observations.

**Figure 9. Cohort Effect by Age Parameters $c_d$**




The observed slower trend in mortality at older ages for males is modeled in this RH parameterization by an underlying higher trend at older ages (b parameters), offset by a starting group of cohorts who had lower mortality rates to begin with (u parameters), the effects of which increased sharply at older ages (c parameters). The LC plus cohorts model for female mortality is an even stranger combination of offsetting effects.

## 4. FIXING THE FITS

One of the problems with the fits above is the sparsity of data in many cohorts. Another is the high correlation among parameters in the female model, which is discussed later. For the recent cohorts, the only way to add data is to wait. For the older cohorts, however, there is data available. Extending the data to include all calendar years of death for cohorts 1882 to 1915 with death ages 55 to 89 is possible, and this increases the number of observations to 35 for each such cohort. This would be expected to give better estimates for those cohort parameters $u_{t-d}$, but also for the $c_d$ parameters that modify the cohort parameters for age effects, and indirectly on all the other parameters, which may have less flexibility to fit to random fluctuations in the data. The original data will be referred to as the partial data and the expanded set as the full data.

The per-parameter penalty log(N)/2 for BIC goes up to 3.763, with 1855 observations, from 3.569 for the 1260 observations in the partial data. The full data need calendar-year parameters $h_t$

starting with 1937 (assuming zero at 1936) instead of from 1972, so the models have 33 more parameters. There are no additional a, b, c or u parameters. Thus the LC model now has 137 parameters, and the RH negative binomial models have 242 parameters. The full sequence of models above was fit, but now the NB1 fits better for males. To resolve this, more distributions were fit for the RH model. The NB3 is intermediate between the NB1 and NB2; the Poisson-inverse Gaussian (PiG) is similar to the negative binomial, but is more skewed; and the Sichel is a three-parameter generalization of the PiG, which can be more or less skewed than the PiG but not less skewed than the NB. These distributions are discussed further in Appendix 1. The results are:

**Table 2. Triangle Fit Comparisons**

| | NLL | | Parameters Added | BIC Needed Improvement | BIC Improvement | |
|---|---|---|---|---|---|---|
| **Model** | **Female** | **Male** | | | **Female** | **Male** |
| **LC Pois** | 21,726 | 24,047 | | | | |
| **RH Pois** | 15,452 | 16,630 | 103 | 388 | 6274 | 7417 |
| **RH NB1** | 13,176 | 13,567 | 1 | 3.8 | 2276 | 3063 |
| **RH NB2** | 13,172 | 13,576 | 0 | 0 | 4 | -9 |
| **RH NB3** | 13,163 | 13,568 | 0 | 0 | 9 | -1 |
| **RH PiG** | 13,172 | 13,567 | 0 | 0 | -0.4 | 0.015 |
| **RH Sichel** | 13,172 | 13,565 | 1 | 3.8 | 0 | 2.2 |

Again the RH model provides a tremendous improvement in the Poisson fit, as does moving from Poisson to negative binomial. The NB3 is the best fit for females, but the NB1 is the best NB for males. The difference between the NB models is that VM, the variance/mean ratio, is fixed at $1+\beta$ for the NB1, is $1+\mu/r$ for the NB2, and is $1+ (\mu/r)^{1/2}$ for the NB3. For females the cell means range from 6000 to 44,000. With the fitted parameters, this gives VM of 5.1 for NB1, 2.4 to 11.1 for NB2, and 3.4 to 7.4 for NB3, which gives the best fit. For males the NB1 VM is 6.1. Another version of the NB discussed in Appendix 2 fits slightly better with a range for VM of 5.3 to 6.7, but uses an additional parameter which does not give enough better fit to justify it.

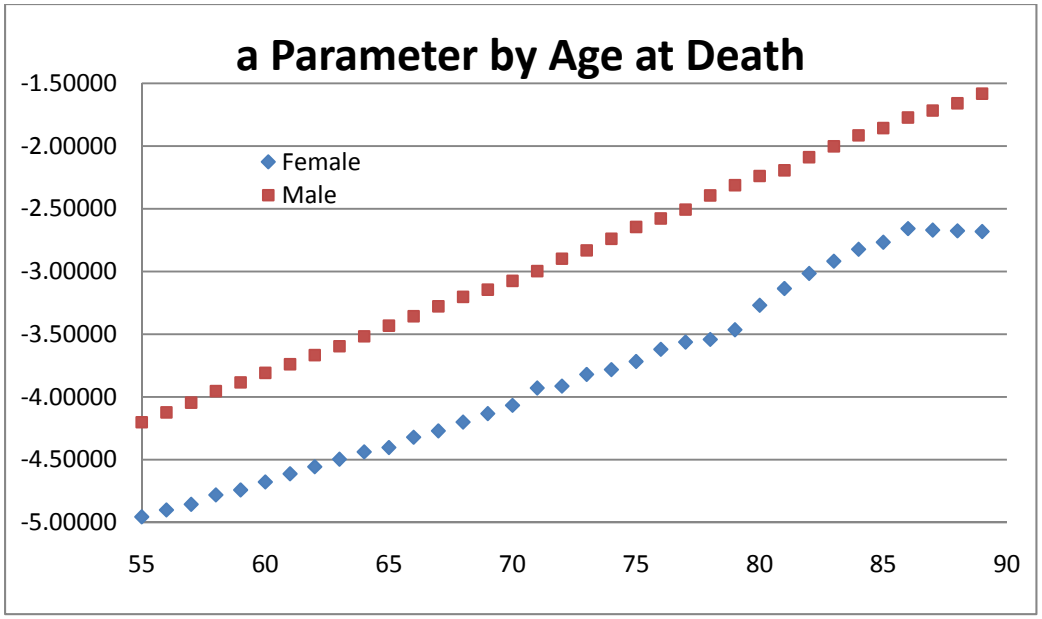
The improvements shown for the last three models are from the better of NB1 and NB2. The PiG and Sichel models also have 1, 2 and 3 versions like the NB. For females, the 2 version of the PiG was found to be slightly worse than the NB2, indicating that the additional skewness was not helpful. The corresponding Sichel has the NB2 as a limiting case, but otherwise has higher skewness than the NB2 with the same mean and variance. The fact that it did not give any improvement over the NB2 suggests that, if anything, less skewed distributions may fit better for females.

For males the PiG, version 1, was very slightly better than the NB1. The Sichel fit even better with an intermediate skewness. However, the improvement in NLL is problematic. At 2.2 it is less

than the 3.8 required by the BIC, but better than 1, which the AIC requires, or 2.0, which the HQIC calls for. There is a good deal of literature suggesting that BIC is too stringent in rejecting parameters. Burnham and Anderson (2004) make a strong push for AIC and the small sample AIC, based on the idea that the sample is not generated from the model being fit, but rather the model is a fairly compact representation of a more complex process. For a sample size of N and p parameters, the small sample AIC penalizes the NLL by Np/(N–p–1). With N = 1855, the additional penalty for the 243rd parameter over the 242nd is 1.32. Thus the AIC, HQIC and small sample AIC all support the additional parameter for the Sichel distribution in this case. Thus it will be taken as the best-fitting model.

The parameters shown below are from the best-fitting Sichel model for males and NB3 model for females. It appears that the full data helps with the male model but does not solve the problems with correlation in the female model.

**Figure 10. Base Mortality—a Parameter**




The a parameters in Figure 10 look reasonable for males but strange for females, especially the decline for the oldest ages. The calendar-year h parameters in Figure 11 also appear reasonable for males but trend upward for females. In this parameterization for females, the downward mortality trend over time ends up as a cohort trend, partially offset with an opposing calendar-year trend.

**Figure 11. Calendar-Year h Parameters**

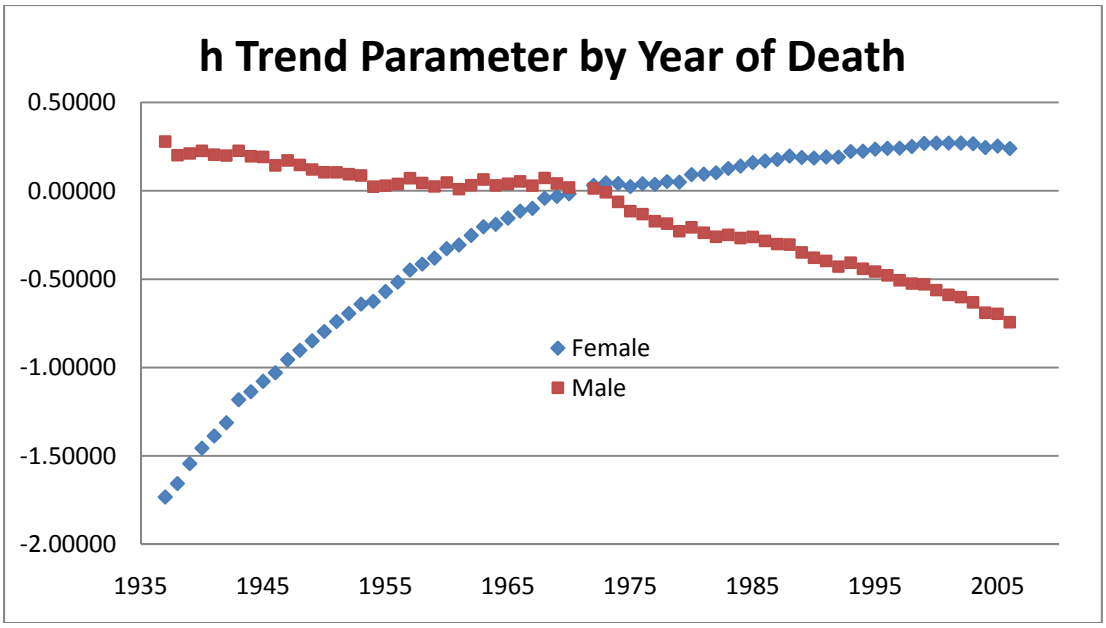
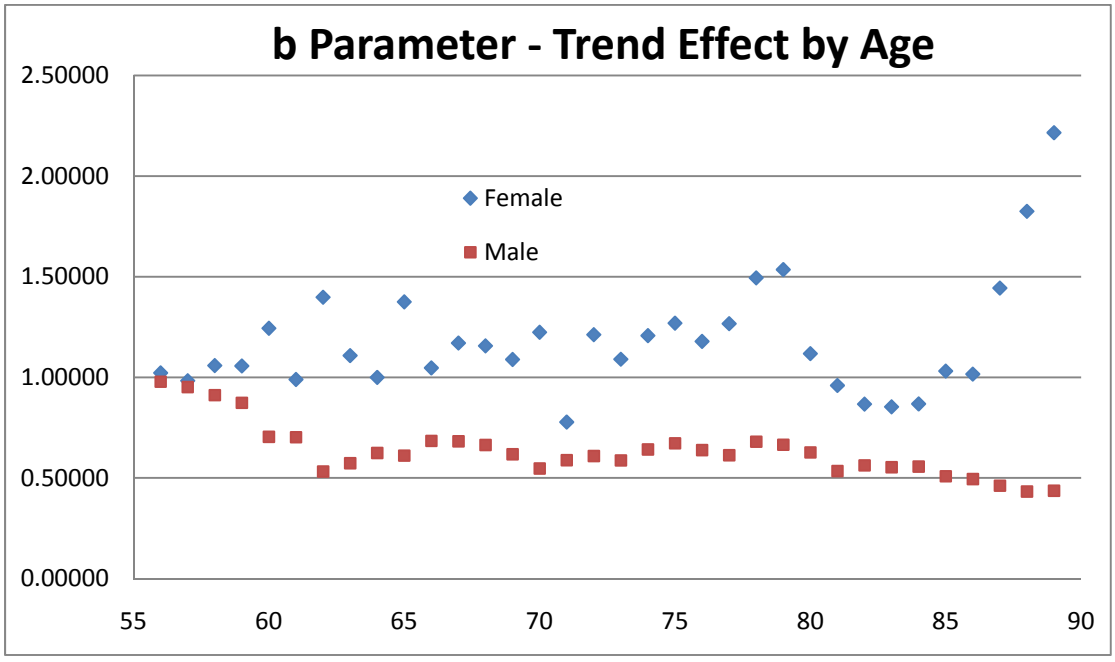



**Figure 12. Trend Effect by Age at Death – b Parameters**

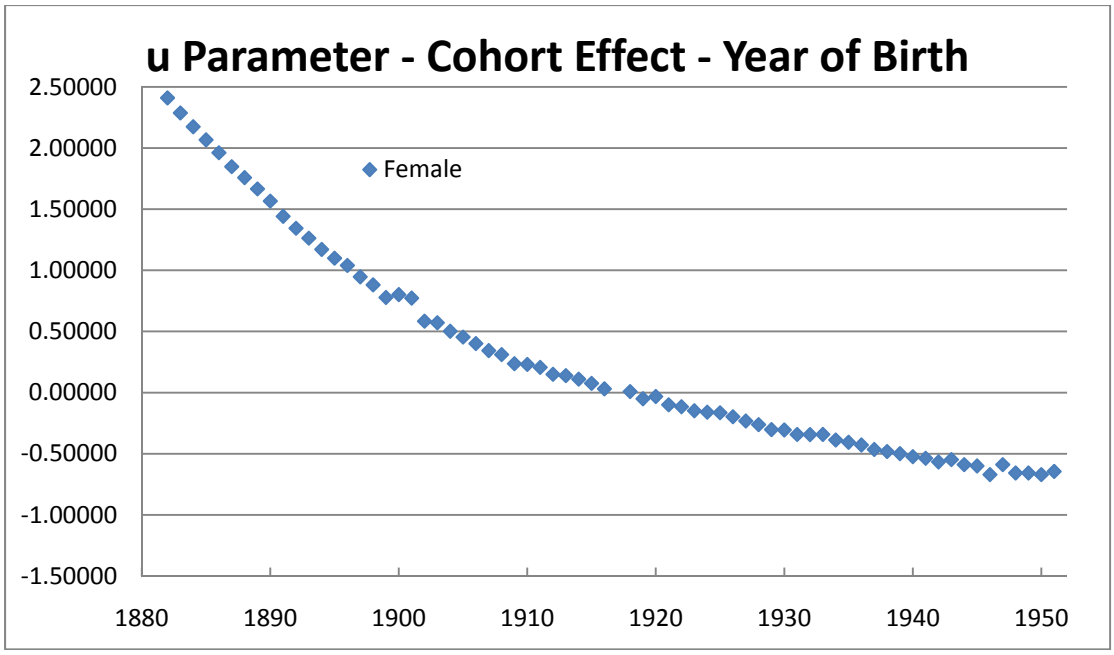



For the b parameters in Figure 12, the sharp upward movement at the oldest ages for females probably has something to do with the lower base mortality at the corresponding points.

The female cohort trend is in Figure 13, which shows a sharp downward trend in mortality in the

direction of later years. This overwhelms the upward trend by calendar year to produce an overall downward trend in mortality, which matches the data, but is not intuitive as an explanation of the data. The h and u parameters with the full cohorts for females are mirror images of what they are for the partial cohorts. This is discussed further below.

**Figure 13. Female Cohort Effect u**




The male cohort parameters are on a completely different scale and so are graphed separately in Figure 14. The full and partial cohort parameters are consistent for males and so can be graphed together. The effect of conditioning on attaining various ages is clearer in the partial cohorts, where the conditioning is on progressively older ages, peaking in about 1910. There is a similar but much smaller effect in the full cohorts, perhaps due to a changing significance on the fact of attaining age 55. In both cases, there is an increase in the mortality in the most recent cohorts, but this is based on very few data points. Also for the male model, the c parameter, set to 1 at age 55, stays that low only for a few ages then goes to much higher values at older ages, as shown in Figure 15. Thus, the higher cohort parameters for the latest cohorts are getting relatively low c parameters applied, and are not likely to remain so low when more data comes in, with higher c parameters.

**Figure 14. Male Cohort Effects for Partial and Full Cohorts**

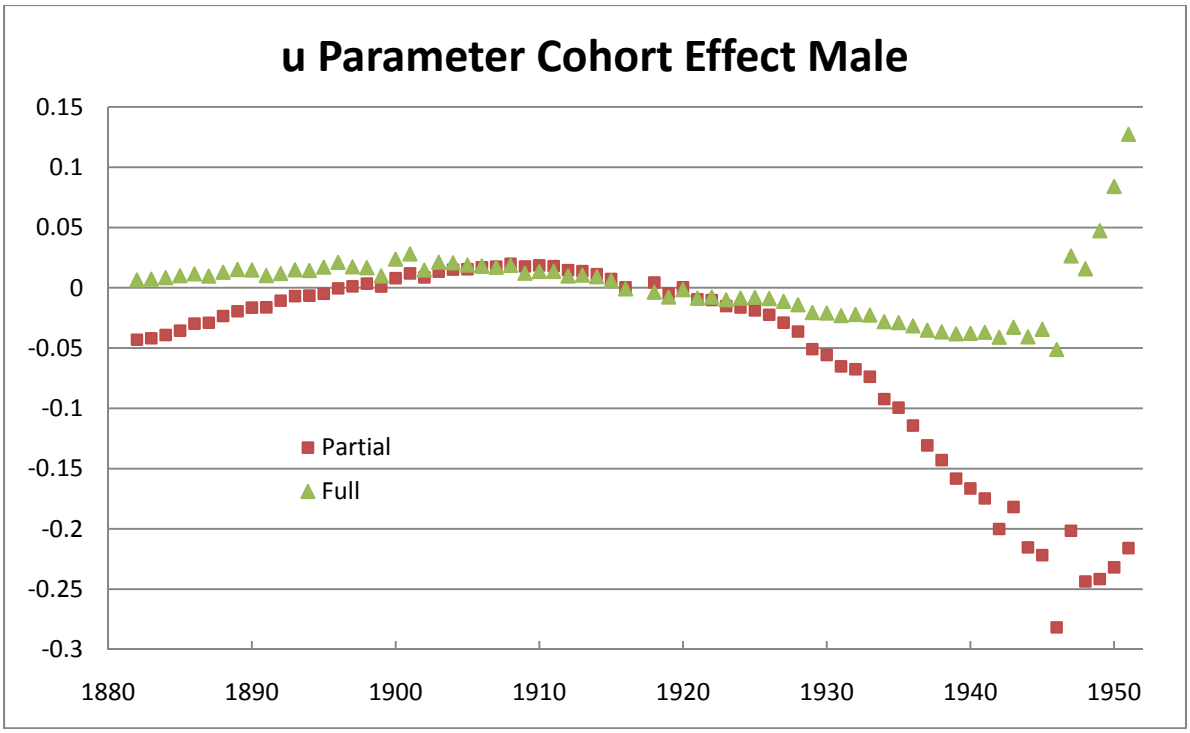
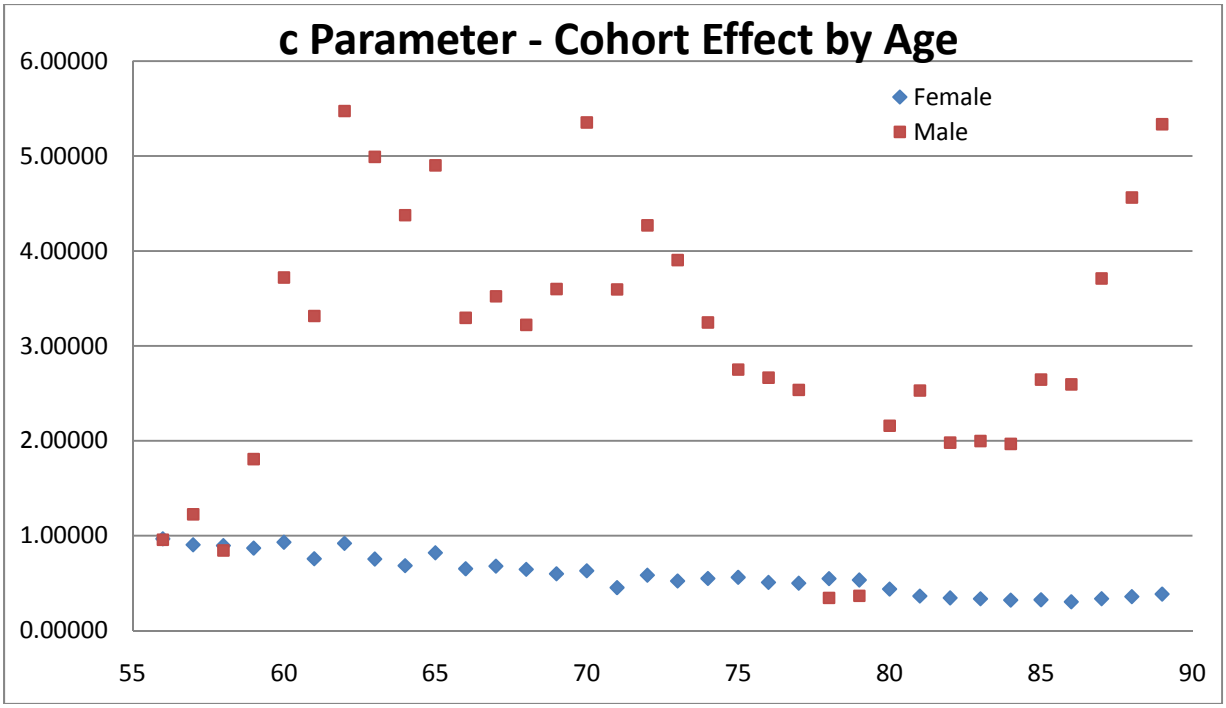



**Figure 15. Age Impact on Cohorts—c Parameter**



An idea of the statistical significance of the parameters can be gained by estimating the parameter covariance matrix as the inverse of the Fisher information matrix from the MLE estimation. Recall that this is the matrix of all 2nd partial derivatives of the NLL. This yields parameter standard deviations and so t-statistics for each parameter and also covariances, and so correlations, among parameters.

For the male triangle parameters, virtually all the parameters had t-statistics with absolute values above 2. The few exceptions are parameters very close to zero, usually near points that were forced to be zero. With 242 parameters, there are over 25,000 correlations, so they are not printed here. However, the averages of the absolute value of the correlations by type of parameter (excluding parameters with themselves) are shown in Table 3 for males and Table 4 for females.

**Table 3. Average Absolute Value of Correlations by Parameter Types—Males**

| | a | h | b | u | c |
|---|---|---|---|---|---|
| a | 52.7% | 31.9% | 36.1% | 43.7% | 25.3% |
| h | 31.9% | 35.6% | 22.3% | 45.9% | 37.3% |
| b | 36.1% | 22.3% | 42.7% | 27.3% | 22.0% |
| u | 43.7% | 45.9% | 27.3% | 79.0% | 67.3% |
| C | 25.3% | 25.3% | 22.0% | 67.3% | 67.4% |

**Table 4. Average Absolute Value of Correlations by Parameter Types—Females**

| | a | h | b | u | c |
|---|---|---|---|---|---|
| a | 98.9% | 98.2% | 9.4% | 97.6% | 34.1% |
| h | 98.2% | 98.9% | 8.4% | 98.4% | 31.6% |
| b | 9.4% | 8.4% | 40.9% | 8.5% | 30.5% |
| u | 97.6% | 98.4% | 8.5% | 97.9% | 31.3% |
| c | 34.1% | 34.1% | 30.5% | 31.3% | 56.5% |

The extremely high correlations among the a, h and u parameters in the female model make the individual parameters highly questionable. There could be many local maxima of the likelihood function, and there is no guarantee that the parameters found are a global maximum. Even if they are, the correlations make the parameter values unstable. In fact, the partial and full datasets gave oppo-

site but similarly offsetting directions for the female calendar-year and cohort trends.

This shows up in the t-statistics as well, which are near 1 in absolute value, so not significant, for all the h and u parameters in the female model.

Moving to the full cohorts then appears to improve the male model, which has reasonable parameters and correlations among parameters, as well as significant t-statistics. For the female model, the high correlations (which, though not shown, are similar for the partial cohorts) make the fit problematic.

Usually when there are high correlations, the solution is to leave out some variables. But the greatly improved fit of the RH model over the LC model appears to rule out omitting the cohort parameters. Parameter reduction through smoothing would still leave quite problematic parameter values as well. One option may be to keep the cohort parameters but not the calendar-year parameters, making the trend a purely cohort matter. It does not seem likely that this would give a good fit, but it might be worth trying.

Another option would be to set the base mortality a parameters as the average or some weighted average of the mortality rates for each age in the full data. This was actually Lee and Carter's initial recommendation. This would give the other parameters less opportunity for mischief. A similar approach could be to use a parameterized curve, like Makeham or splines, for the base mortality. Yet another possibility might be to multiply the cohort and calendar-year parameters, and then apply a single age parameter to the product. This type of model is used extensively in casualty loss reserving, but has had mixed results (informally communicated) in mortality studies.

## 5. MORTALITY CURVES

The raw mortality rates for each year of death are somewhat noisy, and so cannot be readily compared graphically. However fitting mortality curves, like Makeham curves, to each year smoothes the data and lets the trends stand out more clearly. Here a generalized Makeham (GM) function is fit to the raw death rates, although fitting to force of mortality is more typical. Richards (2008) discusses some such generalizations, based on earlier work by Beard (1959) and Perks (1932). Using a curve to fit the $a_d$ parameters requires a log transform, and the form used here takes 4 parameters $\alpha$, $\beta$, $\theta$, $\gamma$:

$$a_d = \theta + \log[(1+\alpha\beta^d)/(1+\gamma\beta^d)]. \tag{5.1}$$

Fitting such curves with four parameters to the log death rates in each year 1971 – 2006 results in the use of 144 parameters, compared with 104 for LC and 207 for RH with partial cohorts. Using the best-fitting negative binomial, the following values of the NLL were produced:

Table 5. Comparative Fits Including GM

| | NLL | | Parameters Added | HQIC Needed Improvement | NLL Improvement | |
|---|---|---|---|---|---|---|
| **Model** | **Female** | **Male** | | | **Female** | **Male** |
| LC | 9444 | 9652 | | | | |
| GM | 9482 | 9553 | 40 | 79 | -38 | 99 |
| RH | 8748 | 8972 | 63GM, 103LC | 124GM, 202LC | 696 LC | 581 GM |

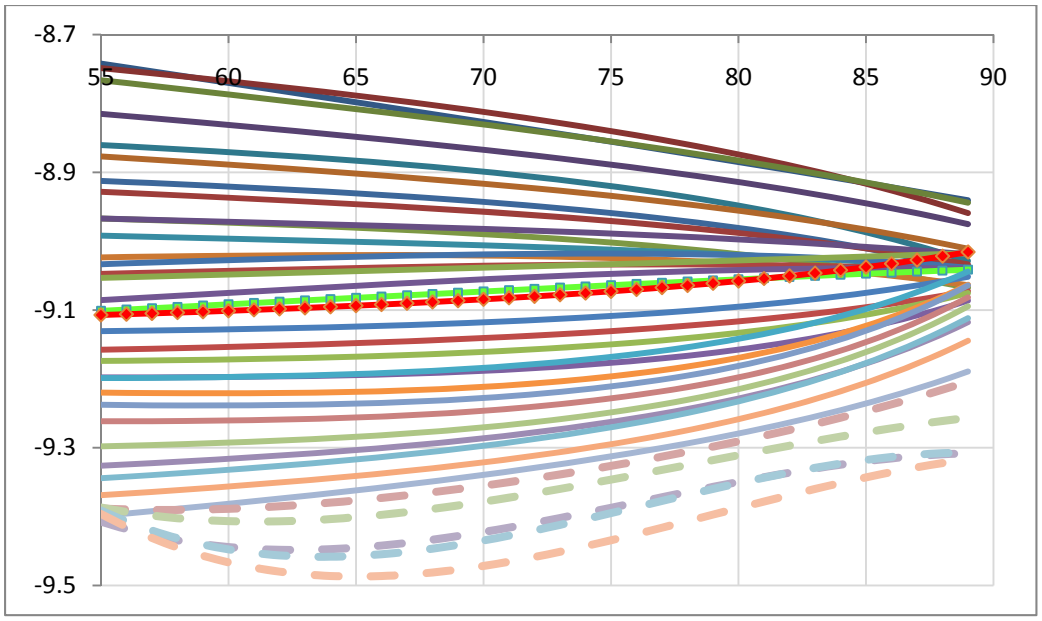
The goodness of fit test here is the HQIC, which requires an improvement in NLL of log(log(sample size)) for each extra parameter, where here the sample size is 1260, requiring an improvement of 1.96555 per parameter. This is intermediate between the AIC and BIC. As can be seen, every model fit the female data better than the male data, and the RH model gave the best fit to both data sets, even though it is of dubious interpretation here. The generalized Makeham curve fit better than LC for the males, where the mortality curve was changing more over time, but LC fit better for females.

Nonetheless, for both males and females, the curves provide continuous versions of the mortality functions for each year which are smooth enough to show all years on a chart, thus providing some insight into what the changes in the mortality functions have been.

The male curves in Figure 16 (with age/12 subtracted) actually divide into three periods. First for 1971 until 1987, which is the light line with the square markers, the curves are straight or downward-curving. Then starting in 1988 (dark with diamond markers) the curves bend upward. Until around 2001 or 2002 (first dotted curve) the mortality at age 55 is steadily improving, but the improvement at the other end of the curves is slower and sometimes non-existent. Then somewhere around 2000 to 2002 the improvement at age 55 stops and the improvement at the older ages accelerates. The last three years show a different shaped curve from the earlier years.

The changes in shape show why LC has problems fitting this data, but the fact that the biggest changes were at the ends of the lines shows why RH can give a big, albeit artificial, improvement in the fit. The graph suggests that projecting future changes in longevity has a high degree of uncertainty involved. Should you just project the last five years, or from 1988 on, or average improvements in mortality over all the data? This could make quite a difference, especially at some ages.

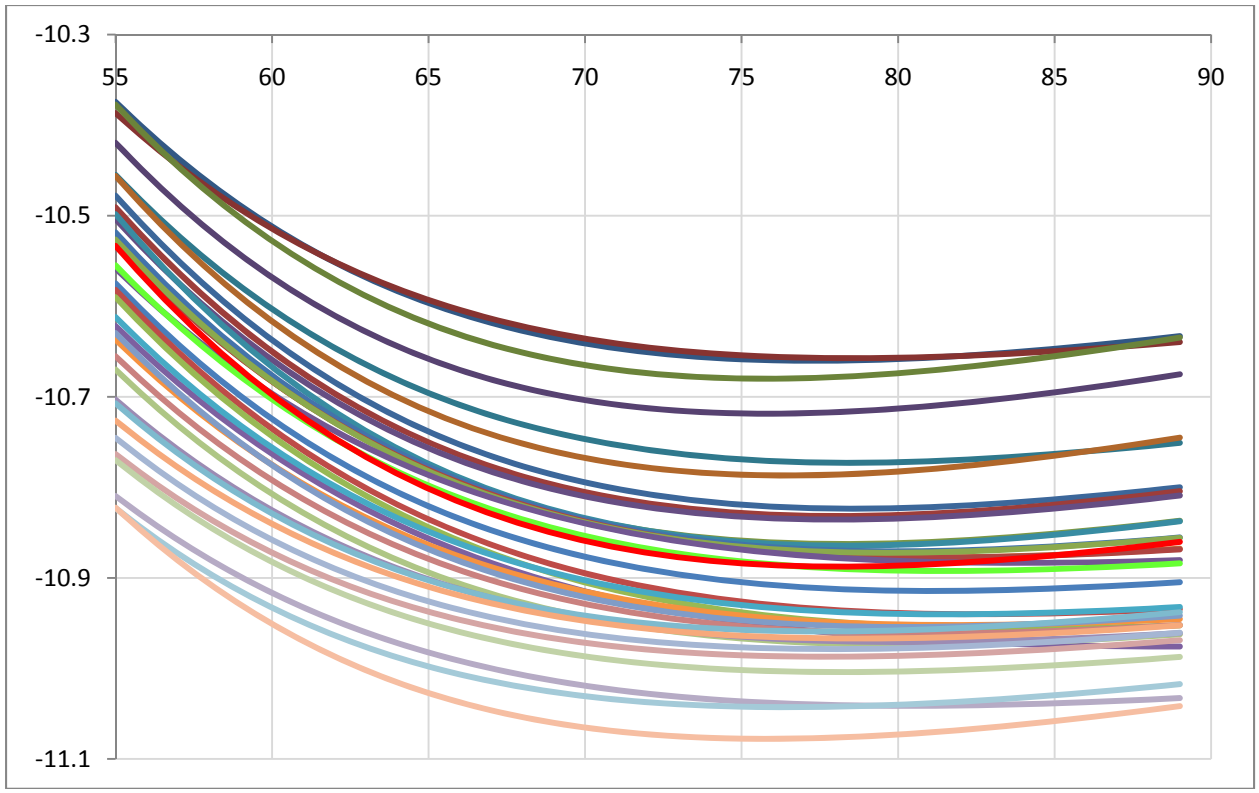
**Figure 16. Generalized Makeham Male**




The recent lack of improvement at age 55 is particularly problematic. That could be related to the recent reduced access to health care in the US for people under 65. If so, you would expect it to eventually improve over time as access improves. At the other end of the curve, it might be reasonable to assume that the older ages will improve at the same rate as most of the curve, as there seems to be a trend in that direction over quite some time. Nonetheless this is an assumption imposed on the projection process and thus adds to the projection uncertainty.

The generalized Makeham model did not fit as well as LC for females, but the fits in Figure 17 still provide some insights. Here age/10 was subtracted to remove the upward trend. It is apparent that there has not been as much change in the shapes of the curves as in the male model. What does stand out, however, is variation in the rate of mortality improvements across the age groups. For instance, for ages 75 and above, there were fairly long periods with very little improvement in mortality, punctuated here and there with years of substantial improvement. Ages 65 and below, on the other hand, had much more steady generally small improvements. As with the male data, there has been little improvement at age 55 in the latest few periods. Also since about 2000 there has been somewhat similar year-to-year improvements in the male and female graphs, even by age.

**Figure 17. Generalized Makeham Female**




As with the male model, this graph brings out some problems in projecting future trends. Can you assume the greater improvement in the last 5 or 6 years will now continue? Would a time-series model with highly fluctuating rates of improvement be better at the older ages? Perhaps in both genders it would be appropriate to calculate trends under different assumptions then include all the scenarios, with selected weights, in the overall longevity improvement uncertainty model.

## 6. PROJECTION RISK

Projection risk can be calculated for a particular dataset of annuitants, which is not what is available here, but some general observations on how to carry out such a calculation using LC and RH models are presented.

To begin, the calendar-year trend levels have to be projected. Standard time-series methodologies produce ever-widening ranges as the trend continues. However, here there is another wrinkle, as the h parameters being trended are estimated parameters, and so are observed with error. An area of regression studies is errors-in-variables models, which has a number of potential methods. If the variances of the h parameters have been estimated and they are relatively constant, then a simple reasonable simulation of a future level could assume that same variance, and first simulate the future

levels with errors and then simulate actual future levels from there using that variance.

A number of mortality modelers have used AR1 models for the annual difference in levels, usually with a negative autocorrelation, to project the trend. Most of these do not take into account the errors-in-measurement issue, however. For an independent series measured with a constant error variance, differencing induces an autocorrelation of -50 percent, arising from the same error having opposite signs in consecutive observations, so the AR1 model may be distorted by the induced autocorrelation. Other mortality projection studies have used the Kalman filter, which recognizes measurement errors, to project the levels, but the simple Kalman filter is based on a random walk, which can have too much autocorrelation. An alternative to AR1 and the Kalman filter is state-space models, which provide common generalizations of both.

If projections are needed for cohorts not in the study, then trending of cohorts also has to be considered. Even the use of the recent cohort parameters should take into account their potential measurement errors, perhaps with a state-space model.

Parameter uncertainty can be implemented by simulating the parameters from the covariance matrix from the Fisher information matrix, which gives an estimate of the covariance matrix of the parameters. Asymptotically the parameters have a multivariate normal distribution with this covariance matrix, so they can be simulated using the normal copula, Cholesky decomposition, etc. However, even though the error distributions are asymptotically normal, they may not be normal for a finite sample, and other distributions could be used to simulate parameter risk, perhaps gamma, which is the exact error distribution for some models, and approaches the normal asymptotically. Other distributions that approach the normal could also be used. One criterion is that the normal should not be used if there is too much probability that a parameter that has to be positive could be simulated as negative from the normal.

Once a routine is in place to simulate parameters and to trend the h and u parameters, the number of deaths can be simulated from the negative binomial or Sichel distribution. If a routine to do this is not available, probably simulating from a transformed gamma with the same first three moments would not be too far off.

Model risk is a more difficult issue. The RH-Sichel model appears fairly reasonable for the male data, but the cohort parameters for the last several cohorts are questionable, being based on few observations. Parameter uncertainty would be large for such parameters. Perhaps using the models but including extra parameter uncertainty for model risk would give usable results.

## 7. SUMMARY AND FUTURE DIRECTIONS

The Lee-Carter model allows only highly constrained shifts in the shape of the mortality curve over time, and adding cohort effects gives much better fits. However these are found to generate new problems, such as potential over-fitting, instability for projections, and highly correlated and insignificant parameters. Also, the negative binomial fits better than the Poisson, which has been seen before and is likely to be a standard result. The best form for the NB is not consistent, however, and may differ for different datasets, depending on how contagion actually applies. For males, the Sichel distribution is better still.

Model risk is an issue, since the RH model can fit well at the ends of the age range using cohort parameters based on few observations. Using full cohorts can reduce this possibility at the older ages but not at the youngest ages. Also the RH parameters can be highly correlated, as in the female model, suggesting that some other model should be found, possibly by reducing the number of parameters.

Projections of mortality risk under current methodologies are thus likely to be unreliable. But better-fitting models are not likely to solve this problem as the RH model fits extremely will. Perhaps other models can be found with fits intermediate between LC and RH but with more parameter stability than RH.

## ADDENDUM

Now 2007 data is available, and some of the recent trends are continuing. Mortality for ages 65+ continued to improve compared to 2006, but for ages in the mid-50s, the lack of improvement continued. Whether this is just a random fluctuation or some underlying trend, such as obesity or reduced access to medical treatment, is yet to be established.

## REFERENCES


[1.] Beard, R. E. 1959. "Note on some mathematical mortality models." In: The Lifespan of Animals, G. E. W. Wolstenholme and M. O'Connor (eds.), Little, Brown, Boston, 302-311.

[2.] Burnham, K.P., and Anderson, D.R. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." Sociological Methods Research 33: 261–304.

[3.] Hannan, E., and Quinn, B. 1979. "The Determination of the Order of an Autoregression." Journal of the Royal Statistical Society B 41: 190–195.

[4.] Lee, R.D., and Carter, L.R. 1992. "Modeling and Forecasting U.S. Mortality." Journal of the American Statistical Association 87: 659–675.

[5.] Perks, W. 1932. "On some experiments in the graduation of mortality statistics." Journal of the Institute of Actuaries 63: 12-40.

[6.] 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.

[7.] Richards, S.J. 2008. "Applying Survival Models To Pensioner Mortality Data." Presented to the Institute of Actuaries, 25 February 2008, http://www.actuaries.org.uk/sites/all/files/documents/pdf/sm20080225.pdf.

[8.] Rigby, R.A., D.M. Stasinopoulos, and Akantziliotou, C. 2008. "A Framework for Modelling Overdispersed Count Data, Including the Poisson-Shifted Generalized Inverse Gaussian Distribution." Computational Statistics and Data Analysis 53: 381–393.

## APPENDIX 1. COUNT DISTRIBUTIONS

The negative binomial distribution has two parameters, $r$ and $\beta$, with mean $r\beta$ and variance $r\beta(1+\beta)$. In the full data there are 1855 cells, and when the negative binomial is used, each cell has a value of $r$ and $\beta$. The mean $\mu = r\beta$ is the value given by the RH model, but how $r$ and $\beta$ vary across cells depends on how the model is set up. In the NB1, it is assumed that every cell has the same value of $\beta$, so the ratio of variance to mean is $1+\beta$ for every cell. In the NB2, every cell is assumed to have the same value of $r$, with $\beta$ set to $\mu/r$, which gives variance to mean ratio $1+\mu/r$, which is higher for the cells with higher means. However there are many other ways the parameters can vary across cells. For instance, suppose there is a constant $q$ for all cells, with $r$ and $\beta$ given by $r = q\mu^{1/2}$ and $\beta = \mu^{1/2}/q$. Then the mean is still $r\beta = \mu$, and the variance to mean ratio for a cell is $1+ \mu^{1/2}/q$. This is what is called the NB3 in the text. Its variance/mean ratio is still higher for the larger cells, but not by as much as in the NB2.

This can be generalized to the NBp distribution, which adds a parameter $p$ to control the variance/mean ratio. It sets $r = q\mu^{1-p}$ and $\beta = \mu^p/q$. The mean is again $r\beta = \mu$, but now the variance to mean ratio for a cell is $1+ \mu^p/q$. The value of $p$ can be found by MLE. For males, the resulting value of $p$ is 0.2, but the NLL is not enough better to justify the additional parameter by any of the information criteria. For females, the $p$ is 0.53, but again this did not improve the NLL enough to justify the extra parameter. It might be argued that the NB3 already has an extra parameter of $p = ½$, but this is a bit ambiguous as the parameter is not free to be fit. In this case the NB3 fits the female data by enough better to justify an additional parameter.

When fitting a single NB distribution to a dataset, all of these forms are the same. The difference comes when fitting a number of distributions to a number of cells where a common relationship of variance and mean is desired. The NBp forms discussed here by no means exhaust the possible such relationships. In general, if the variance/mean ratio desired is $1+G(\mu)$, just set $r = \mu/G(\mu)$ and $\beta = G(\mu)$. For instance, $G(\mu) = q \log(\mu)$ might work in some cases, possibly even for the male data in this paper.

The Poisson—inverse Gaussian (PiG) distribution can be derived analogously to the NB as a Poisson mixture, but now the Poisson parameter is mixed by the inverse Gaussian instead of the gamma. Again it has 1, 2, 3 and p versions, etc. The inverse Gaussian is 50 percent more skewed than the gamma with the same mean and variance, and the PiG inherits this greater skewness, although not by the same ratio. The third central moment divided by the mean is the 3rd moment analogue of variance/mean for count distributions. For the negative binomial this is $1+3\beta+2\beta^2$, while for the PiG it is $1+3\beta+3\beta^2$. For $\beta=5$, which is fairly typical in the fits here, that gives 66 for the NB and 91 for the PiG, both of which would have variance/mean = 6.

The Sichel distribution is a generalization of the PiG and is a Poisson mixed by a generalized inverse Gaussian. It can be more or less skewed than the PiG but not less than the NB, which is a limiting case. It uses the modified Bessel function of the second kind (sometimes called the third kind), $K_\nu(t) = \frac{1}{2}\int_0^\infty x^{\nu-1} \exp\left[-\frac{1}{2}t(x+x^{-1})\right] dx$.

The Sichel distribution with parameters r, $\beta$ and $\nu$ can be most readily expressed with two auxiliary parameters c and s, with $c = K_\nu(r)/K_{\nu+1}(r)$ and $s^2 = 1+2\beta c$. The probability function at j is:

$$p_j = \frac{(r\beta c)^j}{s^{\nu+j} j!} \frac{K_{j+\nu}(rs)}{K_\nu(r)}$$

This is a reformulation of the version given in Rigby et al. (2008). Their parameters can be mapped from these by taking $\alpha = rs$, $\sigma = 1/r$, $\mu = r\beta$, and $c = 1/c$.

The PiG is just the case $\nu = -½$, for which $c = 1$. The Sichel mean is still $r\beta$ and the variance is $\mu(1+h)$ with $h = 2\beta c(\nu+1) + \mu(c^2 - 1)$. For the PiG, this simplifies to $h = \beta$. The Sichel ratio of the third central moment to the mean is $\mu_3/\mu = 1+ 2\mu(\beta c - h) + 3h + 2\beta ch(\nu+2)$.

The $\nu$ parameter can be very different than $-½$, and for the male data here was estimated as 2155. The $\beta$ parameter was close to 6, and r was set at $\mu/\beta$. The resulting third moments were usually intermediate to those of the PiG and NB.

## APPENDIX 2. FITTING NOTES

With several distributions to be fit, routines were sought that did not use derivatives of the NLL or could use numerical derivatives. The R package subplex uses an efficient form of the simplex algorithm, and was found useful in getting rapid improvement in the NLL from initial guesses. However it seemed to have difficulty in final convergence, often ending up in a region where the NLL was changing very slowly but was not near a minimum. Running subplex two or three times with default settings usually helped a good deal.

From there the optim routine in the Stats package was found to be useful in proceeding more toward a minimum. The optim option used most often was BFGS with gr=NULL, which takes fast approximate numerical derivatives of the NLL to find the best direction for improvement. Usually it would start off with only small improvements, but usually ended up finding a region where more rapid improvement was possible, then slowing down again near to convergence. Relative and absolute convergence criteria of 1e-17 and 1e-12 were used, which may be beyond machine precision. However the routine would converge, although usually not to a true minimum.

The next step was to define a gradient function of the parameters using numerical derivatives from the numDeriv package. This is a slower but more accurate gradient, and using BFGS with it always improved the fit. The problem is that the convergence is defined by the NLL not changing

much, which does not always end up with all derivatives very close to zero. Since the 2nd derivatives at the minimum are needed for the information matrix, it seemed a good idea to make the derivatives reasonably close to zero. For this the routine dfsane from the BB package was found helpful. In perhaps 50 iterations it could find points close to the optim parameters but with a reduction of 2 or 3 orders of magnitude in the largest (absolute) derivatives. It usually produced only very small changes in the NLL from what optim had yielded, however.

For the Bessel functions, the base R package function does not work with high values of the index (say $\nu > 1500$). There is a Bessel package available for Windows in R-Forge. It has a function besselK.nuAsym that does work for large values of the index, but not for small values. It needs an additional package Rmpfr, which is available on CRAN.

There are recursive formulas for the PiG and Sichel probabilities, but these are awkward at best for probabilities for tens of thousands of events.

The parameter constraints that force some parameters to be zero or one are different from much of the literature, which uses constraints on the sums of parameters. However doing it this way helps guarantee that the information matrix is not singular, which is necessary for its inversion.

(Yilu Zhang and Lina Ma helped research the R methodology for fitting distributions used here.)