Stochastic Trend Models in Casualty and Life Insurance

Spencer M. Gluck, FCAS, MAAA
Gary G. Venter, FCAS, ASA, CERA, MAAA

Abstract

Quantifying liability risk for insurance companies requires projections of distributions of future contingent costs. Past patterns are typically forecast to continue but with the possibility of deviations to various extents. Trends into future calendar years for existing populations of people or events involve inherent uncertainties, but the forecasts are needed in both casualty and life insurance, and the understanding/quantification of those uncertainties is essential to risk analysis and management. With an emphasis on the somewhat analogous issues of calendar-year payment trend in casualty and mortality trend in life, we review existing models, provide some extensions and discuss some pitfalls. In particular, we find that unrecognized time series behavior in casualty payment trends can significantly contribute to reserve risk, and provide techniques for identifying and measuring it. In life, time series in mortality trend is typically modeled, but we find that the fitting of historical calendar-year trends creates autoregressive effects in the estimated trends that may not exist in the underlying trends.

Keywords:

Loss Reserves, Mortality, Trend, Auto Regressive Models, Time Series.

1. Introduction

Quantifying liability risks involves projections of the costs of existing commitments into the future. One component of such projections is quantifying trends and variability of trends. In casualty insurance, the cost level is subject to stochastic trends, whereas in life and annuity business, the stochastic trends for mortality rates are most influential. We review some of the models used to quantify risks in both branches, note some areas where improvements in standard methods are needed, and in some cases suggest alternative approaches.

One factor often overlooked in both branches is the degrees of freedom lost in the model-building process. If you think of the data as using up degrees of freedom to pull the model towards itself, as suggested by the work of Ye (1998), standard methods calculate the degrees of freedom used up by the parameterization process. However the model-selection process is another avenue the data has to pull the model towards it. For instance, parameters can trend over time with trend breaks inserted by the modeler when needed to better fit the data. We introduce a parameter-smoothing filter based on generalized weighted averages to automate this process, which eliminates the subjective selection of trend breaks and allows calculation of the degrees of freedom used up in this aspect of model building in addition to those used up in fitting.

In many casualty models, trends occurring in calendar years (i.e., on the diagonal of the triangle) are not explicitly identified, and are forecast only implicitly. Even when explicitly modeled, the possibility of time series behavior in the trends has rarely been discussed. On the other hand, mortality trends in life models are sometimes fit by an ARIMA(1,1,0) process—that is, an AR(1) process is fit to first differences of calendar-year levels. We show that lag 1 autocorrelation is induced in first differences of estimated parameters by the estimation process even when it does not exist in the actual mortality levels.

Section 2 reviews existing models. Section 3 introduces possible improvements in casualty liability stochastic trend models. Section 4 briefly does the same for mortality models. Section 5 introduces the general weighted average (GWA) filter. Section 6 illustrates an application of the GWA filter to casualty calendar-year payment trends and discusses the implications for reserve risk analysis. Section 7 discusses induced autocorrelation and approaches to dealing with it. Section 8 concludes.

2. Background on Existing Models

In casualty reserves, implicit calendar-year trends are built into most models through a stretching out of the development pattern arising out of historical trends. The risk of changes in future trend can be superimposed on the development risk from any model as well. However, the explicit modeling of historical trends within the data comes out of the multiplicative fixed-effects models (row factor times column factor) dating back at least to Bailey and Simon (1960) for ratemaking and Hachemeister and Stanard (1975) in reserving.

In life insurance, the Lee-Carter 1992 model is a multiplicative fixed-effects model for a transform of the mortality rates. Verbeek (1972) and Taylor (1977) extended the casualty model to include row, column and diagonal effects, and Renshaw and Haberman (2006) do this for mortality models, although not as simple factors. Barnett and Zehnwirth (2000) formulate the Taylor model as trends instead of levels, which allows for reducing the number of parameters. They work in paid incremental losses, which they model as lognormal, based on an argument of the accumulation of multiplicative random effects. We also look at models with the residual variance proportional to the mean, which despite the arguments for lognormal, seem to fit actual data better in many cases.

Multiplicative Fixed-Effects Model (MFE)

The general framework of cross-classified models postulates that the expected value of a cell in an array is the product of a row and a column effect. Some notation is needed to discuss this.

The n+1 columns of a triangle are numbered 0, 1 , … n and denoted by the subscript d. The rows are also numbered from 0 and denoted by w. The last observation in each row of a full triangle will then have w+d=n. The cumulative losses in cell w,d are denoted $c_{w,d}$ and the incremental losses by $q_{w,d}$.

For the MFE model, $E[q_{w,d}]$ is $U_w g_d$, where U_w and g_d are the row and column parameters, respectively. Note that increasing each g by the same factor and dividing each U by that factor does not change the mean for any cell. To have specificity, it is often convenient to have the g's sum to 1. Then U_w can be interpreted as the ultimate loss for year w and g_d the fraction that are at lag d.

Assuming that the distribution around the cell mean is lognormal, each cell's observation is $\log [q_{w,d}] = \log U_w + \log g_d + \varepsilon_{w,d}$, which is a linear model with a normal error term, and so estimable by regression. This was already studied by Kremer (1982). On the other hand, if the distribution is normal, so $q_{w,d} = U_w g_d + \varepsilon_{w,d}$, the model is non-linear. Mack (1991) linked this model of development triangles to MFE models in classification ratemaking, such as those in Bailey (1963), Bailey-Simon (1960), etc. These models can be estimated by a generalization of fixed-point iteration called Jacobi iteration, using $g_d = \sum_{w=0}^{n-d} U_w q_{w,d} \Big/ \sum_{w=0}^{n-d} U_w^2$ and $U_w = \sum_{d=0}^{n-w} g_d q_{w,d} \Big/ \sum_{d=0}^{n-w} g_d^2$. This is just the result of alternatively treating the g's and the U's as known constants, so the model temporarily becomes a simple factor model in the other parameter.

Poisson Distributed Losses with Fixed Severity

In the life case, it is natural to assume that deaths follow a Poisson distribution. In the casualty case, a convenient starting point for multiplicative fixed-effects models is to assume the error terms follow the Poisson – constant severity (PCS) distribution. This is the aggregate loss distribution consisting of a Poisson frequency and a constant severity. In this context it assumes all claims or payments in all cells are the same size, call it *b*. This of course is rarely the case, but the model has some advantages. First, it is a distribution of aggregate claims, which most triangles consist of. However its historical appeal is that an MFE model estimated by MLE gives the same reserve estimate as the chain ladder.

In the pure Poisson case, the agreement of methods was shown by Hachemeister and Stanard (1975) although that finding was not published formally until Kremer (1985) in German (translated into Russian as well) and Mack (1991) in English. Renshaw and Verrall (1998) extend this to the over-dispersed Poisson, which in generalized linear model terminology is defined as any member of the exponential family whose variance is proportional to its mean. However the only distribution meeting this criterion is the PCS.

Giving the same answer as the chain ladder is not a particularly useful criterion for evaluating models, but it starts from a familiar base. Thus the $q_{w,d}$ will be assumed approximately PCS distributed for MFE models here.

For the PCS model, a cell with frequency λ has mean $b\lambda$ and variance $b^2\lambda$. For the MFE implementation then $b\lambda_{w,d} = U_w g_d$.

This model is applied here to incremental losses, so that the observation $q_{w,d}/b$ is Poisson with mean $U_w g_d/b$. The loglikelihood function[1] can be shown to be:

$l = C + \sum \left(\frac{q_{w,d}}{b} \ln \frac{U_w g_d}{b} - \frac{U_w g_d}{b} \right)$, where $C = -\sum \ln \Gamma(1 + q_{w,d}/b) \equiv -\sum \ln [(q_{w,d}/b)!]$. Taking derivatives, the MLE estimates can be expressed as: $g_d = \sum_{w=0}^{n-d} q_{w,d} \Big/ \sum_{w=0}^{n-d} U_w$ and $U_w = \sum_{d=0}^{n-w} q_{w,d} \Big/ \sum_{d=0}^{n-w} g_d$, which do not depend on b. The difference in this iteration from that of Bailey-Simon is due to using the Poisson assumption instead of the normal. Technically, the Poisson

[1] Note that this requires not fitting just one Poisson distribution but (n/2 +1)(n+1) of them, defined by 2n+1 row-column parameters plus *b*. But MLE applies to fitting multiple distributions with the same parameters.

probabilities are zero unless $q_{w,d}$ is an integral multiple of *b*. However Mack (2002), Chapter 1.3.7, shows that there is a continuous analogue of the Poisson that can be scaled by *b* and gives estimates close to the PCS. When the PCS is applied in a continuous setting it can be thought of as using this distribution.

The MLE formulas can be solved by iteration, starting with some values then solving alternatively for the g's and U's until the results converge. If the resulting g's do not sum to 1, it is easy to just divide each by their sum and multiply each U by the same sum. Starting at the upper right corner of the triangle and working back can show that these estimates correspond to the chain-ladder calculation. Essentially the U's are the last diagonal grossed up to ultimate by the development factors and the g's are the factors converted to a distribution of ultimate. The fitted incrementals are then the g's applied to the U's, and can be calculated by using the development factors to back cumulatives down from the last diagonal.

Modeling Calendar-Year (Diagonal) Effects

It is fairly common that triangles are not adequately modeled by the MFE model including only row and column factors, as evidenced by residual unmodeled diagonal effects. There are a number of possible underlying phenomena. Claims inflation (economic or social) may operate on the calendar year of payment. Inflation operating on accident year is in the *U* parameters, but if inflation that varies by calendar year is reflected only in the accident year parameters, high and low residuals can show up by diagonal. Diagonal effects can also be a result of accelerated or stalled claim department activity in a calendar year that would often be made up for in a later year or years, so more than one diagonal is affected. Gradual lengthening or shortening of the payment pattern may manifest as changing calendar-year trends. Significant patterns in residuals among diagonals would suggest that calendar-year phenomena are operating. Modeling them can provide better fits. Failing to account for such effects when they are present can lead to misestimating row and column parameters. Failing to account for the risk associated with future calendar-year effects can lead to unrealistically narrow forecast distributions.

Taylor (1977), following Verbeek (1972) discusses a method for estimating calendar-year effects, which he calls the separation method. For some decades after that, models of calendar-year effects were informally called separation models, even when they did not use that technique.

In the lognormal MFE model given by $q_{w,d} = U_w g_d h_{w+d}(1+\eta_{w,d})$, taking logs gives $\log q_{w,d} = \log U_w + \log g_d + \log h_{w+d} + \varepsilon_{w,d}$, which is a linear multiple regression model.

Barnett and Zehnwirth (2000) set up a model framework of this type, but they change levels to trends, which facilitates parameter reduction. They denote $\log U_w$ by α_w and express $\log g_d = \sum_{k=1}^{d} \gamma_k$ and $\log h_{w+d} = \sum_{t=1}^{w+d} \iota_t$. This makes $\gamma_d = \log[g_d/g_{d-1}]$, for instance. Thus it makes sense to call γ a trend in the direction of later lags. If the lag parameters g are trending upwards or downwards by an exponential curve for several columns, the same γ can be used for those columns, reducing the number of parameters in the model. Similarly the ι's are trends over calendar years and may be constant for a few years, reducing the number of calendar-year trend parameters.

Barnett and Zehnwirth emphasize that this is not a model, but a model development framework. It would never make sense to estimate a parameter for each accident year, each development year and each diagonal, as those effects are overlapping. The art of building a model in this framework is to find the combination of effects that best explain the data while maintaining statistical standards of parsimony of parameters. However as noted, such a model building process itself adds implicit parameters.

Improving Parsimony

The MFE model and its deterministic analog, the chain-ladder method, have sometimes been criticized as being over-parameterized. The problem becomes most evident in the estimation of the U (row) parameters for immature accident years and the g (column) parameters toward the tail of the triangle. The potential weakness of the chain-ladder method in this regard is often carefully attended to by practicing actuaries.

Immature accident years often receive a great deal of attention. Within the deterministic approaches, the Bornhuetter-Ferguson method (1972) is a fairly general solution of stabilizing chain-ladder estimates with external information, but the source of the external information is unspecified. Stanard (1985) and Buhlmann (1983) each independently proposed a more specific solution, in which the accident year levels U are effectively merged into a single parameter, assuming that the accident years are first adjusted to a common level (colloquially, the "Cape Cod" method). Potentially unstable development factors toward the tail may be addressed with ad hoc curve fitting (e.g., Sherman (1984)), or simply careful examination with judgmental smoothing as necessary.

The Barnett/Zehnwirth (2000) framework allows for improved parsimony in the accident year direction by allowing the U parameters to be compressed to a single parameter or to a reduced number of parameters, each applying to a contiguous block of rows. As previously discussed, trends in the columns and diagonals can likewise be modeled for blocks of columns or diagonals, creating continuous piecewise exponential models for those effects. In each case, the "break points" defining such blocks are user-specified as a part of the fitting process.

As an alternative to the accident year blocks, de Jong and Zehnwirth (1983) introduced the Kalman filter to the actuarial world. Filtering creates levels for the rows that are partially codependent. The effect is row levels that do not change sharply, but that may drift as necessary to fit the data. There may be substantial advantages: improved fit while maintaining parsimony, a more elegant and plausible model of changing levels and a reduction in the need for arbitrary modeling judgments that create implicit, unrecognized parameters. Gluck (1996) introduced the "Generalized Cape Cod" method, providing a similar accident year smoothing effect within deterministic methods.

Within the existing Barnett/Zehnwirth specification, filtering is done only for the accident year levels.

3. Casualty Stochastic Trend Models

We now describe our approach to modeling the casualty triangle. The natural data for the model is typically a triangle of incremental loss payments, along with a corresponding vector of relative exposures. Our starting point is a trend parameter structure based on the work of Barnett and Zehnwirth (B/Z, 2000).

We reiterate previously introduced notation before we use it:

Data

$q_{w,d}$ The data triangle, typically incremental loss payments, for accident year w and lag (or development year) d; $w = 0 \ldots n$; $d = 0 \ldots n - w$

The calendar year of payment is $k = w + d$

E_w Exposures for accident year w.

The model will be fitted to the losses per exposure, i.e. $y_{w,d} = q_{w,d} / E_w$

Parameter Structure

Each type of parameter is displayed with a subscript, w, d or k for the accident year (AY), development year (DY) and calendar year (CY) directions, respectively. For the moment, the model is presented as if each subscripted parameter were different for each value of the subscript, although this would never be the case in practice. In fact, the model as stated is over-determined and would require an additional constraint. We will return to these matters in the subsection "Parameter Reduction."

For the parameter types, we adopt the B/Z notation. The $\boldsymbol{\alpha_w} = \ln(U_w)$ are AY levels, while the $\boldsymbol{\gamma_d}$ and $\boldsymbol{\iota_k}$ are trends, in the DY and CY directions, respectively, also on log scale. The model is then:

$$y_{w,d} = \exp\left(\alpha_w + \sum_{i=1}^{d}\gamma_i + \sum_{j=1}^{w+d}\iota_j\right) + e_{w,d}$$

Probabilistic Structure

The error terms $\boldsymbol{e_{w,d}}$, and therefore the incremental payments themselves, are assumed to be stochastically independent.

In the form of the distributions, we now depart from B/Z. Whereas they use a log-linear model with lognormal multiplicative errors, we instead assume that the incremental payments $\boldsymbol{q_{w,d}}$ are distributed PCE ("Poisson Constant Severity," also known as the over-dispersed Poisson or ODP). We have several reasons for our preference;

- The most important distinction is in the variance assumption. With the PCS, variance is presumed to be proportional to the mean – the natural behavior of sample sums. In particular, as we move toward the tail of the triangle with small means, the CVs increase and the distributions become skewed.

- In contrast, the starting assumption of the log-linear form of the model is constant CV. We have found this to be typically unrealistic in triangle data, creating significant residual heteroscedasticity (in the DY direction) which must be corrected using an additional heteroscedasticity model.

- Where expected values are small, in the tail (and occasionally the "nose"), actual values of zero are not unexpected nor are they uncommon. The PCS reacts as intended, while the lognormal distribution treats zero as impossible and values near zero as extreme negative outliers.

Recalling that the model is expressed in terms of $y_{w,d}$ rather than $q_{w,d}$, the variance of $y_{w,d}$ is also assumed to be inversely proportional to the exposure E_w. Thus

$$\mathrm{var}(y_{w,d}) = \sigma^2 * E(y_{w,d}) / E_w \quad ,$$

where σ^2 is a constant scale factor.

Estimation

The model can be estimated as a weighted Generalized Linear Model (GLM), using the E_w's as weights. The fitting procedure returns maximum likelihood estimates of the parameters under the distributional assumptions as previously described. The model easily converts to the formulation required for GLM, a combination of a link function and a linear predictor. In this case, the link function is the logarithm, ln:

$$y_{w,d} = \ln^{-1}(\eta_{w,d}) + e_{w,d}$$

and the linear predictor is:

$$\eta_{w,d} = \alpha_w + \sum_{i=1}^{d} \gamma_i + \sum_{j=1}^{w+d} \iota_j$$

In fact, the parameters can also be estimated without using GLM. As briefly discussed in Section 2, the MLE solution for each of the three parameter types individually is quite simple, and the overall MLE solution can be arrived at by iteratively solving for each of the three types in succession.

Given parameter estimates, we have fitted values $\hat{y}_{w,d}$ at each point of the historical triangle and forecast, and estimate the corresponding variances as follows:

$$\text{var}(y_{w,d}) = \sigma^2 * E(y_{w,d}) / E_w \approx \hat{\sigma}^2 * \hat{y}_{w,d} / E_w$$

Note that variance is proportional to the mean and inversely proportional to the weights. For the estimated variances, $E(y_{w,d})$ is approximated by the fitted value $\hat{y}_{w,d}$, and the constant scale factor σ^2 is estimated as follows:

$$\hat{\sigma}^2 = \sum_{w,d} \left((y_{w,d} - \hat{y}_{w,d})^2 * \frac{E_w}{\hat{y}_{w,d}} * \left(1 \Big/ 1 - \frac{\partial \hat{y}_{w,d}}{\partial y_{w,d}} \right) \right)$$

The last multiplier in the above equation is a correction to reverse the bias in the squared residuals caused by degree of freedom restrictions. The sum of the partial derivatives is the effective number of parameters from the generalized degrees of freedom approach of Ye (1998).

Parameter Reduction

For each of the three parameter types, we may apply one of two methods to reduce the number of parameters. The first is to allow the parameters to be identical for contiguous blocks of years in any direction by selecting “break points” at which the parameter value changes. The second is to apply filtering (parameter smoothing), which allows the values to vary from year to year, but reduces the effective number of parameters by creating time-related dependency among the values. The filtering algorithm is described in a subsequent section.

The innovation versus B/Z is the availability of filtering in the CY and DY directions. In both cases, the filtering makes for more convenient model fitting, and eliminates the need for judgmentally selecting break points.

Filtering the CY trend has significant additional implications. We can now consider CY trend as continuously changing, a time series, rather than constant for periods of time with occasional sharp turns as necessary to fit the history. Observing the process of changing trends in the past can provide the basis for a model of potentially changing trends in the future, essential to an understanding of risk in a dynamic environment.

Filtering the DY trend also provides convenience and the elimination of break points, but there are challenges. The DY trend filter will not be used herein.

When the individual AY levels are modeled as independent or with the filter, the AY levels can have sufficient freedom to exhibit an overall trend. Combined with trends in the DY and CY directions, there is one too many free trends for a two-dimensional array, and an additional constraint is required. We have typically chosen to constrain the AY levels to have an overall trend of zero.

Forecasting the CY Trend

The reserving problem does not require forecasts of AY levels beyond the data. Extrapolating a tail may involve forecast DY trend, but we will not address that issue here. But by definition, all forecast payments involve a forecast of future CY trend. Explicit recognition of CY trend is absent from the most common reserve methodologies in the casualty area, and many actuaries are uncomfortable with the need to insert a judgment that has a substantial impact on the forecast. Nevertheless, the future is forecast in all methodologies, and the ability to observe trends in the historical data allows for an informed forecast. The CY forecast that is implicit in the chain-ladder may in fact be reasonable, but cannot be presumed reasonable without examination. As will be illustrated, CY filtering can provide additional information to inform the CY trend forecast.

Predictive Distributions

While detailed description of techniques for forecast distributions is not within the scope of this paper, the need to provide information about forecast distributions has been a driver of the development of probabilistic reserve models. The model structure, together with the fitted parameters and residuals, provide an estimated probabilistic structure for the loss process. Predictive distributions are intended to represent the distribution of potential outcomes, including variability due to random process risk occurring in the future, and potential inaccuracy of the forecast due to parameter estimation error. Process risk is typically measured via simulation. The two most common approaches to parameter error are closed form solutions based on the estimated parameter variance-covariance matrix and bootstrapping.

England and Verrall (2001) described an application of bootstrapping using the previously described MFE model. Relying on the finding that chain-ladder estimation reproduces MLE estimates (under appropriate conditions), they demonstrated how such an analysis can be performed without advanced statistical software, and the approach has achieved fairly widespread popularity. We have applied a similar approach to bootstrapping, but with parametric re-simulation of the data rather than classic re-sampling.

Neither approach provides a measure of the uncertainty resulting from the imperfection of the model structure itself (specification error), which may be considerable. In particular, model selection judgments that are made by examination of the data are in effect parameters embedded in the model structure, the aforementioned "break points" being a case in point.

We return to the uncertainty regarding the future CY trend. The possibility of changes in CY trend is a process risk, but unlike the typically modeled process risk, it is non-diversifying. Without a provision for such risk, casualty loss reserve models are static, reflecting only the risk that would exist in a world that never changes. Since nearly all well-known models are static in this regard, an additional provision for this risk would create a more realistic risk assessment. A likely approach is to superimpose a mean zero time series on the forecast distribution of any such model. The CY trend filter can provide a basis for selecting/parameterizing the future CY trend time series.

ILLUSTRATION: U.S. Industry Workers' Compensation

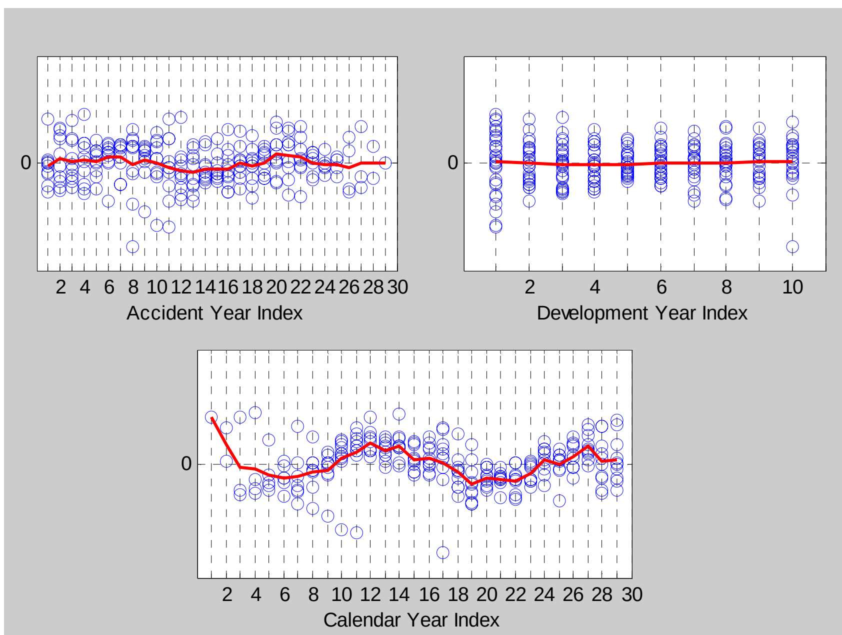
We present sample output based on an analysis of industry paid loss data for workers' compensation derived from published Schedule P information. Note that it is not based directly on published industry aggregates, but is the result of fairly extensive work to eliminate errors and distortions. The paid loss triangle is in fact a trapezoid, having only 10 development years, but 29 accident years, from 1979 to 2007. The exposure measure is actual net premiums, which we do not treat as a reliable exposure measure.

The first illustrative model allows independent AY levels, independent DY trends and no CY trends. This is the MFE model for which the parameter estimates can be derived by the chain-ladder. For relative overall fit, we record two parameter-penalized measures, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We also record the effective number of parameters as previously defined.

AIC: 6,244 BIC: 6,377 # Parameters: 38

To test the validity of the fitted model, we examine residual plots. If the model fits in accordance with its probabilistic assumptions, the "deviance residuals" will be approximately independently, identically, normally distributed. The single most important criterion for the residuals is randomness—the lack of any identifiable unmodeled pattern. We plot the residuals against each of the three directions of the triangle. The mean residual in each column is also plotted to aid the eye in the search for unmodeled patterns.

Figure 3.1.
3-way Deviance Residuals – MFE

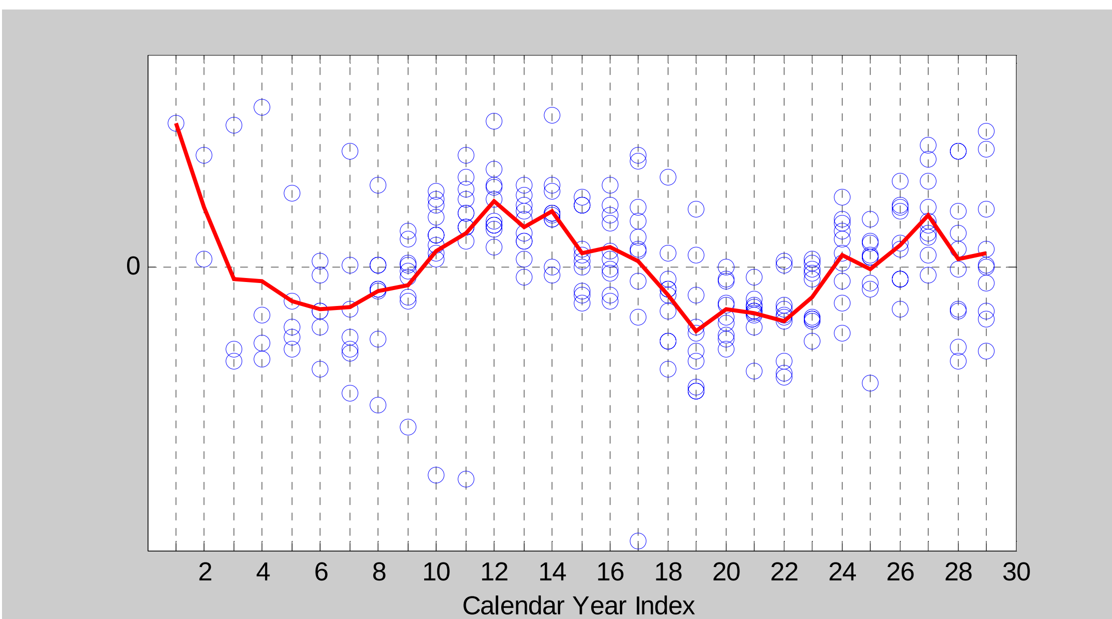


With a separate parameter for every AY and every DY, it is unsurprising that the model fits quite closely, on average, in each of those directions. However, the CY plot is clearly not random.

Given the lack of a useful exposure base, parameter reduction in the AY direction is not available. In this case, the volume and stability of the database makes this less of a problem than is typically the case in individual company data. In the DY direction, we have only nine trends and an unusually large number of observations for each, and so again, the common over-parameterization issues are not relevant, and we will not illustrate parameter reduction here either.

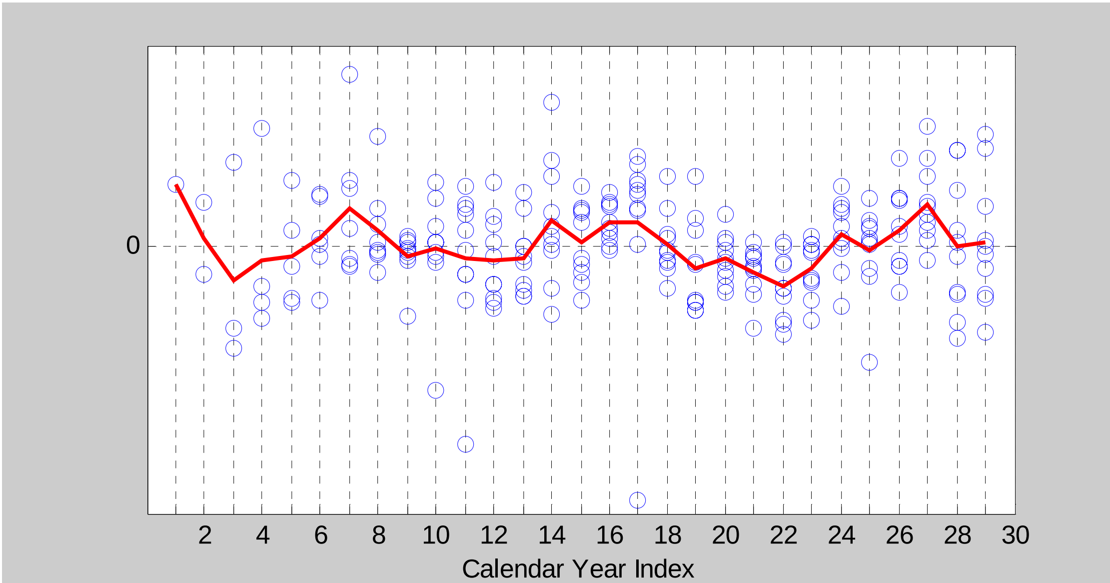
The CY direction clearly needs attending to. In most applications, we would add CY parameters but significantly reduce parameters in other directions, ending with a lower parameter count than the chain-ladder, but in this case, we are only adding parameters. We first use the break-point approach. Revisiting the CY residual plot, we select break points that would allow us to fit the residuals as a series of connected lines. Break points are selected at CYs 7, 12 and 19. We will also consider whether a break should be added at CY27 (i.e., 2005).

Figure 3.2.
CY Deviance Residuals – MFE



The new CY residual plot indicates that the CY trends fit the data better:

Figure 3.3.
CY Deviance Residuals – Trend Breaks (7, 12, 19)



The new statistics are as follows:

AIC: 6,157 BIC: 6,301 # Parameters: 41

Although we have three additional parameters, both the AIC and BIC have decreased, indicating that the improvement in fit is worth the extra parameters according to either measure. Note that the BIC will always have a larger parameter penalty than the AIC, but there is no consensus as to which measure is superior.

In Table 1 below, we display the different fitted CY trends, their corresponding time periods and standard errors. The "Original" is as described above, while the "Alternative" has an additional parameter break at CY27 (2005). Note that when evaluating whether breaks should be added, the statistical significance of a potential break can be tested, in addition to the AIC/BIC statistics and scatter plots.

Given the history of sharp turns as modeled, the best picture for future trend may not be easy to see, but the potential trend volatility must be considered when we evaluate forecast risk. We will return to this example in Section 6, applying filtering to the CY trends.

TABLE 1
Trend Parameters and Standard Errors

Original			**Alternative**		
Calendar Period	**Annual Trend**	**Std. Error**	**Calendar Period**	**Trend**	**Std. Error**
1979:1985	- 4.0%	0.7%	1979:1985	- 4.1%	0.7%
1985:1990	+3.8%	0.6%	1985:1990	+3.8%	0.6%
1990:1997	- 3.4%	0.4%	1990:1997	- 3.5%	0.4%
1997:2007	- 0.3%	0.3%	1997:2005	+0.3%	0.4%
			2005:2007	- 4.6%	1.4%

4. Mortality Stochastic Trend Models

The standard starting point for modeling mortality is now the Lee-Carter (1992) model. 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 death 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. Let a_d be the average over column d of log $m_{w,d}$. Suppressing the error term, the Lee-Carter model in this notation can be expressed as: $x_{w,d} = \log m_{w,d} - a_d = b_d h_{w+d}$. This is an MFE model by age at death d and calendar year of death w+d. Renshaw and Haberman (2006) extend this for what is called cohort, or year-of-birth, effects: $\log m_{w,d} = a_d + b_d h_{w+d} + c_d u_w$. Here a_d is a parameter to be estimated for each age of death, but is usually close to the average over w of the log $m_{w,d}$.

Typically the deaths in a cell are modeled as Poisson in $m_{w,d}E_{w,d}$. However Venter (2008) and others have found that Poisson is often too light-tailed. Variance does appear to be proportional to mean, as the Poisson requires, but too many residuals are too far from the mean to make Poisson plausible. Perhaps a negative binomial would be more appropriate. This is in line with contagion effects in mortality data. Environmental effects like war, weather and disease can cause peaks and valleys in the death rates.

The calendar-year mortality level in these models is given by the h parameters. Venter (2008) and others have found that an AR(1) model fits the differences in parameters well, but as discussed below, this could be an artifact of the fitting process. If the AR(1) model is considered as part of the overall mortality model, h becomes a random effect instead of a fixed effect. For computational tractability, however, most analysts estimate the AR(1) model as a separate step after fitting the h parameters to the data. Lee and Carter actually use a random walk model for the h parameters. Other time-series models have been tried as well.

The death rates $m_{w,d}$ are linked to the mortality rates $q_{w,d}$ by $m_{w,d} = -\log(1 - q_{w,d})$. Cairns etal. (2007) compare several mortality models, and find that modeling $\text{logit}(q) = \log(q) - \log(1 - q) = \log(1 - e^{-m}) + m$ is also a successful approach. In particular, they look at the model

$\text{logit}(q_{w,d}) = g_{w+d} + h_{w+d}d_1 + k_{w+d}d_2 + u_w$, where d_1 is d less the average d in the study, and d_2 is d_1^2 less the variance of the d's in the study. They find this model provides about as good a fit as Renshaw-Haberman, but its parameters are more consistent when fit to different subsets of the data, which is also a motivation for using d_1 and d_2 instead of letting the moments be absorbed in the g parameter. However now trends are needed for the g, h and k parameters.

5. General Weighted Average (GWA) Filter

We formulate the problem as follows:

1. We have a time-ordered column vector of observations $B = b_1, \ldots, b_n$. In the specific application, these observations are parameter estimates.

2. B is an unbiased estimate of the unobserved true parameters $D = d_1, \ldots, d_n$. Thus, $b_j = d_j + y_j$, with $E(y_j) = 0$.

3. The covariance matrix for the y's is C with i, j element $C_{i,j} = E(y_i y_j)$.

4. We postulate that the values of the vector D are neither identical nor fully independent, but rather exhibit time-related dependence. The essential property of that dependence is that it is stronger for points close in time and weaker for points distant in time. We capture that property by postulating the simplest possible time series, a stationary (mean zero) random walk. Thus, $d_{j+1} = d_j + x_j$ with $E(x_j) = 0$, and $E(x_j^2) = s_j$. The x's are mutually independent and independent of the y's.

The filter is a smoothing mechanism to help us detect the signal, D, through the noise, C. The random walk that is postulated in step 4 above is not necessarily the best description of the time series process that may exist in D; rather, it is a means for the filter to reflect time-related dependence while imposing very little of its own structure on the outcome. Once we can see D more clearly, we can further explore its true nature. We will return to this point in the example.

The s_j's may be referred to as adaptive variances, and we will tend to use a single value for all of them. The degree of smoothing will be controlled by the relationship between the s_j's and the variances in C. At the extremes, $s_j = 0$ makes the values a single parameter, while the values approach separate, independent parameters as s_j becomes large.

Defining the General Weighted Average

An average is the simplest case of least squares regression, in which there is a single parameter, the constant. Thus, if there are n observations, the design matrix for the average is simply an n x 1 column vector of ones. Call it A.

Analogously, a weighted average is weighted least squares regression with this same design matrix. Weighted regression is appropriate when the observations are independent but not equally reliable. Typically, the weights are inversely proportional to the variances of the observations, which results in a minimum variance estimate for the parameters (in this case the average). Credibility weighting is actually inverse variance weighting.

Finally, the GWA is the simplest case of general least squares regression. In the general case, the observations are no longer presumed to be independent. Instead, the variance-covariance matrix of the observations, V (n x n), is explicitly accounted for. For the GWA, the design matrix A is as specified above, and the weights for the average is the 1 x n vector W:

$$W = (A'V^{-1}A)^{-1}A'V^{-1}.$$

Equivalently, the vector, $A'V^{-1}$ may be directly normalized by dividing by its sum. If the observations are the vector B, the GWA is simply WB. Note that $A'V^{-1}$ expresses the weights as a reciprocal variance. This is analogous to credibility theory, where the credibility $P/(P+s^2/t^2)$ results from normalizing the inverse variances P/s^2 and $1/t^2$ to sum to 1 by dividing by their sum.

It is instructive to consider the GWA of the original parameter estimates, B, weighted by the estimated covariance matrix, C. If B and C have been estimated by least squares regression, then the GWA of B is identical to the least squares estimate of a single parameter, b, replacing the multiple parameters, b_i .[2]

Completing the Filter

We require estimates of the n elements of D. The estimate of d_k will be a GWA of the vector B, where the relevant covariance matrix V_k reflects the errors associated with using each of the b_j's as estimates of d_k. Thus the i, j element of V_k is $v_{k;i,j} = E[(b_i - d_k)(b_j - d_k)]$

$$= E[(b_i - d_i + d_i - d_k)(b_j - d_j + d_j - d_k)] = E[(y_i + d_i - d_k)(y_j + d_j - d_k)]$$

$$= E[\, y_i y_j + y_i\,(d_j - d_k) + y_j\,(d_i - d_k) + (d_i - d_k)(\, d_j - d_k)].$$

For j > k, $d_j - d_k = \sum_{l=k}^{j-1} x_l$, and for j < k, $d_j - d_k = -\sum_{l=j}^{k-1} x_l$.

$v_{k;i,j} = C_{i,j} + E[\,(d_i - d_k)(\, d_j - d_k)]$, due to the independence of the y's and x's.

There are now three cases:

- If k is between i and j, inclusively, then $v_{k,i,j} = C_{i,j}$
- If i and j are both greater than k, then $v_{k,i,j} = C_{i,j} + s_k + \ldots + s_{\min(i,j)-1}$
- If i and j are both less than k, then $v_{k,i,j} = C_{i,j} + s_{\max(i,j)} + \ldots + s_{k-1}$

all following from the mutual independence of the x's.

---

[2] The statement does not necessarily hold when B and C are estimated using other means, such as MLE. In the examples of this paper, it is very nearly accurate.

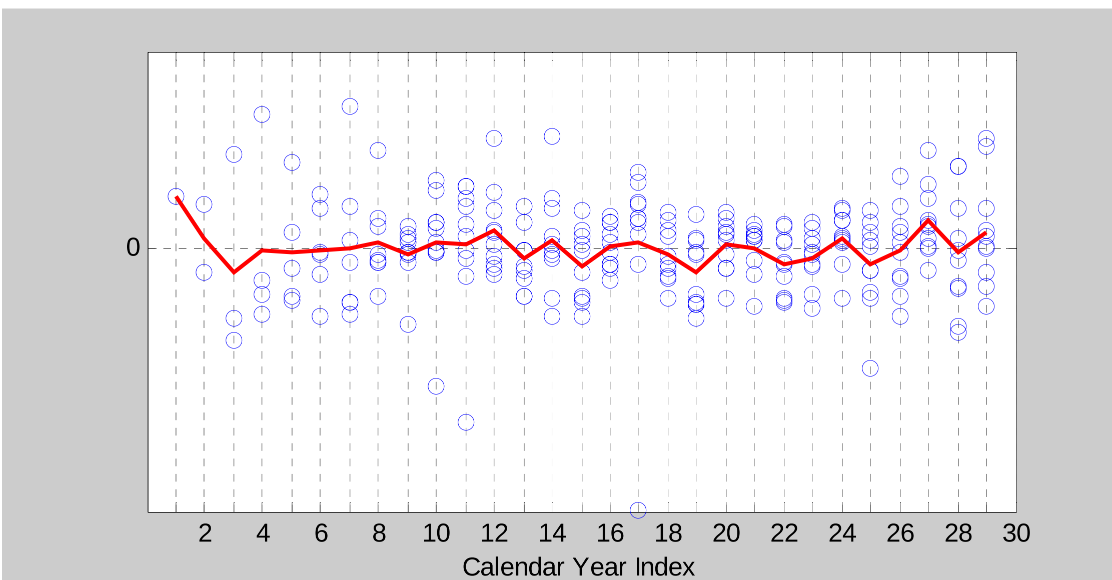
The adaptive variance(s), s is often selected judgmentally. Several techniques have been suggested for estimating s from the data, by comparing the observed relationships among the estimates B with the relationships that would be predicted by C alone. Another alternative is to find the value based on the goodness of fit of the smoothed parameters to the data, using a penalized likelihood function to test goodness of fit, where the number of parameters comes from the generalized degrees of freedom approach of Ye (1998).

The above formulation applies for a single direction of the triangle or parameter type. To filter multiple parameter types, we apply the GWA filter recursively.

6. Application of the GWA Filter: Casualty Example

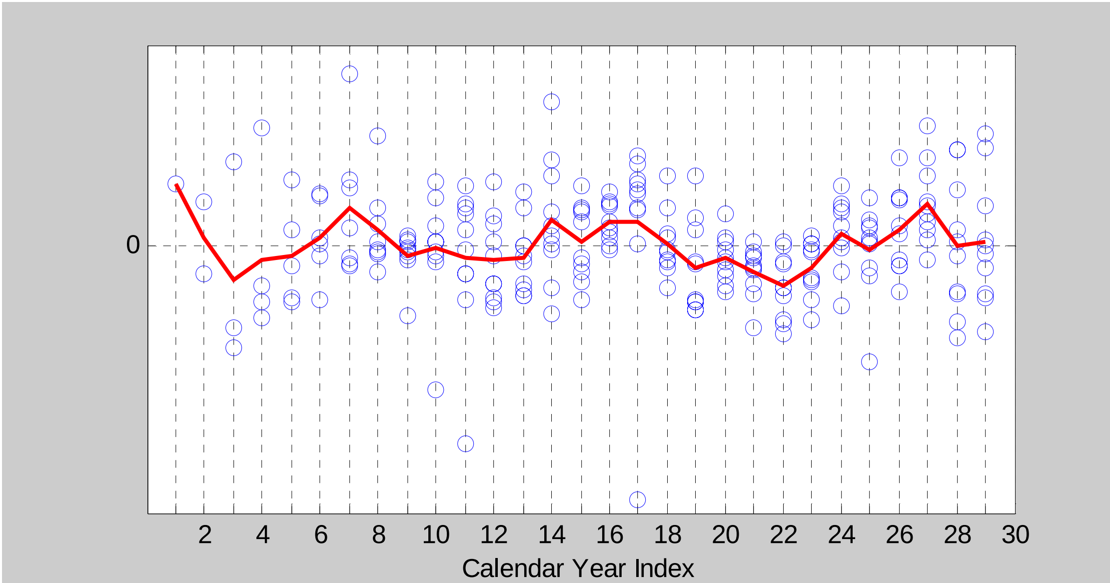
We now apply the GWA filter to CY trends in the industry workers' compensation database. Other aspects of the model have remained unchanged: separate parameters for every AY level and every DY trend. We describe the model as having "dynamic trends," meaning continuously changing. We have transformed the adaptive variance s to a scale ranging from zero to one, indicative of the degree of smoothing (smaller values create more smoothing). We first examine the CY residual plot for the dynamic trend model, with smoothing parameter 0.4.

Figure 6.1.
CY Deviance Residuals – Dynamic Trend (0.4)



To the eye, this seems to be a significant improvement versus the previous 3 – break model, which we repeat for comparison.

Figure 3.3.
CY Deviance Residuals – Trend Breaks (7, 12, 19)



We next compare the fit statistics and parameter count.

Dynamic Trend (0.4):	AIC: 6,127	BIC: 6,284	# Parameters: 44.7
Trend Breaks (7, 12, 19):	AIC: 6,157	BIC: 6,301	# Parameters: 41

The generalized degree of freedom parameter count is now fractional. The filter with smoothing parameter 0.4 is equivalent to adding 3.7 more parameters as compared to the 3 – break model, but both the AIC and BIC show substantial improvement.

The next figure plots the fitted trends and allows us to see the operation of the filter. The independent trends show us the unsmoothed trends if we allowed an independent trend parameter each calendar year. Note the apparent lag-1negative correlation of the independent trend estimates. This is in fact an expected characteristic of estimation error when the parameters estimated are trends. The anticipated negative lag-1 correlation is reflected in the parameter covariance matrix, and therefore corrected for by the GWA filter. This issue will be discussed further in Section 7.

Recalling the discussion of Section 5, we use a stationary random walk as a mechanism within the GWA filter; we do not imply that a random walk is the best description of the underlying time series process. The observed smoothed pattern doesn't appear at all like a random walk, and in fact fits quite well to a second order auto-regressive process (AR-2). This process can be forecast forward, producing the following mean forecast for the next nine calendar years:

Figure 6.2.
Independent and Smoothed Trend Estimates

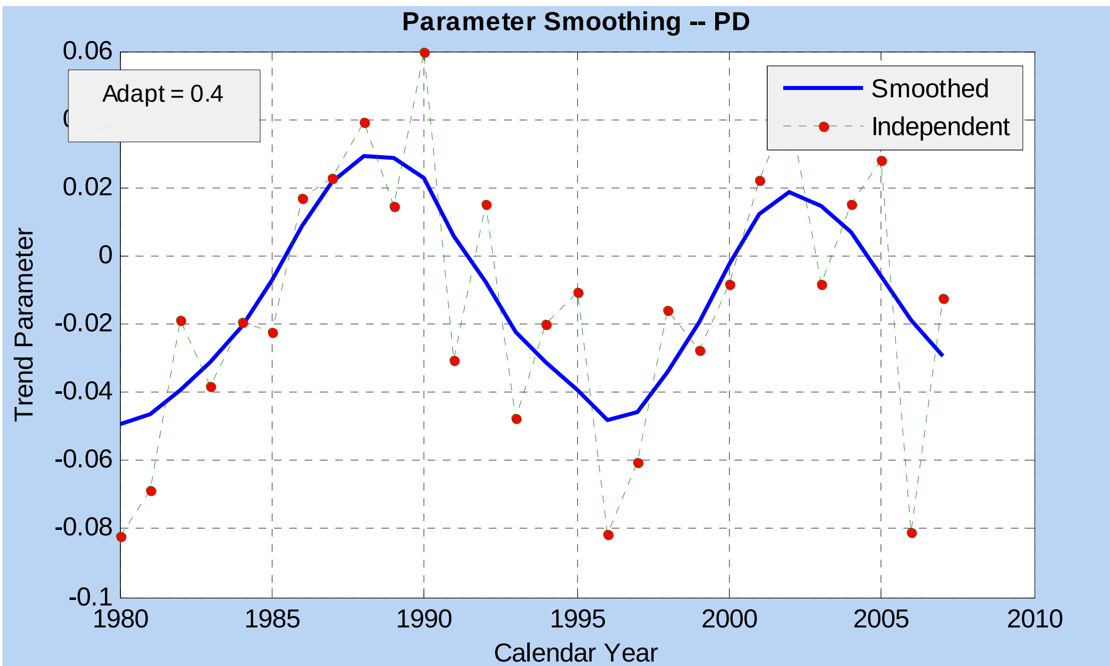
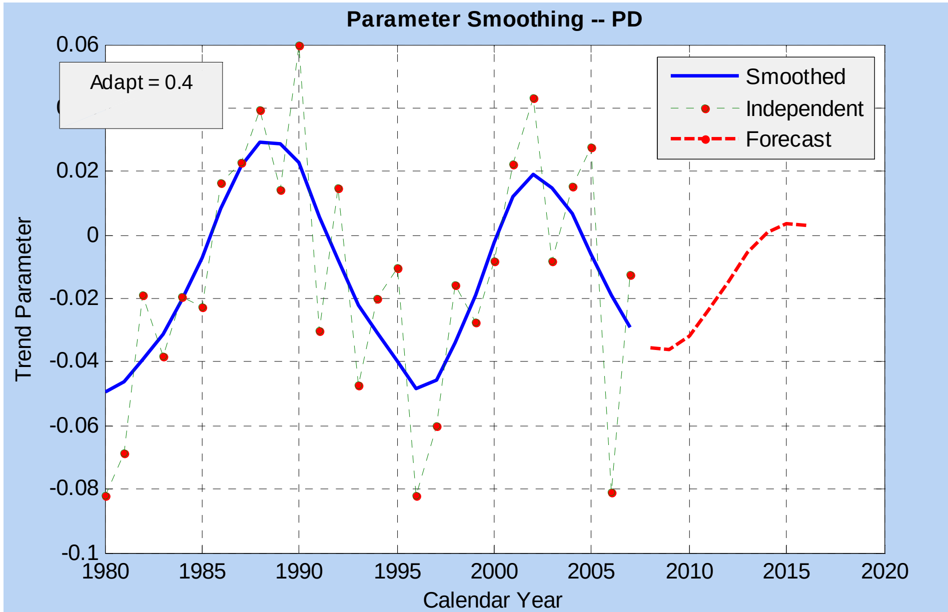


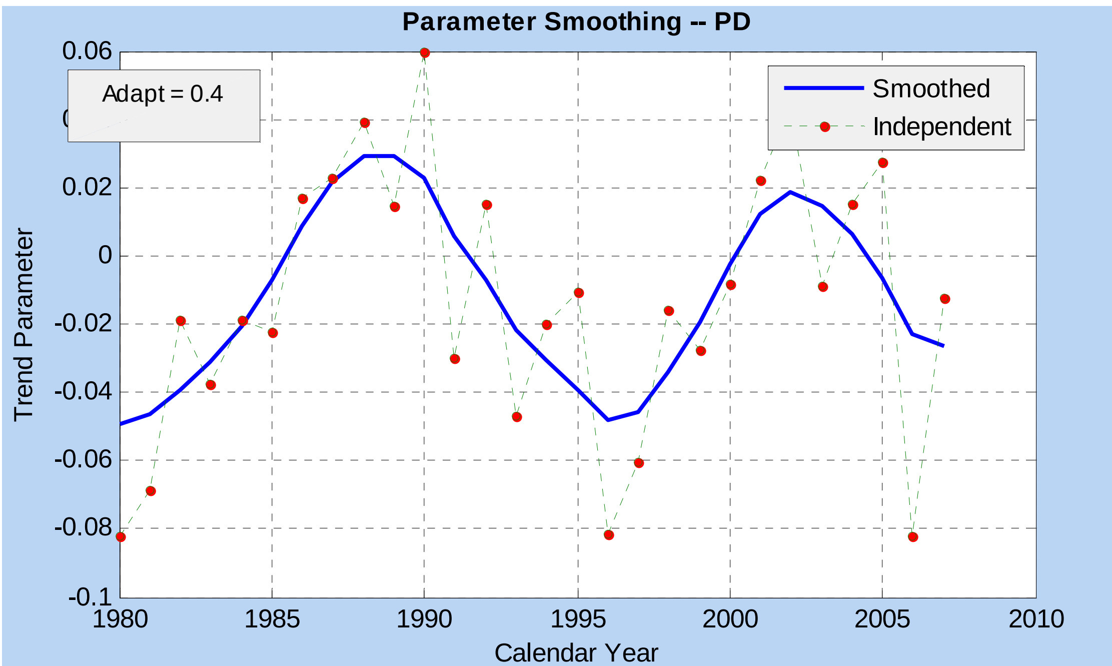
Figure 6.3
Smoothed Trends and AR-2 Forecast Trends



Selecting this trend forecast appears easier than the more difficult judgment required with the trend breaks model. Significant changes in CY trends have been observed, but the pattern of the volatility has been remarkably regular—a fact that we can hardly be confident of for the indefinite future. Nevertheless, back-testing on this dataset demonstrated dramatically improved forecast accuracy when compared to the chain-ladder. It is not necessarily likely that the same degree of improvement can be expected in general.

It has been noted that the stationary random walk assumption imposes little of its own structure on the smoothed pattern, but the mean zero assumption can have influence. The AR-2 model in this case exhibits significant local trends, although this behavior would not be evident in an AR-1 process. In any case, the filter typically picks up trends with little distortion, since points both before and after the smoothed point are weighted together. However, in the AR-2 case, there may be a problem at the endpoint, since all weight goes to points on one side, and the fitted value may be drawn back toward the mean. This can make a difference if the time series is being forecast, since the last values are influential. We have found that when the selected time series is AR-2, forecast accuracy is improved with an adjustment that eliminates the possible flattening effect for the last one or two values. In the prior figures (6.2 and 6.3) the last 2 values were adjusted. In contrast, Figure 6.4 includes the original last smoothed values, illustrating the flattening effect.

Figure 6.4.
Smoothed Trends (including last values)



Implications for Forecast Distributions

In the casualty environment, changing trends are the norm, not the exception, and the evidence in this large and stable data set is quite convincing. A dynamic environment is the null hypothesis, not to be rejected without extremely convincing evidence to the contrary. The traditional chain-ladder method may produce a perfectly plausible base estimate, but when the primary purpose becomes risk analysis, static models are contradicted by both common sense and the available statistical evidence.

We bootstrapped the MFE model of the industry workers' compensation data to forecast the prediction error of just the next five calendar years of payments. The typical standard error was about 3.0 percent. We then back-tested the chain ladder method, getting 17 actual readings of the real forecast errors. The actual mean square error ranged from about 8 percent to 11 percent, depending on the number of factors averaged. The fewest factors, just one or two, produced the most accurate forecasts. The longer term averages on which the bootstrapping is based were at the high end. Only about 10 percent of the true variance was predicted.

Bootstrapping the dynamic trend model on the same basis resulted in a typical forecast error of about 2.8 percent due to the better fitting model. Adding the simulated error of the fitted AR-2 model increased the forecast error to 4.5 percent; about 60 percent of the variance is attributed to the time series risk in a five-year forward forecast. The percentage will increase for longer forecasts. In this case, the forecast error of 4.5 percent was reasonably consistent with the back-tested accuracy of the method, but it would still be reasonable to assume that there is significant unmodeled specification error, even with the improved model.

The understatement of forecast error for the MFE model is exacerbated by the size and stability of the database: the measured process and parameter risks are small and the unmodeled risks dominate. With noisier data sets the degree of understatement may be smaller. In addition, the over-parameterization of the MFE model may, by happy coincidence, counteract the problem to some degree.

7. Autocorrelation Induced by Estimation

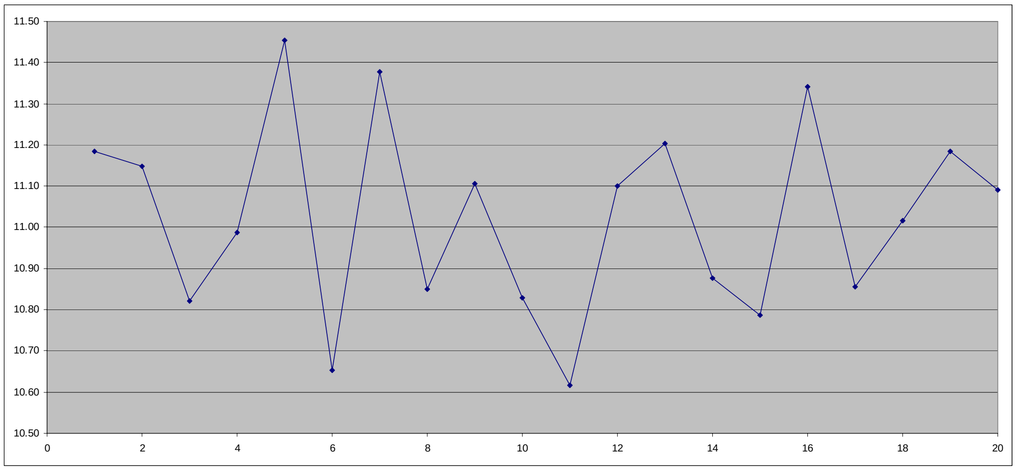
If a time series of parameters is estimated with independent errors, then the first differences are negatively autocorrelated at lag 1 due to a common parameter getting different signs in consecutive differences. For instance, $\mathrm{Cov}(b_3 - b_2, b_2 - b_1) = E[(b_3 - b_2)(b_2 - b_1)] - E(b_3 - b_2)E(b_2 - b_1) = E[b_3b_2 - b_2^2 - b_1b_3 + b_1b_2] - Eb_3Eb_2 - (Eb_2)^2 - Eb_1Eb_3 + Eb_1Eb_2. = -\mathrm{Var}(b_2)$. This can be compared to the variance of a single difference $\mathrm{Var}(b_2 - b_1) = \mathrm{Var}(b_2) + \mathrm{Var}(b_1)$. If all the individual variances are equal, the covariance of two adjacent differences is then – ½ of the variance of a single difference.

Thus an apparent auto-regressive process can be created by first differencing a process with independent observations. If this process is one of parameter estimates, there may be no auto-regression in parameters being estimated but an apparent lag 1 correlation of -50 percent in the increments. A possible response would be to do a GWA smoothing of the parameters before fitting a time-series model.

As a simplified example, a linear trend of +10 percent per year on a beginning value of 100 was projected for 21 years by simulation, including a normally distributed mean zero error in each trended value (not in the increment) with standard deviation of 2 percent. This might correspond to a series of parameters of an upwardly trending series each fit with that much parameter error. The autocorrelation of the 20 first differences was then computed. Over 500 simulations, the mean autocorrelation was -49 percent, with a standard deviation of 16 percent and a skewness of 44 percent.

The mean 1st difference should average to 11. These are graphed for the first simulation in Figure 7.1. The alternating up and down shifts are typical of a lag 1 autocorrelated process.

Figure 7.1.
Differences of Simulated Cost Levels



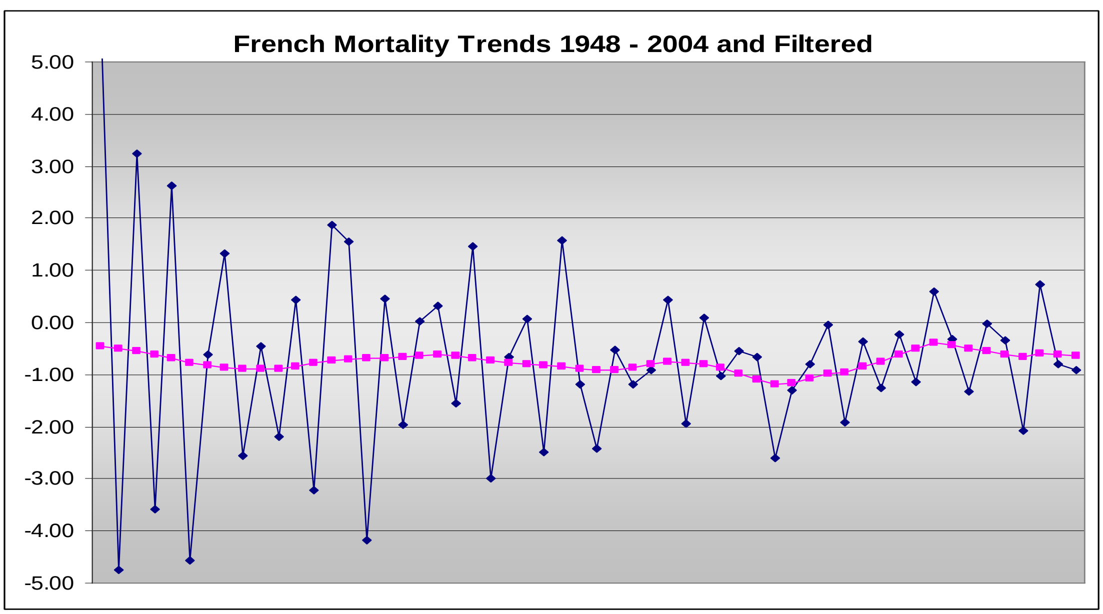
GWA can be applied, although this is a more simplified case than is usually used. Since there is no change in the underlying trend, each $s_j = 0$. The variance of any difference is $2(0.02)^2 = 0.0008$. Two consecutive differences have covariance -0.0004. Other covariances are zero. With this, every variance matrix V_j will be the same, so the smoothing will produce a constant value. For the first simulation this constant is 11.002. This is similar to, but, due to recognizing correlations, slightly different from, the mean difference in that simulation of 11.024.

The autocorrelation induced by estimation may lead to fitting an AR1 model to 1st differences of estimated level parameters. This can occur in casualty insurance if calendar-year cost levels are estimated individually and then a time-series model sought for the increments. The same thing can happen in modeling mortality by calendar year and then looking for a model to project changes in mortality into future calendar years. In mortality modeling there is usually a big enough triangle available to fit individual levels for each year, so several authors have followed this approach. Venter

(2008) found an autocorrelation of about 65 percent in calendar year mortality trends for 1948–2004. However, the first few years were very noisy, with a highly autocorrelative pattern in the increments. For the period 1954–2004, the lag 1 autocorrelation was 51 percent. This is close to what would be expected from a level trend. Thus standard linear trends, with standard confidence bands, might give a more reasonable projection of future mortality risk than would using an AR1 model.

However a similar autocorrelation could arise from a trend that changed occasionally over the period. If trends can do that, there is more risk in future trend. This could be modeled by a doubly mean-reverting stochastic process, or perhaps by a linear trend with occasional jumps in the trend that are generated by a compound Poisson process that generates the time to the next jump and th size of the jump. GWA filtering can provide a view of the process that can help judge what models are reasonable. Figure 7.2 illustrates this with a fairly high degree of smoothing for the French mortality trends from Venter (2008). In this case, a trend that persists for a random amount of time and then changes might seem reasonable.

Figure 7.2.
French Mortality Trends and DWA Smoothing



8. Conclusion

Forecasts significantly forward in calendar time are required in both casualty and life insurance, and understanding the uncertainty inherent in such forecasts is an essential element in risk analysis and risk management.

Time series can provide an appropriate structure for a model of time-related forecast risk. Actuaries in both fields measure trends in calendar time within their data sets, and must be alert to time series phenomena. But in each case, the true process is partially obscured by estimation error. In addition to obscuring the true process, estimated trends tend to exhibit substantial autoregressive effects that arise from the estimation process rather than the underlying process. This fact should be recognized in selecting projection models.

The technique of filtering can assist in detecting the true underlying process within the noise of estimation error, allowing a more valid model of time-related risk.

References

[1] Bailey, R.A. 1963. "Insurance Rates with Minimum Bias." PCAS 50: 4–11.

[2] Bailey, R.A., and Simon, L.J. 1960. "Two Studies in Automobile Insurance Rate Making." ASTIN Bulletin 1: 192–217.

[3] Barnett, G., and Zehnwirth, B. 2000. "Best Estimates for Reserves." PCAS 87: 245–321.

[4] Bornhuetter, R., and Ferguson, R. 1972. "The Actuary and IBNR." PCAS 59: 181–195.

[5] Buhlmann, H. 1983. "Estimation of IBNR Reserves by the Methods Chain Ladder, Cape Cod, and Complimentary Loss Ratio," unpublished.

[6] 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, Pensions Institute Discussion Paper PI-0701.

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

[8] de Jong, P. 2006. "Forecasting Runoff Triangles." North American Actuarial Journal 10(2): 28–38.

[9] de Jong, P., and Zehnwirth, B. 1983. "Claims Reserving, State-Space Models, and the Kalman Filter." Journal of the Institute of Actuaries 110:157–181.

[10] Gluck, S.M. 1997. "Balancing Development and Trend in Loss Reserve Analysis." PCAS 84: 482–532.

[11] 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.

[12] Hachemeister, C.A., and Stanard, J.N. 1975. "IBNR Claims Count Estimation with Static Lag Functions." (Lecture), ASTIN Colloquium, Portimão, Portugal, Sept. 30–Oct. 3, 1975.

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

[14] Hewitt, C.C. 1966 "Distribution by Size of Risk-A Model." PCAS 53: 106–114.

[15] Klugman, S.A., Panjer, H.H., and Willmot, G.E.. 2004. *Loss Models: From Data to Decisions*, 2nd ed. Hoboken, N.J.: Wiley.

[16] Kremer, E. 1982. "IBNR Claims and the Two Way Model of ANOVA." Scandinavian Actuarial Journal 1: 47–55.

[17] Kremer, E.. 1985. *Einfuhrung in die Versicherungsmathematik*. Göttingen, Germany: Vandenhoek & Ruprecht.

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

[19] Long, J.S., and Ervin, L.H. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." The American Statistician 54: 217–224.

[20] Mack, T. 1991. "A Simple Parametric Model for Rating Automobile Insurance or Estimating IBNR Claims Reserves." ASTIN Bulletin 21(1): 93–109.

[21] Mack, T. 1993. "Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates." ASTIN Bulletin 23(2): 213–225.

[22] Mack, T. 1994. "Measuring the Variability of Chain Ladder Reserve Estimates." CAS Forum (Spring): 101–182.

[23] Mack, T. 2002. *Schadenversicherungsmathematik*. 2nd ed. Karlsruhe, Germany: Verlag Versicherungswirtshaft.

[24] Murphy, D.M. 1994. "Unbiased Loss Development Factors." PCAS 81: 154–222.

[25] 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.

[26] Renshaw, A.E., and Verrall, R.J. 1998. "A Stochastic Model Underlying the Chain Ladder Technique." British Actuarial Journal 4: 903–923.

[27] Sherman, R. 1984. "Extrapolating, Smoothing, and Interpolating Development Factors." PCAS 71: 122–155.

[28] Stanard, J. 1985. "A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques." PCAS 72: 124–153.

[29] Taylor, G.C. 1977. "Separation of Inflation and Other Effects from the Distribution of Non-Life Insurance Claim Delays." ASTIN Bulletin 9(1–2): 219–230.

[30] Taylor, G.C. 2000. *Loss Reserving: An Actuarial Perspective*. Boston: Kluwer Academic Publishers.

[31] Taylor, G.C., and Ashe, F.R. 1983. "Second Moments of Estimates of Outstanding Claims." Journal of Econometrics 23: 37–61.

[32] Venter, G.G. 1998. "Testing the Assumptions of Age-to-Age Factors." PCAS 85: 807–847.

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

[34] Venter, G.G. 2007. "Generalized Linear Models beyond the Exponential Family with Loss Reserve Applications." CAS E-Forum, Summer (forthcoming).

[35] Venter, G.G. 2008. "Triangles in Life and Casualty." AFIR Colloquium Papers.

[36] Verbeek, H.G. 1972. "An Approach to the Analysis of Claims Experience in Motor Liability Excess of Loss Reassurance." ASTIN Bulletin 6(3): 195–202.

[37] Weissner, E.W. 1978. "Estimation of the Distribution of Report Lags by the Method of Maximum Likelihood." PCAS 65: 1–9.

[38] Ye, J. 1998 "On Measuring and Correcting the Effects of Data Mining and Model Selection." Journal of the American Statistical Association 93(441): 120–131.

Spencer M. Gluck, FCAS, MAAA, can be reached at *spencer.m.gluck@guycarp.com* or *spencergluck@yahoo.com*.

Gary G. Venter, FCAS, ASA, CERA, MAAA, is actuary in residence at Columbia University. He can be reached at *gary.venter@gmail.com*.