DISCUSSION PAPER PI-0802

Evaluating the Goodness of Fit of Stochastic Mortality Models

Kevin Dowd, Andrew J.G. Cairns, David Blake,
Guy D. Coughlan, David Epstein, and Marwa Khalaf-Allah

September 2008

ISSN 1367-580X

The Pensions Institute
Cass Business School
City University
106 Bunhill Row London
EC1Y 8TZ
UNITED KINGDOM

http://www.pensions-institute.org/

lifeMetrics

DISCLAIMER

Additional information is available upon request. This report has been partially prepared by the Pension Advisory group, and not by any research department, of JPMorgan Chase & Co. and its subsidiaries ("JPMorgan"). Information herein is obtained from sources believed to be reliable but JPMorgan does not warrant its completeness or accuracy. Opinions and estimates constitute JPMorgan's judgment and are subject to change without notice. Past performance is not indicative of future results. This material is provided for informational purposes only and is not intended as a recommendation or an offer or solicitation for the purchase or sale of any security or financial instrument.

Evaluating the Goodness of Fit of Stochastic Mortality Models

Kevin Dowd[*], Andrew J.G. Cairns[#], David Blake[♣]

Guy D. Coughlan, David Epstein, Marwa Khalaf-Allah[♦]

September 2008

Abstract

This study sets out a framework to evaluate the goodness of fit of stochastic mortality models and applies it to six different models estimated using English & Welsh male mortality data. The methodology exploits the structure of each model to obtain various residual series that are predicted to be iid standard normal under the null hypothesis of model adequacy. Goodness of fit can then be assessed using conventional tests of the predictions of iid standard normality. The models considered are Lee-Carter's 1992 one-factor model, a version of Renshaw-Haberman's 2006 extension of the Lee-Carter model to allow for a cohort effect, Currie's 2006 age-period-cohort model, which is a simplified version of the Renshaw-Haberman model, the Cairns-Blake-Dowd 2006 two-factor model and two generalised versions of the latter that allow for a cohort effect. For the data set considered, there are some notable differences amongst the different models, but none of the models performs well in all tests and no model clearly dominates the others.

Key words: goodness of fit, mortality models

[*] Centre for Risk & Insurance Studies, Nottingham University Business School, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom. Corresponding author: Kevin.Dowd@nottingham.ac.uk. The authors thank Lixia Loh and Liang Zhao for excellent research assistance.

[#] Maxwell Institute for Mathematical Sciences, and Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom.

[♣] Pensions Institute, Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom.

[♦] Coughlan, Epstein and Khalaf-Allah: Pension ALM Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom.

1. Introduction

In an earlier study, Cairns *et al.* (2007) examined eight different stochastic mortality models. These models – variously labelled M1 to M8 – were fitted to both English & Welsh and US male mortality data (over ages 60-89), and their results led the authors to conclude that five of these models – M2, M3, M6, M7 and M8 – provided the "best fits" to the two data sets based on a set of qualitative and quantitative criteria.[1] M2 is Renshaw and Haberman's generalisation of the one-factor Lee-Carter model (labelled M1) to incorporate a cohort effect (Renshaw and Haberman, 2006), M3 is the age-period-cohort (APC) model which is a simplification of M2 (Currie, 2006; see, also, Osmond, 1985 and Jacobsen *et al*., 2002), and M6, M7 and M8 are different generalisations of the two-factor model of Cairns, Blake and Dowd (2006) (labelled M5) that also incorporate a cohort effect.

The earlier study used both qualitative and quantitative evaluation criteria, where the latter consisted primarily of Bayesian Information Criterion (BIC) rankings complemented by nesting tests in cases where one model is a special case of another. The present study builds on this work in proposing a more complete and systematic methodology for establishing the quantitative goodness of fit (GOF) of a subset of the above models based on formal hypothesis testing. The objective is twofold: (a) for each model, to determine the complete set of testable implications that follow from the null hypothesis that the model provides a good fit to the historical data; and (b) to systematically test whether those predictions actually hold for one particular data set.

More specifically, we use what we know about the structure of each model to construct the following series that are predicted to be (at least approximately) independently and identically distributed standard normal (hereafter abbreviated to 'iid N(0,1)') under the null hypothesis:

[1] In particular, they found that for the English & Welsh males data, the Bayes Information Criterion ranked the models as follows: 1=M8, 2=M7, 3=M2, 4=M6, 5=M3, 6=M1, 7=M4 and 8=M5. They then dropped M4 (P-splines; Currie *et al*., 2004) from further analysis, in part because of these findings and in part because of its inability to project future stochastic mortality rates. They went on to obtain the following ranking of the remaining models on US data: 1=M2, 2=M7, 3=M3, 4=M8, 5=M6, 6=M1, 7=M5.

- Standardised mortality rate residuals or *mortality residuals* for short. The mortality residuals are the differences between the realised (or actual) mortality rates for any given set of ages and years and their model-generated equivalents (i.e., fitted values). Once standardised, these are predicted to be approximately iid N(0,1) under the null hypothesis.
- Standardised residuals of the model's unobservable state variables (SVs) or *SV residuals* for short. The SVs are the stochastic factors driving the dynamics of the model, and, once standardised, are also assumed to be approximately iid N(0,1).
- Standardised residuals for the prices (or fair values) of mortality-dependent financial instruments derived from the model (or *price residuals* for short), where the residuals concerned are the differences between these prices and their model-based equivalents, and these too should be approximately iid N(0,1) under the null hypothesis.

In particular, we apply the GOF framework to six models, namely, M1, M2B (a version of M2), M3B (a version of M3), M5, M6 and M7.[2,3] Of these, models M2B, M3B, M6 and M7 involve a cohort effect, whereas M1 (Lee-Carter) and M5 (Cairns-Blake-Dowd) do not. Each model was estimated using a single data set involving LifeMetrics data for the mortality rates of English & Welsh males[4] for ages from 64 to 89 and spanning the years 1961 to 2004.[5] As such, the results presented herein are not necessarily representative of what might be obtained for other data sets. They do however serve to illustrate the methodology and potential pitfalls in certain stochastic mortality models.

[2] M2B and M3B are the versions of M2 and M3 that assume an ARIMA(1,1,0) process for the cohort effect.
[3] We do not consider M8 in this paper because the results presented in Cairns *et al.* (2008) suggest that its forecasts on US mortality data are unreliable.
[4] See Coughlan *et al.* (2007) and www.lifemetrics.com for the data and a description of LifeMetrics. The original source of the data was the UK Office for National Statistics.
[5] The under-64s were excluded because it is the mortality rates of older people that are of the greatest financial significance to pension funds and annuity providers – and this is our main interest in conducting this series of studies on stochastic mortality models – and the mortality rates of those over age 89 were excluded because of poor data reliability. We would also emphasise that models M5-M7 were specifically designed for the higher age ranges, whereas the other models considered in this study were designed to fit younger ages as well.

The paper is organised as follows. Section 2 explains our notation. Section 3 outlines and implements the testing framework for each model's mortality residuals. Section 4 does the same for each model's SV residuals. Section 5 provides some test results for the price of an illustrative mortality-dependent financial contract, namely a period term annuity. Section 6 concludes.

2. Notation

We begin with some notation, and distinguish between the following mortality rates:

- $q(t,x)$ = true (and unobserved) mortality rate, i.e., the probability of death between times t and t+1 for individuals aged x at time t;
- $\tilde{q}(t,x)$ = crude estimate of year-t mortality rate based on observed deaths and exposures data;
- $\bar{q}(t,x)$ = estimated year-t mortality rate based on data up to and including year t, and using a specified mortality model (i.e., the fitted value from the model).

The crude mortality rate $\tilde{q}(t,x)$ is linked (by assumption) to the crude death rate, $\tilde{m}(t,x)$, via $\tilde{q}(t,x)=1-exp\left(-\tilde{m}(t,x)\right)$.

The models we consider involve the following SVs:

- $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$ are the true (unobserved) age, period and cohort effects given that the relevant specified model is true;
- $\bar{\beta}_x^{(i)}$, $\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$ are their estimates given data from years t_0 to t_1 and ages x_0 to x_1, and which are used to calculate the $\bar{q}(t,x)$;
- $\hat{\beta}_x^{(i)}$, $\hat{\kappa}_t^{(i)}$ and $\hat{\gamma}_c^{(i)}$ are their one-step ahead forecasts given data from years t_0 to t_1-1 and ages x_0 to x_1.

The cohort effects are estimated for years of birth c_0 to c_1, where the year of birth is equal to $t-x$.

3. Assessing the Goodness of Fit of the Mortality Residuals

Assessing goodness of fit involves three stages: estimation, implementation and testing.

3.1. Estimation

We start by selecting a lookback window on which to base our initial estimates. We choose a rolling 21-year window comprising the current and previous 20 years' historical observations.[6]

We then estimate the model and obtain estimates of the unobserved SVs $\overline{\beta}_x^{(i)}$, $\overline{\kappa}_t^{(i)}$ and $\overline{\gamma}_c^{(i)}$ and model-based estimates of the mortality rate $\overline{q}(t,x)$. In the present context, the sequence of 21-year rolling windows gives us estimates for 24 years between 1981 and 2004.[7]

The mortality residual is calculated as the difference between $\tilde{q}(t,x)$ and $\overline{q}(t,x)$. If the underlying random variable, the number of deaths, follows the assumption of a Poisson distribution, then the distribution of deaths can be approximated by a normal distribution as the number of deaths gets 'large', as seems reasonable when we consider the size of the male population of England & Wales. If a model's estimates are adequate, the mortality residuals should also be approximately normal. The standardised mortality residuals – found by subtracting the residual mean and dividing

[6] We could also have a chosen a window that expands over time to take account of the fact that our data accumulate over time. Having started with 20 observations to obtain our estimates for 1981, we might have used 21 observations to obtain estimates for 1982, and so forth. However, an expanding window would complicate the underlying statistics. A rolling fixed-length window is more straightforward to deal with.

[7] Note that each additional year on the length of the rolling window would reduce our number of observations by 1. A 21-year rolling window seems to strike a reasonable balance between the conflicting needs, on the one hand, for the window to be long enough to provide statistically reliable estimates, and, on the other, to have enough observations to be able to carry out convincing tests.

the result by the residual standard deviation – are then predicted to be approximately iid N(0,1).[8]

By way of example, and to make our discussion of estimation issues more concrete, consider the case of model M1 (whose structure is set out in equations (1) and (2) below):

1. We first take the exposures and deaths data from 1961 to 1981 and fit the model to obtain estimates for the age effects $\overline{\beta}_x^{(1)}$ and $\overline{\beta}_x^{(2)}$ and the period effect $\overline{\kappa}_t^{(2)}$ (see equation (1) below).
2. We then insert these into (1) to obtain the model-based death rate $\overline{m}(t,x)$ and thence the model-based mortality rate $\overline{q}(t,x)$ and the mortality residual $\tilde{q}(t,x)-\overline{q}(t,x)$ for 1981.
3. We repeat this process using data for 1962-1982 to get the mortality residual for 1982; we repeat again using data for 1963-1983 to obtain the 1983 mortality residual, and carry on in the same manner until we use data for 1984-2004 to obtain the 2004 mortality residual.

The other models are estimated in comparable ways.

3.2. Implementation

We now summarise the implementation for each model[9] in turn.

Model M1

Model M1, the original Lee-Carter model, postulates that the true underlying death rate, $m(t,x) = -\log(1-q(t,x))$, satisfies the following equation:

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} \qquad (1)$$

[8] For convenience, we use the term 'tested for iid N(0,1)' as shorthand for 'tested for the predictions of iid N(0,1)', where these predictions are those of a zero mean, a unit variance, a zero skewness, a kurtosis equal to 3, and, of course, independent and identically distributed.

[9] For additional details on each model, see Cairns *et al*. (2007, 2008).

where the state variable $\kappa_t^{(2)}$ follows a one-dimensional random walk with drift (Lee and Carter, 1992):

$$\kappa_t^{(2)} = \kappa_{t-1}^{(2)} + \mu + CZ_t^{(2)} \tag{2}$$

in which μ is a constant drift term, C is a constant volatility and $Z_t^{(2)}$ is a one-dimensional iid N(0,1) error.

Now let $D(t,x)$ be the number of deaths between *t* and *t*+1 at age *x* last birthday, and let $E(t,x)$ be the corresponding exposures. From these, we calculate the crude death rates $\tilde{m}(t,x) = D(t,x)/E(t,x)$. Given the Poisson assumption about deaths and given that the expected number of deaths is large, the number of deaths is approximately normal with mean and variance both equal to $\bar{m}(t,x)E(t,x)$. It follows that the standardised mortality residuals

$$\varepsilon(t,x) = \frac{\tilde{m}(t,x) - \bar{m}(t,x)}{\sqrt{\bar{m}(t,x)/E(t,x)}} \tag{3}$$

should be approximately iid N(0,1) under the null hypothesis.[10,11] Moreover, we would expect this prediction to hold both when we follow any given age from one year to the next and when we compare the death rates for different ages during the same year. Thus, the matrix of $\varepsilon(t,x)$ terms should be approximately iid N(0,1) in both dimensions.[12]

[10] We say 'approximately' in part because we are using estimates of the SVs rather than their true values, in part because there are likely to be measurement errors in the data (e.g., estimates of exposures are likely to be subject to errors) and in part because the assumed Poisson process with a fixed arrival or mortality rate at any point in time is likely to be an over-simplification of reality.

[11] The reader will also note that (3) strictly refers to death-rate rather than mortality-rate residuals. However, the former will have the same distribution as the latter, so for expositional purposes it is convenient to ignore the difference between them.

[12] This prediction holds for each mortality model.

We have $26 \times 24 = 624$ observations in the $\varepsilon(t,x)$ matrix (i.e., we have observations for each of 26 different ages spanning 64 to 89, over 24 different years spanning 1981 to 2004).[13]

Model M2B

This model, which is a particular extension of the Lee-Carter model to allow for a cohort effect, postulates that $m(t,x)$ satisfies:

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_c^{(3)} \tag{4}$$

where the state variable $\kappa_t^{(2)}$ follows (2) and $\gamma_c^{(3)}$ is a cohort effect where $c = t - x$ is the year of birth. We follow Cairns *et al*. (2008) and CMI (2007) and model the cohort effect, $\gamma_c^{(3)}$, as an ARIMA(1,1,0) process that is independent of $\kappa_t^{(2)}$:[14]

$$\Delta\gamma_c^{(3)} = \mu_\gamma + \alpha_\gamma\left(\Delta\gamma_{c-1}^{(3)} - \mu_\gamma\right) + \sigma_\gamma Z_c^{(\gamma)} \tag{5}$$

Model M3B

This model is a simplified version of M2B. It postulates that $m(t,x)$ satisfies:

$$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_c^{(3)} \tag{6}$$

where the variables (including the cohort effect) are the same as for M2B.

Model M5

M5 is a reparameterised version of the Cairns-Blake-Dowd (CBD) two-factor mortality model (Cairns *et al.*, 2006). The model postulates that $q(t,x)$ satisfies:

[13] This is also the case for every model.

[14] Cairns *et al.* (2008) found that a better statistical fit was provided by an ARIMA(0,2,1) model (labelled M2A). However, this alternative model was much less reasonable from a biological perspective than M2B, which used ARIMA(1,1,0). Note too that the gamma-related parameters in (5) – μ_γ and so forth – are specific to $\gamma_c^{(3)}$, but we do not make this explicit in order to avoid cumbersome notation (i.e., $\mu_{\gamma^{(3)}}$, etc.). We apply the same principle throughout the paper.

$$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) \tag{7}$$

where $\bar{x}$ is the average of the ages used in the dataset, and where the state variables now follow a two-dimensional random walk with drift:

$$\kappa_t = \kappa_{t-1} + \mu + CZ_t \tag{8}$$

where μ is a constant 2×1 drift vector, C is now a constant 2×2 upper triangular 'volatility' matrix (and more precisely, the Choleski 'square root' matrix of the variance-covariance matrix), and Z_t is a two-dimensional standard normal variable, each component of which is independent of the other.[15]

Model M6

M6 is a generalised version of M5 with a cohort effect, i.e.,

$$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_c^{(3)} \tag{9}$$

where the κ_t process follows (8) and the $\gamma_c^{(3)}$ process follows (5). Thus, M6 has the same κ_t process as M5 and the same $\gamma_c^{(3)}$ process as M2B and M3B.

Model M7

Our last model, M7, is another generalised version of M5 with a cohort effect, i.e.,

$$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \sigma_x^2) + \gamma_c^{(4)} \tag{10}$$

[15] The reparameterisation of the original model is $\kappa_t^{(2)} = A_2(t)$ and $\kappa_t^{(1)} - \kappa_t^{(2)}\bar{x} = A_1(t)$, where $A_1(t)$ and $A_2(t)$ are the state variables of the original model. An additional difference between the original CBD model and the reparameterised version M5 is that *x* in M5 refers to age at time *t*, whereas in the original CBD model it refers to age at some initial time 0.

where the state variables κ_t in this case follow a three-dimensional random walk with drift, σ_x^2 is the variance of the age range used in the dataset,[16] and $\gamma_c^{(4)}$ is a cohort effect that is modelled as an AR(1) process.

3.3 Test Results

The hypothesis tests used in this section aim to identify whether the mortality residuals described above are consistent with iid N(0,1) as assumed under the null hypothesis. The residuals are tested in two ways: (i) by year, i.e., aggregated across ages; and (ii) by age, i.e., aggregated across years. Each of these involves four types of tests:

- *t*-test of mean prediction;
- Variance ratio (VR) test – see Cochrane (1988), Lo and MacKinley (1988,1989);
- Normality test based on the skewness and kurtosis predictions – see Jarque and Bera (1980);
- Serial correlation test (based on the test statistic $\rho\sqrt{N-2}/(1-\rho^2)$ which is distributed under the null hypothesis as a *t*-distribution with *N*-2 degrees of freedom).

A 'statistically significant' result for any of these tests – which we take to be any test which produces a *p*-value of less than 1% – indicates inconsistency with iid N(0,1).

As the analysis is quite involved, we defer any detailed discussion of these individual tests to an Appendix and instead report only summary results in the main text. Perhaps the most useful summary results are the percentages of the test results for each model that are significant at the 1% level. These are reported in Table 1.

[16] The generalisation incorporates an additional quadratic age effect as well as a cohort effect.

Table 1: Percentages of $\varepsilon(t,x)$ Test Results Significant at the 1% Level

Model	By Year	By Age	Average	Ranking by Average
M1	31.3	30.8	31.1	=4
M2B	8.3	7.7	8.0	1
M3B	31.3	30.8	31.1	=4
M5	32.3	33.7	33.0	6
M6	15.6	18.3	17.0	3
M7	16.7	14.4	15.6	2

Notes: Based on the Tables in the Appendix. 'Average' refers to the average of the 'By Year' and 'By Age' results.

Under the null hypothesis, we would expect these percentages to be around 1%. Instead, we find that these percentages lie in the range 7.7% to 33.7%. This indicates that all models are problematic showing deviations from iid N(0,1) in a higher-than-expected percentage of tests. However, we can also see that by this criterion M2B gets $\varepsilon(t,x)$ scores that are better than any of the other models. M7 and M6 come second and third respectively. The other models come some way behind M7 and M6 and there is not much to choose between them.

4. Assessing the Goodness of Fit of the State Variable Residuals

4.1 Estimation

The derivation of the test results for the SV residuals is complicated by the fact that the SVs are unobservable. We therefore need to obtain estimates of the unobserved state variables ($\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$) using 21 years of data up to and including year *t*. If we had direct observations of the state variables ($\tilde{\kappa}_t^{(i)}$ and $\tilde{\gamma}_c^{(i)}$) in the same way that we have direct observations of the mortality rates, $\tilde{q}(t,x)$, we could have proceeded in the same way as in the previous section: we would have obtained the period-effect residuals as $\tilde{\kappa}_t^{(i)} - \bar{\kappa}_t^{(i)}$ and the cohort-effect residuals as $\tilde{\gamma}_c^{(i)} - \bar{\gamma}_c^{(i)}$. However, this is not possible because $\tilde{\kappa}_t^{(i)}$ and $\tilde{\gamma}_c^{(i)}$ are not directly observable. We therefore need proxies for these observations, and we obtain these proxies using 1-step ahead forecasts based on a model estimated using 20 years of data up to and including year *t*-1. If we denote these forecasts by $\hat{\kappa}_t^{(i)}$ and $\hat{\gamma}_c^{(i)}$, the estimated period-effect residuals

become $\hat{\kappa}_t^{(i)} - \bar{\kappa}_t^{(i)}$ and the estimated cohort-effect residuals become $\hat{\gamma}_c^{(i)} - \bar{\gamma}_c^{(i)}$. We now standardise each of these series by subtracting its estimated mean and dividing the result by its estimated one-period-ahead standard deviation. The resulting standardised SV residual series are then each predicted to be approximately iid N(0,1) under the null hypothesis.

For each model, we have one or more sets of standardised SV residuals. The number of standardised SV residual series depends on the model – it is equal to the number of period effects (which varies from 1 to 3) and the number of cohort effects (which is either 0 or 1) in each model. The number of standardised SV residual series in each model therefore varies from 1 to 4.

As an aside, the fact that the model is re-estimated for each year in our sample period means that we are working with estimates for μ and C that are regularly updated. Accordingly, in the discussion below, we let $\bar{\mu}_t$ and $\bar{C}_t$ denote their estimates based on data up to and including year *t*.

4.2 Implementation

We now consider each model in turn.

Model M1

For M1, we use (2) to obtain estimated values of $\kappa_t^{(2)}$ (i.e., $\bar{\kappa}_t^{(2)}$) and 1-step ahead forecasts of $\kappa_t^{(2)}$ (i.e., $\hat{\kappa}_t^{(2)}$), viz.:[17]

$$\bar{\kappa}_t^{(2)} = \bar{\kappa}_{t-1}^{(2)} + \bar{\mu}_{t-1} + \bar{C}_{t-1}\bar{Z}_t^{(2)} \tag{11}$$

$$\hat{\kappa}_t^{(2)} = \bar{\kappa}_{t-1}^{(2)} + \bar{\mu}_{t-1}. \tag{12}$$

[17] When we use the 20-year window to obtain the $\hat{\kappa}_t^{(2)}$ forecasts, we need to ensure that any constraints in the estimation process are used in a fashion consistent with way in which the $\bar{\kappa}_t^{(2)}$ estimates were obtained. Thus, for M1, we use the constraints $\sum_{t=1961}^{1980} \kappa_t^{(2)} = 0$ and $\sum_{x=x_0}^{x_1} \beta_x^{(2)} = 1$ for both $\bar{\kappa}_t^{(2)}$ and $\hat{\kappa}_t^{(2)}$.

Substituting (12) into (11) and rearranging gives the standardised SV residuals:

$$\bar{Z}_t^{(2)} = \bar{C}_{t-1}^{-1}(\bar{\kappa}_t^{(2)} - \hat{\kappa}_t^{(2)}). \tag{13}$$

In (13), $\bar{\kappa}_t^{(2)}$ is the estimated value of $\kappa_t^{(2)}$ based on data from *t*-20 up to and including time *t*, and $\hat{\kappa}_t^{(2)}$ is the 1-step ahead forecasted value of $\kappa_t^{(2)}$ based on data from *t*-20 up to and including time *t*-1. This gives us 24 values of $\bar{Z}_t^{(2)}$ and, under the null hypothesis, these are predicted to be iid N(0,1).

Model M2B

For M2B, we obtain the standardised SV residuals $\bar{Z}_t^{(2)}$ using (13), and we model the cohort effect $\gamma_c^{(3)}$ and recover the standardised cohort-effect residuals $\bar{Z}_c^{(\gamma)}$ using (5). Both standardised residual series $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$ are predicted to be iid N(0,1).

We can also test the properties of both sets of estimated residuals simultaneously. Since $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$ should each be iid N(0,1) and independent of each other, statistical theory tells us that the sum of squares of 2 independent N(0,1) variates is distributed as a chi-squared with 2 degrees of freedom. It therefore follows that:

$$\bar{Y}_t = [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2 \sim \chi_2^2$$
$$\Rightarrow \quad \bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{iid U(0,1)} \tag{15}$$

where $F(.)$ is the distribution function for a chi-squared with 2 degrees of freedom. Under the null, the series $\bar{p}_t$ should be distributed as iid standard uniform (or iid U(0,1)). If we wished to, we could then test this series using standard uniformity tests such as Kolmogorov-Smirnov, Kuiper, and Lilliefors, etc.[18] However, testing is easier (and we have more tests available) if we put $\bar{p}_t$ through the following transformation:

[18] For more on these tests, see, e.g., Dowd (2005, chapter 15 appendix).

$$\overline{h}_t = \Phi^{-1}\left(\overline{p}_t\right) \sim \text{iid N(0,1)} \tag{16}$$

where $\Phi(.)$ is the distribution function for a standard normal variable. This transformation gives us an 'observed' series $\overline{h}_t$ that is distributed as iid N(0,1) under the null. We can then test whether $\overline{h}_t$ is iid N(0,1).

Model M3B

The standardised SV residuals for M3B are obtained in exactly the same way as for M2B.

Model M5

For model M5, we use (8) to obtain the 2x1 vector $\overline{\kappa}_t$ and the 1-step ahead forecasts $\hat{\kappa}_t$:

$$\overline{\kappa}_t = \overline{\kappa}_{t-1} + \overline{\mu}_{t-1} + \overline{C}_{t-1}\overline{Z}_t \tag{17}$$

$$\hat{\kappa}_t = \overline{\kappa}_{t-1} + \overline{\mu}_{t-1} \tag{18}$$

$$\Rightarrow \overline{Z}_t = \overline{C}_{t-1}^{-1}(\overline{\kappa}_t - \hat{\kappa}_t) \tag{19}$$

Under the null, each standardised SV residual series – that is, $Z_t^{(1)}$ and $Z_t^{(2)}$ – is iid N(0,1) and independent of the other.

We now test $Z_t^{(1)}$ and $Z_t^{(2)}$ for iid standard normality using conventional tests, and additionally apply a standard correlation test to check the prediction that these have a zero correlation.

As with M2B and M3B, we can also test the properties of both sets of standardised residuals simultaneously. In this case, under the null hypothesis,

$$\overline{Y}_t = [\overline{Z}_t^{(1)}]^2 + [\overline{Z}_t^{(2)}]^2 \sim \chi_2^2$$

$$\overline{p}_t = F\left(\overline{Y}_t\right) \sim \text{iid U(0,1)} \tag{20}$$

$$\Rightarrow \bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{iid N(0,1).} \tag{21}$$

We now test $\bar{h}_t$ for iid N(0,1).

Model M6

Following the same logic, for M6, we obtain

$$\bar{Z}_t = \bar{C}_{t-1}^{-1}(\bar{\kappa}_t - \hat{\kappa}_t) \tag{22}$$

which gives us two sets of standardised SV residuals $Z_t^{(1)}$ and $Z_t^{(2)}$ that are predicted to be iid N(0,1) and independent of each other. And, as with M2B and M3B, we obtain the corresponding standardised cohort-effect residuals that are also predicted to be iid N(0,1). It then follows that

$$\bar{Y}_t = [\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{\gamma}] \sim \chi_3^2$$

$$\Rightarrow \bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{iid U(0,1)} \tag{23}$$

$$\Rightarrow \bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{iid N(0,1).} \tag{24}$$

which we then test for iid N(0,1).

Model M7

M7 is similar but involves three sets of standardised SV residuals - $Z_t^{(1)}$, $Z_t^{(2)}$ and $Z_t^{(3)}$ - rather than two. M7 also involves standardised cohort-effect residuals $\bar{Z}_c^{(\gamma)}$.[19] Applying the same logic as before then gives us:

$$\bar{Y}_t = [Z_t^{(1)}]^2 + [Z_t^{(2)}]^2 + [Z_t^{(3)}]^2 + [Z_c^{(\gamma)}]^2 \sim \chi_4^2$$

$$\Rightarrow \bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{iid U(0,1)} \tag{25}$$

$$\Rightarrow \bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{iid N(0,1)} \tag{26}$$

[19] Note however that $\bar{Z}_c^{(\gamma)}$ now refers to the standardised residual of the $\gamma_c^{(4)}$ process rather than that of the $\gamma_c^{(3)}$. The context makes it clear which gamma process $\bar{Z}_c^{(\gamma)}$ is referring to.

where $F(.)$ is now the distribution function for a chi-squared with 4 degrees of freedom. As in earlier cases, we then test $\bar{h}_t$ for iid N(0,1).

4.3 Test Results

We begin by reiterating that the hypothesis tests used to assess GOF in this section test whether the state variable residuals are (as appropriate) singly or jointly iid N(0,1).

*Model M1*Table 2 presents the sample moments and the test results for M1's standardised SV residual series, $\bar{Z}_t^{(2)}$, and these results are compatible with the null hypothesis of standard normality. However, the null hypothesis of temporal independence is strongly rejected. Altogether, there are 4 p-values reported for M1, and, of these, one is significant at well under the 1% level. If we treat any p-values below 1% as a 'fail', then, by this criterion, M1 has a 'failure' rate of one test out of four or 25%.

Table 2: Results for $\bar{Z}_t^{(2)}$: Model M1

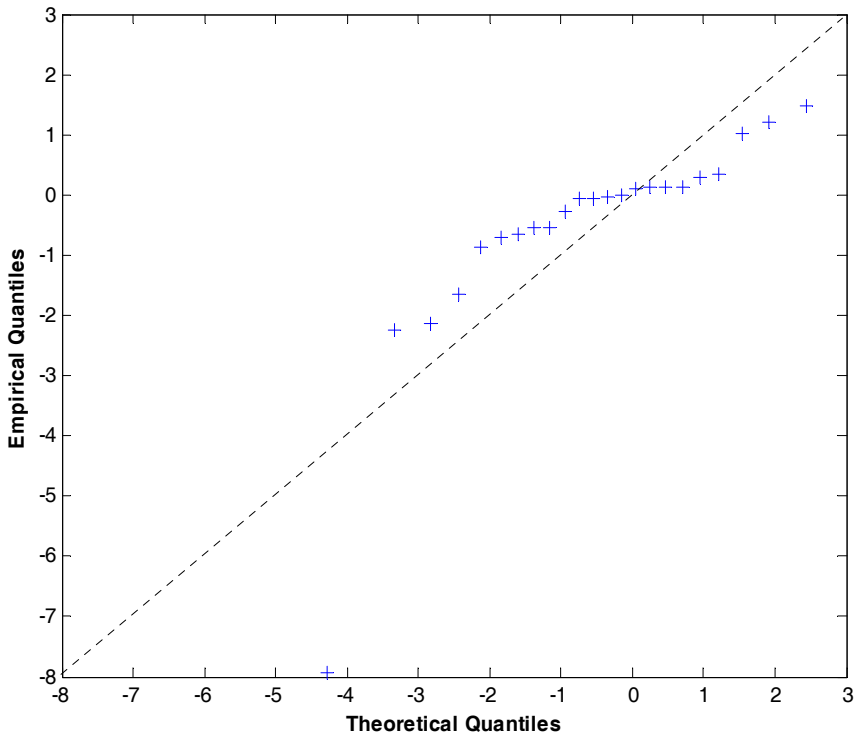
Sample moments	
Mean	-0.385
Variance	1.070
Skewness	0.060
Kurtosis	2.625
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.082
Test of variance ratio prediction	
P-value variance ratio test stat	0.740
Test of normality prediction	
P-value Jarque-Bera test stat	0.925
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	-0.570
P-value correlation	0.001**

Notes: Based on 24 annual observations spanning 1981-2004. All tests are two-sided except for the JB test which is inherently one-sided. If ρ is the correlation coefficient, $\rho\sqrt{N-2}/(1-\rho^2)$ is distributed under the null as a *t*-distribution with *N*-2 degrees of freedom. ****** indicates significance at the 1% level.

Model M2B

Figures 1 and 2 give the QQ plots[20] for the model's two standardised SV residual series, $\overline{Z}_t^{(2)}$ and $\overline{Z}_c^{(\gamma)}$, and Figure 3 gives a plot of empirical vs. predicted $\overline{p}_t$. We can see that all three Figures show extremely poor fits: the two QQ plots have one or more very extreme outliers (especially for the cohort effect plot in Figure 2) and do not lie close to the 45° line; the $\overline{p}_t$ plot in Figure 3 clearly does not lie anywhere close to its predicted 45° line either. There are therefore very clear problems with both this model's standardised residuals series.

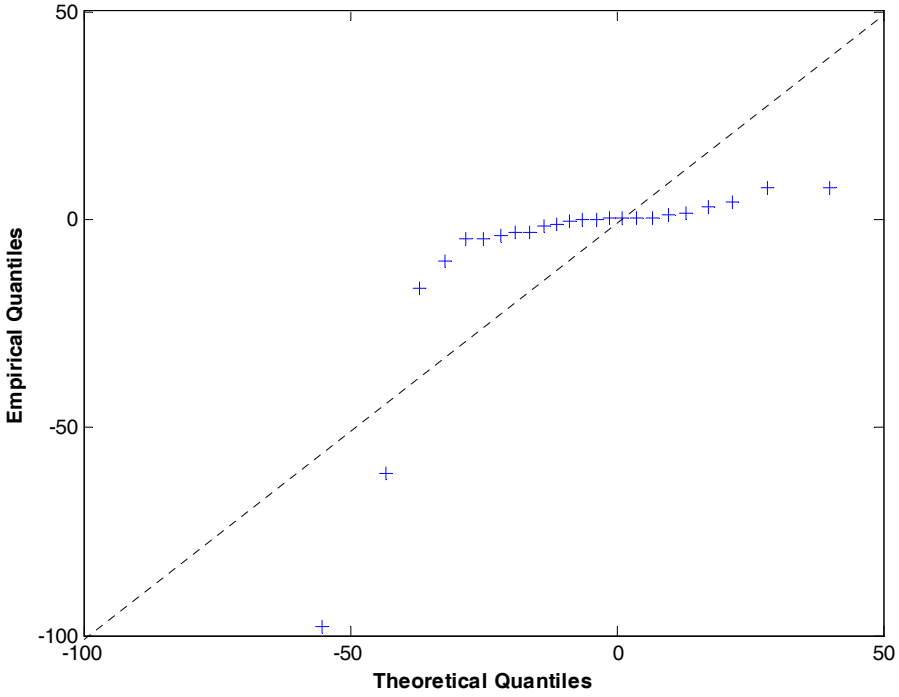
Figure 1: QQ Plot for $\overline{Z}_t^{(2)}$: Model M2B



Notes: Based on 24 annual $\overline{Z}_t^{(2)}$ observations of model M2B over the period 1981-2004.

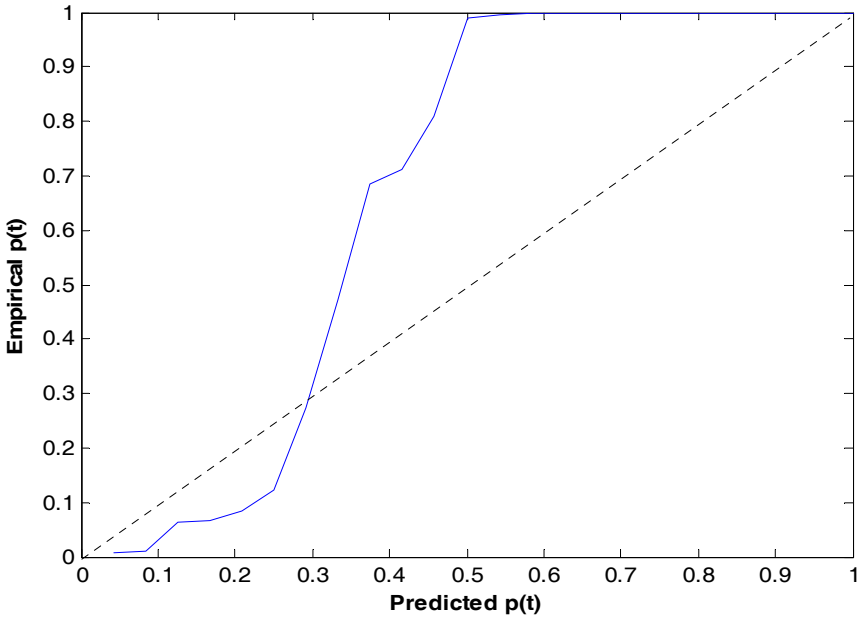
[20] A QQ plot is a plot of the empirical quantiles of a distribution against their predicted counterparts, where the latter in this case are based on the prediction of standard normality. QQ plots give a useful visual indicator of whether the empirical quantiles are consistent with the predicted ones: under the null, we would expect the plots to lie fairly close to the 45° line. Note that we do not report the QQ and associated plots for models other than M2B, as these are all compatible with the underlying null hypotheses. The plots for M2B on the other hand are more informative.

Figure 2: QQ Plot for $\overline{Z}_c^{(\gamma)}$: Model



Notes: Based on 24 annual $\overline{Z}_c^{(\gamma)}$ observations of model M2B over the period 1981-2004.

Figure 3: Plot of Empirical vs. Predicted $\overline{p}_t$: Model M2B



Note: Based on 24 annual $\overline{p}_t$ observations of model M2B. $\overline{p}_t = F\left([\overline{Z}_t^{(1)}]^2 + [\overline{Z}_c^{(\gamma)}]^2\right)$, where $F(.)$ is the distribution function of a χ_2^2.

These impressions are confirmed by the results of Table 3, which presents the sample moments and the test results for each of $\overline{Z}_t^{(2)}$ and $\overline{Z}_c^{(\gamma)}$. Both perform poorly – both series score *p*-values of 0 for the variance and normality tests – and the sample moments of the $\overline{Z}_c^{(\gamma)}$ bear no resemblance to the predictions. Similarly, the $\overline{h}_t$ test results in Table 4 lead us to reject the null hypothesis that the standardised residuals are jointly iid N(0,1).

Note that there are 12 p-values reported for M2B, and of these six are below 1%. If we again treat any p-values below 1% as 'fails', then, by this criterion, M2B has a 'failure' rate of 50%.

It is worth pausing for a moment to consider why M2B produces such poor results. If the model and fitting procedure were robust, then adding in one year's data should only have a small impact on the estimated age, period and cohort effects. However, it was found with M2B – but not with any of the other models considered in this study – that adding one extra year of data could lead the model to jump from one set of fitted values for the cohort effect to a completely different set.[21] This problem is most likely explained by the likelihood function having multiple maxima. The changes in parameter values then reflect a jump in the fitting algorithm from one maximum to another.[22]

[21] These claims are borne out by graphs of fitted parameter values (not included here), which show considerable instability for M2B. By contrast, graphs of the fitted parameter values for other models are all stable. For further discussion of the stability problem, see Cairns *et al*. (2008). The authors of CMI WP 25 encountered similar problems. To quote from their study: "the fitted cohort parameters do not appear to be stable as the age range fitted is changed" (CMI, 2007, p. 18, para 7.18); "when back-testing a dataset or fitting a different age range, we were unable to find a set of starting parameter values that consistently worked for different subsets of the data. Where a number of sets of starting parameter values worked for a particular dataset, we also found that the fitted values could differ materially" (CMI, 2007, p. 19, para 7.21).

[22] These jumps, in turn, lead to the fitted standardised residuals having some very extreme values as shown in Figures 1-3 and Tables 3-4.

Table 3: Results for $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$: Model M2B

Sample moments		
	$\bar{Z}_t^{(2)}$	$\bar{Z}_c^{(\gamma)}$
Mean	-0.450	-7.558
Variance	3.551	545.648
Skewness	-2.813	-3.262
Kurtosis	13.910	13.801
N	24	24
Test of mean prediction[1]		
P-value mean *t*-test stat	0.254	0.127
Test of variance ratio prediction		
P-value variance ratio test stat	0.000**	0.000**
Test of normality prediction		
P-value Jarque-Bera test stat	0.000**	0.000**
Test of temporal independence[2]		
Pearson correlation (*t*+1,*t*)	-0.132	-0.026
P-value correlation	0.534	0.905

Notes: As per Notes to Table 2.

Table 4: Results for $\bar{h}_t$: Model M2B

Sample moments	
Mean	1.498
Variance	5.447
Skewness	-0.433
Kurtosis	1.466
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.005**
Test of variance ratio prediction	
P-value variance ratio test stat	0.000**
Test of normality prediction	
P-value Jarque-Bera test stat	0.212
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	0.167
P-value correlation	0.429

Notes: $\bar{h}_t = \Phi^{-1}(\bar{p}_t)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. Note, however, that in 8 cases, the estimated value of $\bar{h}_t$ was 1. Since the normal inverse of 1 is undefined, these values were reduced to 0.9999 for the purposes of computing the results in this Table. Otherwise as per Notes to Table 3.

Model M3B

Table 5 presents the moments and test results for the standardised residuals for M3B. As with M2B, we have 12 reported p -values, but in this case only two are significant at the 1% level. M3B therefore has a 'failure' rate of 2/12$^{\text{ths}}$ or 16.7%.

Table 5: Results for $\overline{Z}_t^{(2)}$ and $\overline{Z}_c^{(\gamma)}$: Model M3B

Sample moments		
	$\overline{Z}_t^{(2)}$	$\overline{Z}_c^{(\gamma)}$
Mean	0.094	-0.284
Variance	0.872	1.948
Skewness	0.312	-0.524
Kurtosis	2.724	4.417
N	24	24
Test of mean prediction[1]		
P-value mean *t*-test stat	0.625	0.329
Test of variance ratio prediction		
P-value variance ratio test stat	0.722	0.008**
Test of normality prediction		
P-value Jarque-Bera test stat	0.793	0.212
Test of temporal independence[2]		
Pearson correlation (*t*+1,*t*)	-0.620	0.010
P-value correlation	0.000**	0.962

Notes: As per Notes to Table 2.

Table 6: Results for $\overline{h}_t$: Model M3B

Sample moments	
Mean	0.001
Variance	1.886
Skewness	0.820
Kurtosis	4.053
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.998
Test of variance ratio prediction	
P-value variance ratio test stat	0.012*
Test of normality prediction	
P-value Jarque-Bera test stat	0.150
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	0.090
P-value correlation	0.674

Notes: As per Notes to Table 4. $\overline{h}_t = \Phi^{-1}(\overline{p}_t)$, where $\overline{p}_t = F\left([\overline{Z}_t^{(2)}]^2 + [\overline{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. * indicates significance at the 5% level.

Model M5

Table 7 presents the sample moments and the test results for $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$ based on M5, and Table 8 presents the sample moments and test results for M5's $\bar{h}_t$ series. M5 has 13 p-values of which two are significant at the 1% level: M5 therefore has a 'failure rate' equal to 2/13^{ths} or 15.4%.

Table 7: Results for $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$: Model M5

Sample moments		
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$
Mean	-0.342	0.663
Variance	0.808	1.172
Skewness	0.171	0.080
Kurtosis	2.734	2.131
N	24	24
Test of mean prediction		
P-value mean *t*-test stat	0.075	0.006**
Test of variance ratio prediction		
P-value variance ratio test stat	0.550	0.516
Test of normality prediction		
P-value Jarque-Bera test stat	0.910	0.677
Test of temporal independence		
Pearson correlation (*t*+1,*t*)	-0.568	0.143
P-value correlation	0.001**	0.499
Correlation between $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$ ²		
Pearson correlation	-0.029	
P-value correlation	0.892	

Notes: Based on 24 annual observations spanning 1981-2004. All tests are two-sided except for the JB test which is inherently one-sided. If ρ is the correlation coefficient, $\rho\sqrt{N-2}/(1-\rho^2)$ is distributed under the null as a *t*-distribution with *N*-2 degrees of freedom. ****** indicates significance at the 1% level.

Table 8: Results for $\overline{h}_t$: Model M5

Sample moments	
Mean	0.103
Variance	1.507
Skewness	-0.145
Kurtosis	2.605
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.685
Test of variance ratio prediction	
P-value variance ratio test stat	0.112
Test of normality prediction	
P-value Jarque-Bera test stat	0.887
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	0.144
P-value correlation	0.498

Notes: As per Notes to Table 7. $\overline{h}_t = \Phi^{-1}(\overline{p}_t)$, where $\overline{p}_t = F\left([\overline{Z}_t^{(1)}]^2 + [\overline{Z}_t^{(2)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function.

Model M6

Table 9 and 10 present the comparable results for M6. This model has 19 p-values of which three are significant at the 1% level: M6 therefore has a 'failure rate' equal to 3/19$^{\text{ths}}$ or 15.8%.

Table 9: Results for $\overline{Z}_t^{(1)}$, $\overline{Z}_t^{(2)}$ and $\overline{Z}_c^{(\gamma)}$: Model M6

Sample moments			
	$\overline{Z}_t^{(1)}$	$\overline{Z}_t^{(2)}$	$\overline{Z}_c^{(\gamma)}$
Mean	-0.100	0.512	-0.570
Variance	0.844	1.117	2.393
Skewness	0.275	0.383	-0.554
Kurtosis	2.893	2.305	6.037
N	24	24	24
Test of mean prediction[1]			
P-value mean *t*-test stat	0.599	0.026	0.084
Test of variance ratio prediction			
P-value variance ratio test stat	0.645	0.632	0.000
Test of normality prediction			
P-value Jarque-Bera test stat	0.855	0.586	0.005
Test of temporal independence[2]			
Pearson correlation (*t*+1,*t*)	-0.557	-0.077	-0.109
P-value correlation	0.001	0.721	0.609
Correlation between $\overline{Z}_t^{(1)}$ and $\overline{Z}_t^{(2)}$		-0.1038	
***P*-value of correlation between $\overline{Z}_t^{(1)}$ and $\overline{Z}_t^{(2)}$**		0.6271	

Notes: As per Notes to Table 2.

Table 10: Results for $\bar{h}_t$: Model M6

Sample moments	
Mean	0.277
Variance	1.289
Skewness	-0.312
Kurtosis	1.806
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.243
Test of variance ratio prediction	
P-value variance ratio test stat	0.320
Test of normality prediction	
P-value Jarque-Bera test stat	0.404
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	0.016
P-value correlation	0.940

Notes: $\bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. Note, however, that in 1 case, the estimated value of $\bar{h}_t$ was 1, which was reduced to 0.9999 for the purposes of computing the results in this Table. Otherwise as per Notes to Table 9.

Model M7

Table 11 and Table 12 present the corresponding results for M7. For this model we have 23 p-values, of which 3 are significant. Hence, M7 has a failure rate equal to 3/23$^{\text{rds}}$ or 13.0%.

Conclusions to Section 4

The results of applying the state variable GOF tests to the six models are summarised in Table 13, which shows the proportions of test results for each model that are significant at the 1% level. It also shows the implied ranking by this criterion: M7 comes a little ahead of M5, which in turn comes a little ahead of M6 and then M3B. M1 comes somewhat further behind and M2B comes well behind the rest.

Table 11: Results for $\bar{Z}_t^{(1)}$, $\bar{Z}_t^{(2)}$, $\bar{Z}_t^{(3)}$ and $\bar{Z}_c^{(\gamma)}$: Model M7

	Sample Moments			
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	$\bar{Z}_c^{(\gamma)}$
Mean	-0.321	0.345	0.116	0.029
Variance	0.858	0.920	1.472	2.130
Skewness	0.165	0.721	0.287	-0.789
Kurtosis	2.701	2.857	2.340	6.985
N	24	24	24	24
	Test of mean prediction			
P-value mean *t*-test stat	0.103	0.091	0.643	0.923
	Test of variance ratio prediction			
P-value variance ratio test stat	0.684	0.856	0.135	0.002**
	Test of normality prediction			
P-value Jarque-Bera test stat	0.905	0.350	0.682	0.000**
	Test of temporal independence			
Pearson correlation (*t*+1,*t*)	-0.626	-0.079	0.083	-0.284
P-value correlation	0.000**	0.712	0.699	0.162
	Correlations			
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	
$\bar{Z}_t^{(1)}$	1			
$\bar{Z}_t^{(2)}$	-0.242	1		
$\bar{Z}_t^{(3)}$	0.145	-0.226	1	
	P-values of correlations[2]			
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	
$\bar{Z}_t^{(1)}$	1			
$\bar{Z}_t^{(2)}$	0.240	1		
$\bar{Z}_t^{(3)}$	0.492	0.275	1	

Notes: Based on 24 annual observations spanning 1981-2004. All tests are two-sided except for the JB test which is inherently one-sided. If ρ is the correlation coefficient, $\rho\sqrt{N-2}/(1-\rho^2)$ is distributed under the null as a *t*-distribution with *N*-2 degrees of freedom. ****** indicates significance at the 1% level.

Table 12 Results for $\overline{h}_t$: Model M7

Sample moments	
Mean	0.357
Variance	1.464
Skewness	0.862
Kurtosis	4.284
N	24
Test of mean prediction	
P-value mean *t*-test stat	0.161
Test of variance ratio prediction	
P-value variance ratio test stat	0.140
Test of normality prediction	
P-value Jarque-Bera test stat	0.099
Test of temporal independence	
Pearson correlation (*t*+1,*t*)	-0.012
P-value correlation	0.956

Notes: As per Notes to Table 11. $\overline{h}_t = \Phi^{-1}\left(\overline{p}_t\right)$, where $\overline{p}_t = F\left([\overline{Z}_t^{(1)}]^2 + [\overline{Z}_t^{(2)}]^2 + [\overline{Z}_t^{(3)}]^2 + [\overline{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function.

Table 13: Summary of Main Standardised Residual Results for the State Variables

Model	Proportion of test results significant at the 1% level	Implied ranking
M1	25%	5
M2B	50%	6
M3B	16.7%	4
M5	15.4%	2
M6	15.8%	3
M7	13.0%	1

Notes: Based on the results in Tables 2-12.

5. Assessing the Goodness of Fit of Model-based Annuity Price Residuals

Our final test of the adequacy of the models is to test the goodness of fit of the prices (or fair values) of financial assets that depend on model-based mortality forecasts. To illustrate, we consider the case of a period term annuity for males aged 65, payable until age 90.[23] We will assume the cashflows on the annuity are discounted using a fixed discount rate of 4%. We adopt procedures similar to those employed for testing the goodness of fit of the state variables. Take the first 20-year window covering 1961-1980. For this period, each model is used to obtain estimates of the underlying state variables: $\overline{\beta}_x^{(i)}$, $\overline{\kappa}_t^{(i)}$ and $\overline{\gamma}_c^{(i)}$. We then generate 1000 one-period ahead simulations of $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$ (i.e., for 1981). For each simulation and each model, we generate model-based mortality rates, $q(t,x)$, for ages between 65 and 90, and the corresponding period annuity prices, $a(t,x)$.[24] The 1000 simulated values give us an estimate of the one-period-ahead forecast distribution of $a(t,x)$ for each model, and we use this to estimate the mean, $\overline{a}(t,x)$, and the corresponding standard deviation. We then use the crude mortality rates, $\tilde{q}(t,x)$, for 1981 to calculate the "crude" period annuity price, $\tilde{a}(t,x)$. The annuity residual for each model is then $\tilde{a}(t,x)-\overline{a}(t,x)$ and this is standardised by dividing by the standard deviation of the one-period-ahead forecast distribution of the period annuity price for that year. This procedure is repeated for the remaining 20-year windows covering 1962-1981, 1963-1982 etc.

The sample moments and moment-based test statistics for the standardised annuity residuals are given in Table 14, and the main highlights are:

- All models give fairly reasonable sample moments for the residuals.
- M1, M3B, M5 and M6 each fail the iid test at the 1% significance level.

[23] A period term annuity is one that has a fixed term and ignores future mortality improvements. That is, for valuation purposes the annuity's future cash flows are calculated purely from the latest period mortality rates. We consider term annuities ceasing at age 90 because models M1, M2B and M3B, having been fitted to data from ages 60 to 89, predict mortality rates from ages 60 to 89 only. Their semi-parametric structure means that there is no natural way to use them to extrapolate mortality rates beyond age 89.

[24] Period annuity prices are calculated as follows. We define, first, the model-simulated period survival function $S(t,x,y)=\{1-q(t,x)\}\times\{1-q(t,x+1)\}\times\ldots\times\{1-q(t,y-1)\}$. The simulated period annuity price is then defined as $a(t,x)=\sum_{y=x+1}^{90}S(t,x,y)(1+r)^{-(y-x)}$ where $r=0.04$. Crude period annuity prices, $\tilde{a}(t,x)$, are calculated in the same way, replacing $q(t,x)$ by $\tilde{q}(t,x)$.

- M7 and M2B pass all tests at the 1% significance level.

These results suggest that M7 and M2B rank more or less equally first, and the others come afterwards with little to choose between them.

Table 14: Sample Moments and *P*-Values for Standardised Annuity Price Residuals

	M1	M2B	M3B	M5	M6	M7
Sample moments						
Mean	-0.194	-0.022	0.335	0.346	0.140	-0.192
Variance	0.463	0.872	0.803	0.782	0.788	1.305
Skewness	-0.382	0.163	-0.025	-0.163	-0.167	-0.026
Kurtosis	2.995	4.085	2.835	2.949	3.015	3.393
P-values of tests						
Mean test stat	0.176	0.909	0.080	0.068	0.446	0.420
VR test stat	0.027*	0.723	0.536	0.483	0.499	0.298
JB test stat	0.747	0.526	0.985	0.947	0.946	0.924
Corr(t+1,t)	-0.561	-0.381	-0.616	-0.548	-0.584	-0.028
Corr(t+1,t) test stat	0.001**	0.048*	0.000**	0.001**	0.000**	0.896

Notes: Based on males aged 65, payable until age 90, a discount rate of 4% and a sample size of 24. * indicates significance at the 5% level and ** indicates significance at the 1% level. See also notes to Table 1.

6. Conclusions

The present study sets out a framework for systematically evaluating the goodness of fit of stochastic mortality models, and applies it to a set of mortality models estimated using England & Wales male mortality data. If a model fits the data well, certain key residual series – those relating to mortality rates themselves, to the unobserved state variables that drive the dynamics of the model (including the cohort-effect where appropriate), and to the residuals of mortality-dependent financial prices – will, once standardised, be approximately iid N(0,1). We then test whether the relevant series are compatible with iid N(0,1).

We find that none of the models considered in this paper performs well in all sets of tests, and no model performs consistently better than the others. For the particular data set used in this analysis, however, we find that:

- For GOF tests of mortality residuals, model M2B performs best, M7 comes second and M6 third, and M1, M3 and M5 come some way behind.
- For the GOF tests of the state variables, M7, M5, M6 and M3B perform best, in that order, although there is not much to choose between them. M1 comes a little further behind. For its part, model M2B is well behind the others and also provides very poor fits of the mortality state variables.
- For the GOF tests of the annuity price residuals, M7 and M2B emerge as the best models and the other models come some way behind.

Two avenues for further work naturally suggest themselves. The first is to test these findings on other mortality data sets. A second is to evaluate the out-of-sample performance of the models' forecasts, i.e., "backtest" the models, and this is addressed in Dowd *et al.* (2008b).[25]

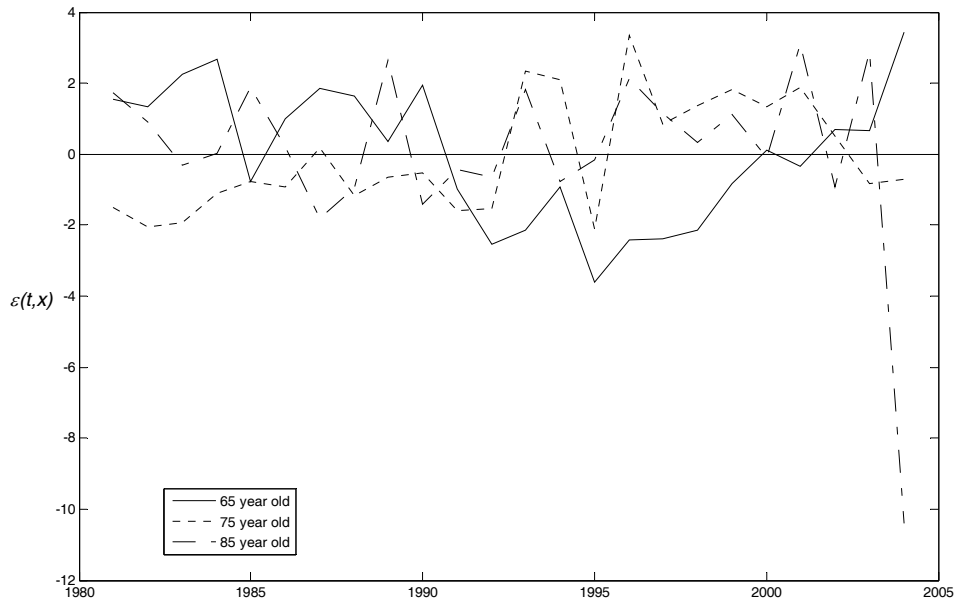
[25] A third avenue of research, which is much more ambitious, is to build a mortality model that is able to take account of the impact of exogenous factors (such as biomedical, environmental, and socio-economic factors) on mortality rates. Wilmoth (1998), for example, applied the Lee-Carter model to cause-of-death data and this would also be an interesting extension to other models.

Appendix: Analysis of the $\varepsilon(t,x)$ Results

Model M1

This Appendix examines the $\varepsilon(t,x)$ results in more detail than was done in the text, and we start with the simplest model, M1. As a preliminary, Figure A1 shows a plot of $\varepsilon(t,x)$ for x = 65, 75 and 85. Note that the plot should be consistent with an N(0,1) random variable under the null hypothesis. This plot shows rather more volatility than we would expect under the null hypothesis, and we get a number of notable outliers, most especially in the plot for 85-year olds for 2004.

Figure A1: Plots of $\varepsilon(t,x)$: Model M1



Tables A1 and A2 present test results for the standardised mortality residuals, $\varepsilon(t,x)$. Table A1 presents results organised year by year, and Table A2 presents results organised by age. They show that deviations from iid N(0,1) are significant for many of the test results. In particular, about 31.3% of test results are significant at the 1% level or below in the case of Table A1, and the corresponding percentage for Table A2 is 30.8%.

The most striking result is that the variances are above (and often well above) their predicted value of 1, which confirms the impression given in Figure 1. Further, the *p*-values for the variance ratio (VR) test are always very low and almost always zero. These findings are consistent with those of Cairns *et al.* (2007).

Tables A1 and A2 also reveal significant serial correlation between errors in the age and period dimensions. Under our null hypothesis, the $\varepsilon(t,x)$ should be independent, so the high correlations here indicate that M1 is failing to model some significant structural dependence between ages and calendar years.

Table A1: $\varepsilon(t,x)$ Results by Year: Model M1

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	0.105	0.084	-0.152	-0.426	0.123	-0.125	-0.423	-0.148	0.206	-0.119	0.102	-0.089
Variance	2.441	1.956	1.727	4.013	1.797	1.903	5.332	3.136	2.534	2.928	1.737	2.953
Skewness	0.366	0.423	0.417	0.916	-0.199	0.691	0.492	0.620	0.515	0.521	0.664	0.492
Kurtosis	2.786	2.564	2.381	3.349	2.054	3.630	2.792	2.178	2.078	2.804	2.316	2.939
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.736	0.762	0.561	0.288	0.644	0.649	0.359	0.674	0.515	0.726	0.696	0.794
P-value VR test stat	0.000**	0.006**	0.027*	0.000**	0.017*	0.008*	0.000**	0.000**	0.000**	0.000**	0.025*	0.000**
P-value JB test stat	0.730	0.612	0.557	0.152	0.566	0.287	0.578	0.302	0.355	0.544	0.298	0.590
Corr (x+1,x)	0.425	0.363	0.628	0.754	0.340	0.300	0.758	0.399	0.423	0.138	0.163	0.129
P-value corr	0.018*	0.052	0.000**	0.000**	0.071	0.120	0.000**	0.029*	0.019*	0.496	0.420	0.527

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	0.187	-0.065	0.107	0.065	0.087	0.035	0.109	-0.071	0.011	-0.022	0.059	-0.174
Variance	2.178	1.896	3.112	2.909	2.493	1.588	2.837	2.124	2.778	4.450	5.344	7.594
Skewness	-0.329	1.095	-0.590	-0.442	-0.332	-0.664	-0.605	-0.123	-0.409	-1.167	-1.255	-1.987
Kurtosis	2.283	5.110	2.610	2.768	2.856	2.381	2.975	2.522	3.702	6.255	6.063	10.069
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.523	0.811	0.759	0.848	0.782	0.889	0.744	0.805	0.973	0.957	0.897	0.751
P-value VR test stat	0.001**	0.009**	0.000**	0.000**	0.000**	0.063	0.000**	0.002**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.599	0.007**	0.433	0.637	0.778	0.313	0.452	0.855	0.533	0.000**	0.000**	0.000**
Corr (x+1,x)	0.357	-0.275	0.561	0.398	0.588	0.676	0.356	0.437	0.218	0.304	0.004	0.058
P-value corr	0.056	0.158	0.001**	0.029*	0.000**	0.000**	0.057	0.014*	0.273	0.114	0.983	0.778

Notes: $\varepsilon(t,x) = (\tilde{m}(t,x) - \overline{m}(t,x)) / \sqrt{\overline{m}(t,x) / E(t,x)}$ and each column refers to the $\varepsilon(t,x)$ results for each given year. VR is the variance ratio test (Cochrane (1988), Lo and MacKinley (1988, 1989)). JB is the Jarque-Bera normality test (Jarque and Bera (1980)). The correlation test is based on the test statistic, $\rho\sqrt{N-2}/(1-\rho^2)$, which is distributed under the null as a *t*-distribution with *N*-2 degrees of freedom.***** indicates significance at 5% level and ****** indicates significance at 1% level. 31.3% of test results are significant at the 1% level.

Table A2: $\varepsilon(t,x)$ Results by Age: Model M1

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	0.283	0.018	-0.179	0.050	-0.043	0.264	-0.042	-0.203	-0.124	-0.129	0.000	-0.070	-0.127
Variance	4.661	3.626	3.496	2.778	2.598	3.747	2.636	2.651	1.895	2.780	2.848	2.521	2.251
Skewness	0.098	-0.148	0.735	-0.109	0.440	1.056	-0.144	0.725	0.225	0.710	0.373	0.576	0.175
Kurtosis	1.804	2.066	4.023	2.726	2.670	3.342	2.465	2.774	3.389	4.109	2.580	2.164	1.714
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.527	0.963	0.644	0.884	0.897	0.510	0.899	0.546	0.662	0.708	1.000	0.830	0.683
P-value VR test stat	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.012*	0.000**	0.000**	0.000**	0.001**
P-value JB test stat	0.480	0.619	0.201	0.941	0.644	0.102	0.831	0.340	0.838	0.197	0.694	0.363	0.411
Corr (t+1,t)	0.745	0.650	0.480	0.635	0.225	0.490	0.239	0.226	-0.063	-0.093	0.135	0.302	0.391
P-value corr	0.000**	0.000**	0.008**	0.000**	0.279	0.006**	0.247	0.276	0.771	0.665	0.524	0.134	0.042*

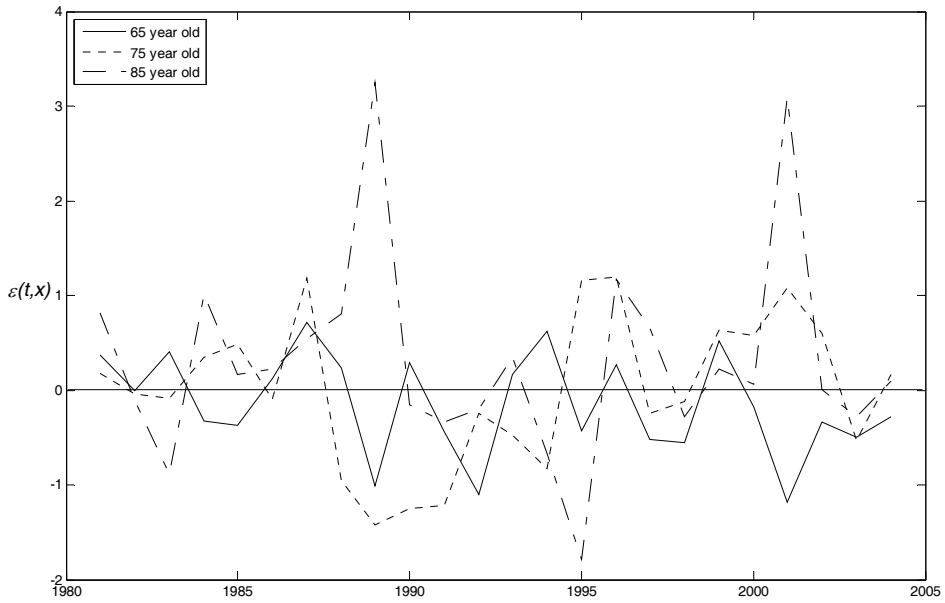
Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	-0.436	-0.231	-0.166	-0.435	-0.331	0.059	-0.253	0.250	0.085	0.244	0.007	0.401	0.530
Variance	2.393	2.202	2.512	2.195	3.090	3.803	3.863	4.388	6.932	1.839	1.617	1.777	2.075
Skewness	0.314	-0.440	0.021	-0.145	0.169	-0.794	-1.564	-2.141	-2.799	-0.660	0.469	-0.025	-0.633
Kurtosis	2.031	3.241	2.386	3.507	3.069	3.552	7.532	10.122	14.135	3.273	1.813	2.096	4.162
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.180	0.453	0.612	0.164	0.366	0.884	0.534	0.564	0.876	0.387	0.977	0.154	0.085
P-value VR test stat	0.000**	0.002**	0.000**	0.002**	0.000**	0.000**	0.000**	0.000**	0.000**	0.017*	0.062	0.025*	0.004**
P-value JB test stat	0.514	0.660	0.827	0.843	0.942	0.243	0.000**	0.000**	0.000**	0.403	0.318	0.664	0.228
Corr (t+1,t)	0.611	0.578	0.407	0.325	0.316	0.227	-0.259	-0.010	-0.455	0.027	0.222	-0.224	-0.017
P-value corr	0.000**	0.000**	0.032*	0.102	0.113	0.275	0.206	0.962	0.013*	0.899	0.285	0.282	0.935

Notes: As per Notes to Table A1, except each column refers to the $\varepsilon(t,x)$ results for each cohort of any given age. 30.8% of test results are significant at the 1% level.

Model M2B

Figure 2 shows the corresponding plot for model M2B. The volatility in these plots is considerably lower than it was for Figure 1, but we get a couple of notable outliers for 85 year olds.

Figure A2: Plots of $\varepsilon(t,x)$: Model M2B



Tables A3 and A4 present test results for the standardised mortality residuals, $\varepsilon(t,x)$, organised by year and age, respectively. These results are much better than for M1, with a much higher percentage of test results consistent with the null hypothesis that the residuals are iid N(0,1). In the case of Table A3, 8.3% of test results are significant at the 1% level, and, in the case of Table A4, the corresponding percentage is 7.7%. These percentages are higher than their predicted values of around 1%, but they are still much lower than those obtained for model M1. It is also noteworthy that the $\varepsilon(t,x)$ variances are much lower than their M1 counterparts and are closer to their predicted value of 1. Also, serial correlations between the $\varepsilon(t,x)$ are generally lower than for M1. The fact that M2B seems to perform better than M1 might suggest that it is important to model the cohort effect in this dataset.

Table A3: $\varepsilon(t,x)$ Results by Year: Model M2B

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	-0.028	0.010	0.038	0.011	-0.019	0.005	0.041	0.017	-0.153	-0.036	-0.092	-0.018
Variance	0.808	0.696	0.802	0.630	1.732	0.714	1.057	0.722	2.329	0.840	0.952	0.520
Skewness	0.002	-0.292	-0.093	-0.002	0.029	0.292	-0.271	-0.799	-0.072	0.127	0.633	-0.208
Kurtosis	2.316	3.524	1.785	2.266	2.322	3.658	3.043	4.492	3.352	1.940	3.528	2.561
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.876	0.950	0.833	0.944	0.941	0.974	0.840	0.921	0.614	0.843	0.635	0.900
P-value VR test stat	0.527	0.266	0.511	0.156	0.026	0.303	0.770	0.320	0.000**	0.615	0.937	0.047*
P-value JB test stat	0.776	0.716	0.441	0.747	0.778	0.658	0.852	0.075	0.925	0.525	0.361	0.820
Corr (x+1,x)	-0.032	0.029	0.225	-0.161	0.451	0.054	0.030	0.001	0.665	0.295	0.201	0.164
P-value corr	0.877	0.886	0.256	0.427	0.011*	0.793	0.886	0.996	0.000	0.127	0.316	0.416

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	0.085	-0.080	-0.005	-0.030	-0.013	0.020	0.092	-0.076	-0.094	-0.043	-0.018	0.035
Variance	0.407	1.037	0.978	0.668	1.028	0.288	0.687	1.509	1.965	0.978	0.856	0.846
Skewness	-0.554	-0.015	0.031	-0.454	-1.021	0.303	-0.561	-1.102	-0.983	-0.030	0.071	-0.635
Kurtosis	3.522	3.243	2.653	2.612	6.144	3.704	2.358	6.093	7.231	4.319	3.335	4.957
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.505	0.692	0.978	0.853	0.949	0.848	0.577	0.755	0.735	0.827	0.921	0.849
P-value VR test stat	0.008**	0.824	0.987	0.215	0.847	0.000**	0.249	0.098	0.005**	0.986	0.661	0.632
P-value JB test stat	0.444	0.968	0.935	0.590	0.000**	0.627	0.404	0.000**	0.000**	0.389	0.931	0.052
Corr (x+1,x)	0.161	-0.130	-0.063	-0.026	-0.077	0.231	0.079	-0.037	-0.010	-0.101	0.037	-0.049
P-value corr	0.426	0.522	0.758	0.900	0.709	0.244	0.699	0.858	0.962	0.622	0.859	0.812

Notes: As per Notes to Table A1. 8.3% of test results are significant at the 1% level.

Table A4: $\varepsilon(t,x)$ Results by Age: Model M2B

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	0.020	-0.147	-0.207	-0.200	0.178	0.299	0.319	-0.108	-0.078	0.086	0.110	0.003	0.019
Variance	0.992	0.284	0.682	0.470	0.939	0.813	0.545	0.867	0.917	1.026	0.799	0.616	0.898
Skewness	-0.086	-0.289	0.143	0.170	-0.600	-0.890	-0.414	-0.412	-2.428	-0.699	0.233	-0.141	1.116
Kurtosis	1.980	2.373	2.397	3.069	3.434	6.213	2.731	3.063	10.896	3.578	2.603	2.210	3.879
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.924	0.189	0.231	0.167	0.376	0.117	0.045*	0.576	0.693	0.682	0.554	0.983	0.922
P-value VR test stat	0.943	0.001**	0.263	0.030*	0.911	0.563	0.077	0.709	0.848	0.854	0.525	0.157	0.795
P-value JB test stat	0.585	0.695	0.801	0.941	0.443	0.001*	0.684	0.711	0.000**	0.318	0.829	0.703	0.056
Corr (t+1,t)	-0.167	-0.049	-0.069	-0.224	0.187	0.052	-0.015	-0.220	0.143	-0.044	-0.350	0.395	0.399
P-value corr	0.429	0.820	0.749	0.280	0.373	0.808	0.945	0.291	0.500	0.838	0.075	0.039*	0.037*

Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	-0.208	-0.108	0.130	-0.244	-0.118	-0.156	-0.205	0.173	0.323	0.032	-0.415	0.111	0.010
Variance	0.776	1.259	0.653	0.642	1.405	2.191	1.244	1.125	1.177	0.885	1.019	1.096	0.890
Skewness	0.091	-1.076	0.557	-0.543	-1.160	-1.207	0.176	-0.239	1.273	0.219	-0.078	0.471	-0.874
Kurtosis	2.494	4.429	2.272	2.668	7.846	5.872	2.250	2.180	5.951	2.227	3.272	2.753	4.399
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.259	0.643	0.439	0.150	0.631	0.610	0.378	0.433	0.159	0.870	0.056	0.608	0.959
P-value VR test stat	0.469	0.363	0.212	0.195	0.188	0.002**	0.388	0.613	0.506	0.760	0.872	0.680	0.774
P-value JB test stat	0.865	0.036*	0.412	0.525	0.000**	0.001**	0.709	0.637	0.001**	0.674	0.952	0.622	0.082
Corr (t+1,t)	-0.074	0.174	0.308	0.044	0.047	0.060	0.104	0.147	-0.026	0.104	0.550	0.364	0.010
P-value corr	0.729	0.409	0.124	0.838	0.827	0.779	0.628	0.489	0.906	0.627	0.001**	0.061	0.963

Notes: As per Notes to Table A2. 14.2% of test results are significant at the 1% level.

Model M3B

Figure A3 presents the rolling $\varepsilon(x,t)$ plots for model M3B. These are similar to, but somewhat worse than, those for M1. Tables A5 and A6 present test results for this model's standardised mortality residuals, and these are very close to those for model M1, with 31.3% and 30.8% of results respectively showing significant deviations from iid N(0,1) at the 1% level.

Figure A3: Plots of $\varepsilon(x,t)$: Model M3B

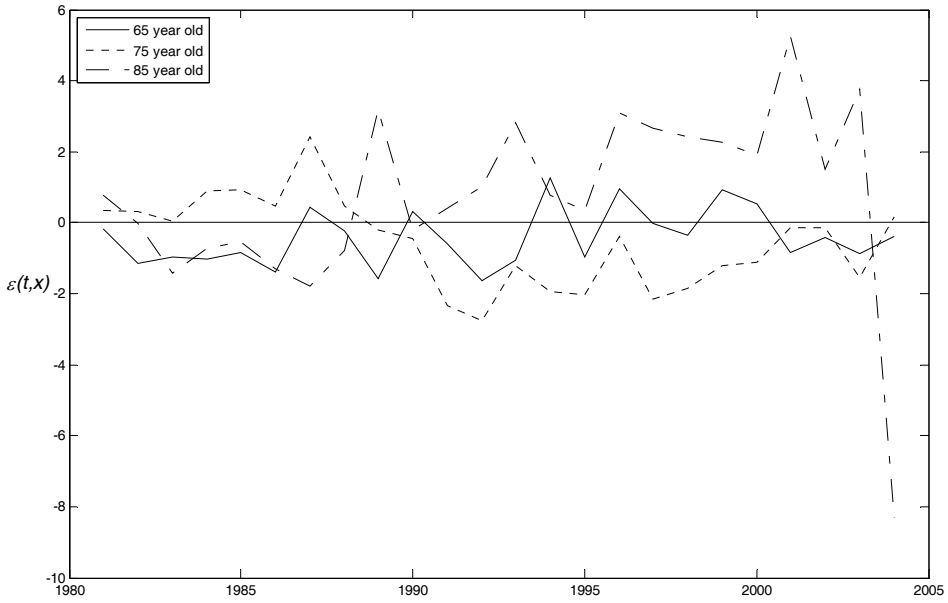


Table A5: $\varepsilon(t,x)$ Results by Year: Model M3B

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	0.085	0.029	-0.053	-0.130	-0.104	-0.124	-0.146	-0.090	0.004	-0.046	0.037	0.001
Variance	1.212	0.956	1.016	2.412	2.178	2.216	2.846	1.968	2.921	1.453	2.440	1.212
Skewness	0.219	-0.008	0.047	-0.009	-0.328	0.353	-0.604	0.127	-0.355	0.111	0.282	-0.643
Kurtosis	2.196	3.308	2.623	2.513	2.445	2.841	2.464	2.525	2.696	2.864	2.666	2.930
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.698	0.880	0.791	0.673	0.722	0.676	0.663	0.745	0.990	0.847	0.904	0.995
P-value VR test stat	0.427	0.949	0.882	0.000**	0.001**	0.001**	0.000**	0.005**	0.000**	0.134	0.000**	0.427
P-value JB test stat	0.635	0.950	0.922	0.879	0.670	0.753	0.389	0.854	0.724	0.964	0.792	0.407
Corr (x+1,x)	0.240	0.434	0.117	0.213	0.482	0.684	0.482	0.420	0.595	0.310	0.487	0.262
P-value corr	0.223	0.015*	0.565	0.284	0.005**	0.000**	0.005**	0.020*	0.000**	0.106	0.005**	0.181

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	0.074	0.041	0.045	0.072	0.066	0.078	0.088	0.059	0.036	0.009	0.003	0.039
Variance	3.050	2.541	3.424	3.453	3.278	2.616	5.279	3.447	5.191	5.304	6.256	5.369
Skewness	-0.081	-0.431	0.274	-0.185	0.121	-0.094	-0.810	-0.679	0.206	-0.474	-0.528	-1.393
Kurtosis	1.748	2.699	1.750	3.638	2.903	2.238	5.270	4.143	4.645	5.439	5.144	9.682
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.832	0.897	0.903	0.844	0.854	0.808	0.846	0.872	0.937	0.985	0.996	0.932
P-value VR test stat	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.422	0.636	0.364	0.745	0.964	0.716	0.015*	0.182	0.210	0.025*	0.045*	0.000**
Corr (x+1,x)	0.750	0.408	0.784	0.589	0.659	0.615	0.435	0.447	0.353	0.302	-0.084	-0.317
P-value corr	0.000**	0.025*	0.000**	0.000**	0.000**	0.000**	0.015*	0.011*	0.060	0.117	0.681	0.098

Notes: As per Notes to Table A1. 31.3% of test results are significant at the 1% level.

Table A6: $\varepsilon(t,x)$ Results by Age: Model M3B

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	-0.575	-0.420	-0.250	-0.508	-0.603	-0.644	-0.500	-0.087	-0.756	-0.921	-0.817	-0.559	-0.639
Variance	1.052	0.662	0.862	0.804	1.772	1.059	1.289	1.409	1.566	1.734	1.524	1.591	2.169
Skewness	-0.102	0.532	0.093	-1.237	-0.056	-0.562	-0.279	0.227	0.110	0.129	-0.061	0.174	1.221
Kurtosis	2.541	2.431	3.265	4.802	3.695	3.461	2.225	2.685	2.724	1.590	2.416	2.775	3.925
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.011*	0.019*	0.200	0.011*	0.037*	0.005**	0.042**	0.723	0.007	0.002	0.004	0.041*	0.045*
P-value VR test stat	0.787	0.227	0.696	0.540	0.025*	0.768	0.319	0.184	0.082	0.032*	0.103	0.072	0.002*
P-value JB test stat	0.882	0.483	0.949	0.009**	0.780	0.478	0.633	0.858	0.940	0.358	0.837	0.918	0.033*
Corr (t+1,t)	0.091	-0.060	-0.127	0.039	0.174	0.147	0.262	0.087	0.333	0.220	-0.037	0.613	0.788
P-value corr	0.673	0.780	0.553	0.855	0.408	0.488	0.201	0.685	0.093	0.291	0.862	0.000**	0.000**

Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	-0.085	-0.021	0.238	0.269	0.673	0.366	0.261	0.359	0.712	0.639	1.014	1.252	1.676
Variance	2.955	1.877	1.822	3.565	2.407	3.472	4.957	6.692	6.989	2.773	3.619	4.568	2.820
Skewness	-0.528	-1.192	-0.675	-2.285	-2.487	-1.593	-1.390	-1.097	-1.532	-0.882	-0.452	-0.554	-0.125
Kurtosis	3.989	4.618	3.349	11.108	11.909	7.911	6.446	5.518	8.029	3.941	2.468	2.743	2.224
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.811	0.941	0.396	0.492	0.044	0.346	0.571	0.503	0.200	0.073	0.016*	0.009**	0.000**
P-value VR test stat	0.000**	0.013*	0.019*	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.351	0.016*	0.378	0.000**	0.000**	0.000**	0.000**	0.004**	0.000**	0.136	0.577	0.524	0.718
Corr (t+1,t)	0.515	0.174	0.209	0.253	0.218	0.005	-0.191	0.135	0.073	0.486	0.744	0.560	0.545
P-value corr	0.003**	0.409	0.315	0.218	0.295	0.980	0.364	0.526	0.734	0.007**	0.000**	0.001**	0.001**

Notes: As per Notes to Table A2. 30.8% of test results are significant at the 1% level.

Model M5

Figure A4 and Tables A7 and A8 give the corresponding plots and standardised residual results for M5. These are very similar to those for M1, with 32.3% and 33.7% of test results respectively being significant at the 1% level.

Figure A4: Plots of $\varepsilon(t,x)$: Model M5

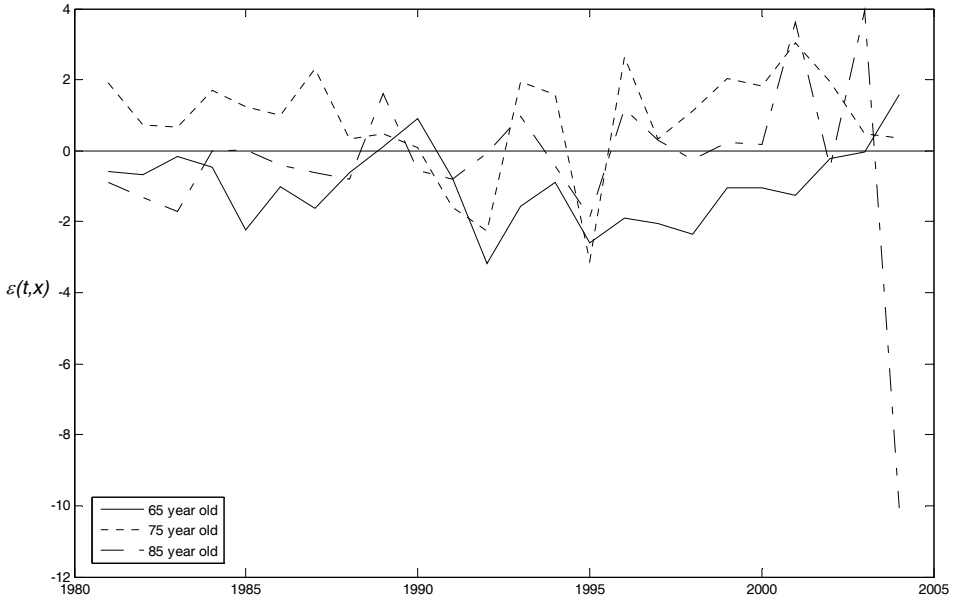


Table A7: $\varepsilon(t,x)$ Results by Year: Model M5

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	-0.172	-0.164	-0.163	-0.195	-0.194	-0.181	-0.155	-0.111	-0.027	-0.082	-0.020	-0.069
Variance	2.062	2.547	1.831	3.813	2.977	4.162	3.227	2.822	1.906	3.149	1.857	2.943
Skewness	-0.179	0.218	-0.387	-0.655	-0.432	-0.226	-0.069	-0.742	0.047	-0.799	0.788	0.026
Kurtosis	2.469	2.096	2.602	2.738	3.697	2.592	1.935	3.277	3.963	3.546	5.040	2.503
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.546	0.605	0.545	0.615	0.572	0.654	0.664	0.738	0.922	0.817	0.941	0.840
P-value VR test stat	0.003**	0.000**	0.014*	0.000**	0.000**	0.000**	0.000**	0.000**	0.008**	0.000**	0.011*	0.000**
P-value JB test stat	0.801	0.579	0.663	0.380	0.513	0.818	0.535	0.291	0.602	0.213	0.027	0.874
Corr (x+1,x)	0.622	0.582	0.540	0.338	0.506	0.411	0.433	0.272	0.028	-0.020	-0.258	0.048
P-value corr	0.000**	0.000**	0.001**	0.074	0.003**	0.023*	0.015*	0.162	0.893	0.923	0.188	0.816

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	-0.025	-0.082	-0.094	-0.080	-0.079	-0.088	-0.074	-0.089	-0.058	-0.064	-0.005	0.011
Variance	1.692	2.603	2.467	2.807	2.593	2.152	3.248	3.807	4.177	4.671	6.124	6.371
Skewness	-0.315	0.310	-0.597	0.201	0.074	0.542	-1.049	-0.212	-1.107	-2.217	-2.429	-2.617
Kurtosis	3.562	3.361	2.129	1.790	3.451	2.570	4.926	4.920	6.545	11.096	12.692	12.897
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.924	0.798	0.762	0.810	0.806	0.763	0.836	0.818	0.887	0.882	0.992	0.982
P-value VR test stat	0.033*	0.000**	0.000**	0.000**	0.000**	0.001**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.680	0.757	0.306	0.415	0.885	0.478	0.012*	0.123	0.000**	0.000**	0.000**	0.000**
Corr (x+1,x)	-0.005	-0.219	0.283	0.326	0.380	0.573	0.367	0.494	0.215	0.305	-0.281	-0.166
P-value corr	0.980	0.270	0.145	0.087	0.040*	0.000**	0.048*	0.004**	0.281	0.113	0.148	0.411

Notes: As per Notes to Table A1. 32.3% of test results are significant at the 1% level.

Table A8: $\varepsilon(t,x)$ Results by Age: Model M5

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	-1.304	-0.985	-1.325	-0.385	-0.506	-0.265	0.066	0.103	0.656	0.898	1.285	0.865	1.070
Variance	3.190	1.208	1.793	1.563	1.804	1.632	2.932	2.871	2.263	2.868	3.016	2.195	1.930
Skewness	0.564	0.188	0.133	-0.440	-0.720	-0.409	-0.693	-0.444	-0.386	-0.534	-0.300	-1.218	0.025
Kurtosis	2.686	3.302	3.050	2.909	4.489	3.000	4.438	2.207	3.827	3.423	3.381	4.583	3.737
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.002**	0.000**	0.000**	0.145	0.078	0.320	0.853	0.768	0.044	0.016	0.001	0.009	0.001
P-value VR test stat	0.000**	0.448	0.022*	0.084	0.021*	0.057	0.000**	0.000**	0.001**	0.000**	0.000**	0.002**	0.009**
P-value JB test stat	0.504	0.890	0.964	0.676	0.117	0.715	0.136	0.492	0.527	0.517	0.777	0.015**	0.761
Corr (t+1,t)	0.383	0.430	0.077	0.078	0.095	0.093	0.001	0.050	-0.105	-0.104	-0.078	0.033	0.213
P-value corr	0.047*	0.022*	0.718	0.715	0.657	0.665	0.996	0.817	0.623	0.628	0.715	0.877	0.307

Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	0.693	0.544	0.869	0.636	-0.258	0.007	-0.129	-0.236	-0.342	-0.419	-1.052	-1.313	-1.619
Variance	1.543	1.164	1.591	2.872	2.048	2.832	4.238	5.685	6.257	1.325	1.949	2.202	1.704
Skewness	-0.574	-0.455	-0.327	-2.113	-2.182	-2.812	-2.921	-2.840	-2.347	0.531	-0.035	0.782	0.080
Kurtosis	2.946	2.534	2.530	9.736	9.965	14.280	14.203	14.423	13.372	3.499	4.117	4.176	2.856
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.012*	0.021*	0.003**	0.079	0.385	0.984	0.762	0.633	0.510	0.088	0.001**	0.000**	0.000**
P-value VR test stat	0.093	0.531	0.072	0.000**	0.004**	0.000**	0.000**	0.000**	0.000**	0.272	0.008**	0.002**	0.038*
P-value JB test stat	0.517	0.593	0.722	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.502	0.535	0.147	0.977
Corr (t+1,t)	0.337	0.052	0.206	0.169	0.322	0.103	-0.314	-0.042	-0.510	-0.336	0.474	0.193	0.672
P-value corr	0.088	0.808	0.324	0.422	0.106	0.631	0.117	0.845	0.004**	0.090	0.009**	0.357	0.000**

Notes: As per Notes to Table A2. 33.7% of test results are significant at the 1% level.

Model M6

Figure A5 and Tables A9 and A10 give the corresponding plots and standardised residual results for M6. These are the best of the models considered so far, with 15.6% and 18.3% of results significant at the 1% level.

Figure A5: Plots of $\varepsilon(t,x)$: Model M6

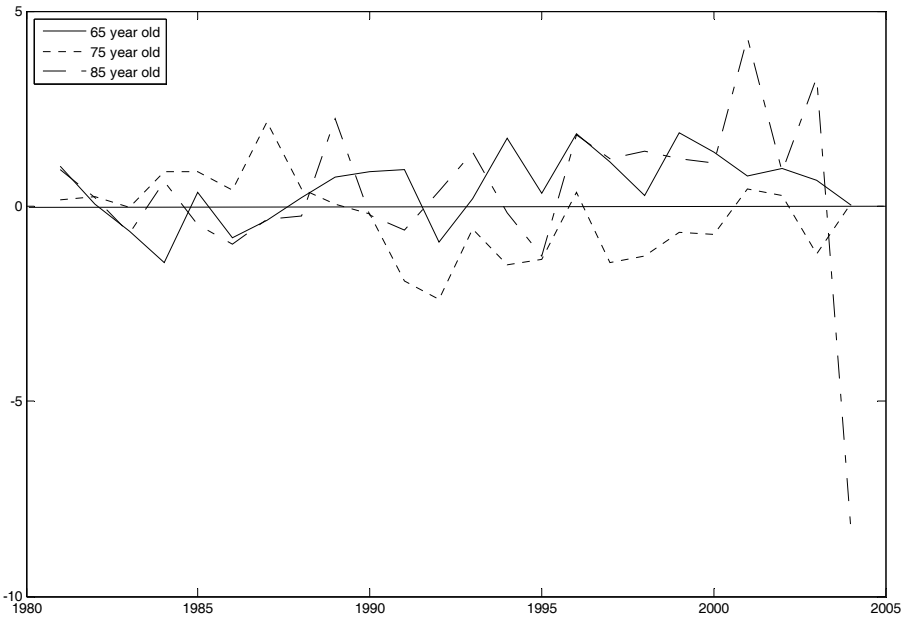


Table A9: $\varepsilon(t,x)$ Results by Year: Model M6

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	0.097	0.055	0.004	-0.055	-0.087	-0.097	-0.089	-0.062	0.003	-0.044	0.031	-0.012
Variance	1.558	1.281	1.091	1.523	1.436	1.848	1.548	1.561	0.925	1.496	0.998	0.657
Skewness	-0.606	0.136	0.105	0.314	-0.284	0.453	0.438	-0.676	0.118	-0.214	0.317	-0.882
Kurtosis	3.022	2.723	2.179	1.893	3.386	3.003	2.246	3.546	3.563	2.645	3.561	4.795
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.695	0.807	0.984	0.823	0.714	0.718	0.719	0.802	0.986	0.855	0.876	0.942
P-value VR test stat	0.075	0.315	0.684	0.091	0.146	0.012	0.079	0.073	0.858	0.106	0.931	0.196
P-value JB test stat	0.451	0.922	0.678	0.416	0.775	0.641	0.485	0.316	0.817	0.846	0.678	0.032
Corr (x+1,x)	0.530	0.390	0.216	-0.132	0.134	0.343	0.187	0.297	-0.009	0.111	-0.199	-0.313
P-value corr	0.001**	0.034	0.279	0.515	0.510	0.069	0.352	0.124	0.964	0.588	0.319	0.102

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	0.052	0.009	0.022	0.053	0.065	0.074	0.096	0.076	0.100	0.080	0.098	0.077
Variance	0.912	1.476	1.098	1.791	1.573	1.498	3.754	2.999	3.630	4.264	5.528	5.349
Skewness	0.253	-0.064	0.610	-1.579	-1.529	-0.766	-2.081	-2.122	-1.686	-2.289	-1.967	-1.596
Kurtosis	2.172	2.953	2.453	8.475	7.316	4.017	9.568	9.939	10.821	11.585	9.949	10.017
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.783	0.972	0.917	0.841	0.795	0.760	0.802	0.825	0.792	0.844	0.833	0.867
P-value VR test stat	0.821	0.118	0.667	0.018	0.068	0.105	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.601	0.990	0.380	0.000**	0.000**	0.160	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
Corr (x+1,x)	0.306	-0.093	0.330	0.141	0.193	0.238	0.267	0.262	0.125	0.141	-0.301	-0.293
P-value corr	0.111	0.650	0.082	0.489	0.335	0.228	0.172	0.181	0.539	0.489	0.118	0.129

Notes: As per Notes to Table A1. 15.6% of test results are significant at the 1% level.

Table A10: $\varepsilon(t,x)$ Results by Age: Model M6

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	0.460	0.466	-0.078	0.486	0.049	-0.009	0.005	-0.197	0.076	0.085	0.255	-0.289	-0.173
Variance	0.975	0.769	1.199	1.237	1.057	0.900	0.867	1.198	0.888	1.424	1.435	1.087	1.720
Skewness	-0.297	-0.341	-0.620	-0.783	-0.106	-0.878	-0.437	-0.784	-0.493	0.446	-0.069	-0.018	1.296
Kurtosis	2.823	2.758	2.329	2.776	3.275	3.939	2.793	4.454	3.117	2.030	2.432	3.074	4.366
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.032	0.016	0.730	0.043	0.817	0.962	0.980	0.388	0.698	0.731	0.308	0.188	0.524
P-value VR test stat	0.990	0.451	0.464	0.398	0.774	0.801	0.710	0.467	0.768	0.171	0.162	0.700	0.034
P-value JB test stat	0.825	0.770	0.370	0.286	0.941	0.138	0.668	0.102	0.610	0.420	0.843	0.997	0.014
Corr (t+1,t)	0.076	0.311	0.301	0.273	0.298	0.332	-0.264	-0.069	0.036	0.213	-0.160	0.463	0.708
P-value corr	0.723	0.120	0.134	0.180	0.139	0.094	0.197	0.750	0.866	0.308	0.448	0.011	0.000**

Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	-0.484	-0.556	-0.090	-0.194	-0.876	-0.352	-0.146	0.121	0.318	0.598	0.326	0.398	0.392
Variance	2.054	1.550	1.542	3.760	2.134	2.878	4.449	5.881	5.234	1.184	1.835	2.176	1.371
Skewness	-0.806	-1.254	-0.550	-1.852	-2.325	-2.594	-2.140	-1.845	-2.241	-0.675	0.012	0.164	0.688
Kurtosis	4.993	5.028	3.331	9.033	11.454	13.435	10.402	10.076	12.235	2.978	3.196	3.335	4.273
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.111	0.039	0.726	0.630	0.007	0.320	0.737	0.808	0.502	0.013	0.250	0.199	0.114
P-value VR test stat	0.004**	0.089	0.094	0.000**	0.002	0.000**	0.000**	0.000**	0.000**	0.493	0.017	0.002**	0.221
P-value JB test stat	0.037	0.006**	0.517	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.402	0.981	0.896	0.173
Corr (t+1,t)	0.275	-0.016	0.116	0.143	0.212	-0.127	-0.347	-0.088	-0.237	-0.039	0.475	0.308	0.554
P-value corr	0.177	0.940	0.587	0.502	0.308	0.550	0.078	0.682	0.252	0.855	0.009**	0.125	0.001**

Notes: As per Notes to Table A2. 18.3% of test results are significant at the 1% level.

Model M7

Figure A6 gives M7's $\varepsilon(t,x)$ plots. These are fairly similar to those of M1. However, the $\varepsilon(t,x)$ in Tables A11 and Table A12 are notably better: the percentages of test results that are significant at the 1% level (hence leading to the rejection of the iid N(0,1) null hypothesis) are a little under about half of what it was for M1, M3 and M5, and close in magnitude to those of M2B and M6.

Figure A6: Plots of $\varepsilon(t,x)$: Model M7

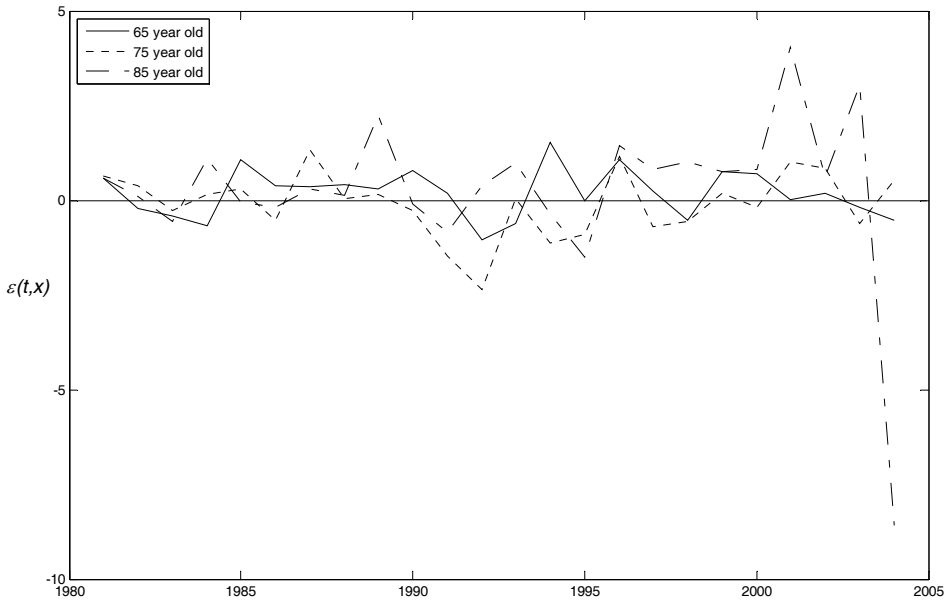


Table A11: $\varepsilon(t,x)$ Results by Year: Model M7

Year =	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Mean	0.013	0.011	0.010	0.012	-0.031	0.019	0.007	-0.008	0.004	-0.018	-0.002	-0.001
Variance	1.070	0.960	0.890	1.216	1.231	1.082	1.039	1.185	0.848	1.282	0.863	0.680
Skewness	-0.339	0.263	0.240	0.158	-0.160	0.275	-0.317	-1.032	0.523	-0.006	0.457	-1.001
Kurtosis	2.990	3.135	2.347	2.313	2.377	3.151	3.563	4.801	3.620	2.075	3.782	4.764
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.950	0.953	0.956	0.957	0.887	0.926	0.973	0.971	0.981	0.936	0.993	0.997
P-value VR test stat	0.738	0.962	0.759	0.420	0.393	0.707	0.817	0.476	0.637	0.314	0.680	0.237
P-value JB test stat	0.780	0.853	0.700	0.733	0.767	0.838	0.677	0.017*	0.449	0.629	0.457	0.021*
Corr (x+1,x)	0.315	0.227	0.049	-0.365	-0.035	-0.054	-0.217	0.064	-0.096	-0.005	-0.348	-0.280
P-value corr	0.100	0.253	0.812	0.050	0.865	0.794	0.276	0.755	0.640	0.981	0.064	0.149

Year =	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Mean	-0.008	-0.016	-0.009	-0.019	-0.018	-0.014	-0.022	-0.013	-0.003	-0.027	-0.015	-0.015
Variance	0.588	1.422	1.053	1.389	0.994	0.983	2.672	2.322	2.784	3.338	4.740	4.802
Skewness	0.051	-0.037	0.810	-1.076	-1.047	-0.141	-1.891	-1.829	-1.826	-2.825	-2.220	-2.086
Kurtosis	2.295	3.577	2.817	7.247	7.206	3.002	9.476	9.862	12.204	13.893	12.098	12.815
N	26	26	26	26	26	26	26	26	26	26	26	26
P-value mean test stat	0.956	0.946	0.964	0.934	0.927	0.944	0.947	0.966	0.993	0.940	0.973	0.973
P-value VR test stat	0.103	0.158	0.781	0.187	0.941	0.974	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
P-value JB test stat	0.760	0.832	0.237	0.000**	0.000**	0.958	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
Corr (x+1,x)	-0.017	-0.137	0.288	-0.072	-0.256	-0.141	-0.060	0.022	-0.140	-0.039	-0.485	-0.472
P-value corr	0.936	0.500	0.137	0.725	0.192	0.488	0.772	0.915	0.489	0.851	0.005**	0.007**

Notes: As per Notes to Table A1. 16.7% of test results are significant at the 1% level.

Table A12: $\varepsilon(t,x)$ Results by Age: Model M7

Current age =	64	65	66	67	68	69	70	71	72	73	74	75	76
Mean	0.187	0.182	-0.348	0.238	-0.161	-0.173	-0.100	-0.241	0.089	0.163	0.394	-0.091	0.077
Variance	0.628	0.399	0.702	0.759	0.936	0.757	0.892	1.267	0.746	1.201	1.190	0.749	0.783
Skewness	0.178	0.078	0.094	-0.820	0.387	-0.998	-0.204	-0.742	-0.391	0.782	0.087	-0.633	0.889
Kurtosis	2.157	2.636	2.152	2.885	3.149	5.351	2.586	3.866	3.372	2.590	2.537	3.765	3.537
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.259	0.171	0.054	0.194	0.424	0.340	0.607	0.305	0.617	0.474	0.090	0.611	0.673
P-value VR test stat	0.173	0.009**	0.302	0.427	0.903	0.423	0.778	0.352	0.398	0.460	0.480	0.403	0.486
P-value JB test stat	0.658	0.925	0.686	0.259	0.733	0.009**	0.844	0.228	0.687	0.271	0.885	0.335	0.178
Corr (t+1,t)	-0.124	-0.034	-0.193	-0.120	0.140	0.223	-0.264	-0.060	-0.201	0.177	-0.288	0.111	0.383
P-value corr	0.561	0.873	0.357	0.573	0.510	0.282	0.198	0.780	0.336	0.401	0.154	0.605	0.047**

Current age =	77	78	79	80	81	82	83	84	85	86	87	88	89
Mean	-0.193	-0.238	0.231	0.102	-0.601	-0.124	0.007	0.177	0.257	0.402	-0.024	-0.113	-0.273
Variance	1.245	1.284	1.306	2.812	1.973	2.911	4.402	5.633	4.963	1.106	1.016	1.120	0.572
Skewness	-1.474	-1.009	-0.242	-2.170	-1.912	-2.031	-1.845	-1.876	-2.616	-0.549	-0.038	0.252	-0.149
Kurtosis	7.383	4.323	2.776	10.325	9.619	10.383	9.421	10.703	14.187	2.664	3.068	2.781	3.004
N	24	24	24	24	24	24	24	24	24	24	24	24	24
P-value mean test stat	0.405	0.314	0.332	0.769	0.047*	0.726	0.987	0.718	0.578	0.074	0.908	0.606	0.090
P-value VR test stat	0.385	0.327	0.297	0.000**	0.007**	0.000**	0.000**	0.000**	0.000**	0.657	0.879	0.625	0.103
P-value JB test stat	0.000**	0.055	0.867	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.517	0.995	0.860	0.956
Corr (t+1,t)	-0.105	-0.137	-0.165	-0.079	0.091	-0.134	-0.370	-0.145	-0.345	-0.185	0.126	-0.205	0.259
P-value corr	0.622	0.520	0.435	0.714	0.670	0.530	0.056	0.494	0.080	0.379	0.554	0.327	0.206

Notes: As per Notes to Table A2. 14.4% of test results are significant at the 1% level.

References

Cairns, A. J. G., D. Blake and K. Dowd (2006) "A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration." *Journal of Risk and Insurance* 73: 687-718.

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

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein and M. Khalaf-Allah (2008) "Mortality density forecasts: An analysis of six stochastic mortality models." Pensions Institute Discussion Paper PI-0801, April.

Cochrane, J. H. (1988) "How big is the random walk in GNP?" *Journal of Political Economy*, 96: 893-920.

Continuous Mortality Investigation (2006) "Stochastic projection methodologies: Further progress and P-Spline model features, example results and implications." Working Paper 20.

Continuous Mortality Investigation (2007) "Stochastic projection methodologies: Lee-Carter model features, example results and implications." Working Paper 25.

Coughlan, G. D., D. Epstein, A. Ong, A. Sinha, I. Balevich, J. Hevia Portocarrera, E. Gingrich, M. Khalaf-Allah and P. Joseph, (2007). "LifeMetrics: A toolkit for measuring and managing longevity and mortality risks." Technical Document (JPMorgan, London, 13 March). Available at www.lifemetrics.com.

Currie, I. D., M. Durban and P. H. C. Eilers (2004) "Smoothing and forecasting mortality rates." *Statistical Modelling*, 4: 279-298.

Currie, I. D. (2006) "Smoothing and forecasting mortality rates with P-splines." Presentation to the Institute of Actuaries (www.ma.hw.ac.uk/~iain/research.talks.html).

Dowd, K. (2005) *Measuring Market Risk*. Second edition. Chichester and New York: John Wiley.

Dowd, K., Cairns, A. J. G., D. Blake, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah (2008) "Backtesting stochastic mortality models: An *ex-post* evaluation of multi-period-ahead density forecasts." Pensions Institute Discussion Paper PI-0803, March.

Jacobsen, R., Keiding, N., and Lynge, E. (2002) "Long-term mortality trends behind low life expectancy of Danish women." *J. Epidemiol. Community Health*, 56: 205-208.

Jarque, C., and A. Bera (1980) "Efficient tests for normality, homoscedasticity and serial independence of regression residuals." *Economics Letters* 6: 255-9.

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

Lee, R. D., and T. Miller (2001) "Evaluating the performance of the Lee-Carter method for forecasting mortality." *Demography* 38: 537-549.

Lo, A.W., and A.C. MacKinley (1988) "Stock prices do not follow random walks: Evidence based on a simple specification test." *Review of Financial Studies* 1: 41-66.

Lo, A.W., and A.C. MacKinley (1989) "The size and power of the variance ratio test in finite samples: A monte carlo investigation." *Journal of Econometrics* 40: 203-38.

Osmond, C. (1985) "Using age, period and cohort models to estimate future mortality rates." *International Journal of Epidemiology*, 14: 124-129.

Renshaw, A. E., and S. Haberman (2006) “A cohort-based extension to the Lee-Carter model for mortality reduction factors.” *Insurance: Mathematics and Economics* 38: 556-70.

Wilmoth, J.R. (1998) “Is the pace of Japanese mortality decline converging toward international trends?” *Population and Development Review* 24:593-600.