ASSESSING LONGEVITY RISK WITH GENERALIZED LINEAR ARRAY MODELS

by

Jillian M. Falkenberg

B.Sc.(Hons), Dalhousie University, 2007

A Project submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of
Statistics and Actuarial Science

© Jillian M. Falkenberg 2009
SIMON FRASER UNIVERSITY
Fall 2009

All rights reserved. However, in accordance with the Copyright Act of Canada, this work may be reproduced, without authorization, under the conditions for Fair Dealing. Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review, and news reporting is likely to be in accordance with the law, particularly if cited appropriately.
APPROVAL

Name: Jillian M. Falkenberg

Degree: Master of Science

Title of Project: Assessing Longevity Risk with Generalized Linear Array Models

Examinining Committee: Dr. Derek Bingham
Chair

________________________
Dr. Gary Parker
Senior Supervisor
Simon Fraser University

________________________
Dr. Yi Lu
Supervisor
Simon Fraser University

________________________
Dr. Thomas M. Loughin
External Examiner
Simon Fraser University

Date Approved: DEC 02 2009
Declaration of Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the “Institutional Repository” link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author’s written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, BC, Canada
Abstract

Longevity risk is becoming more important in the current economic environment; if mortality improvements are larger than expected, profits erode in the annuity business and in defined benefit pension schemes. The Lee-Carter model, although a popular model for mortality rates by age and calendar year, has been critiqued for its inflexibility. A recently proposed alternative is to smooth the mortality surface with a generalized linear array model (GLAM), allowing for an additive surface of shocks. We compare the GLAM and Lee-Carter models by fitting them to Swedish mortality data. Lee-Carter mortality predictions are calculated, and a time series method for GLAM prediction is developed. The predicted mortality rates and associated uncertainties are compared directly, and their impact on annuity pricing is analyzed. Letting future mortality be stochastic, we can calculate the expected value and variance of the present value for various annuities.

Keywords: Generalized linear array model; Lee-Carter model; mortality; annuity; projection; longevity risk
I would like to thank my supervisor Dr. Gary Parker. I feel so fortunate to have had the opportunity to work with him. I have learned a lot about Actuarial Science over the last two years. At times I felt discouraged and overwhelmed, but a visit to his office always seemed to turn everything around and renew my resolve. I would also like to thank Dr. Yi Lu, for teaching me so much in my first semester here at Simon Fraser University. Her understanding and kindness has made my time here an enjoyable experience.

In addition, I would like to thank my friends in the department for making me laugh so much, and being always willing to provide assistance when needed. Finally, a thank you to Matt Richard, for knowing the importance of a hug and a good meal.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approval</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Methods of Extrapolating</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Traditional Models</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Scenarios</td>
<td>2</td>
</tr>
<tr>
<td>1.1.3 Lee-Carter Model</td>
<td>3</td>
</tr>
<tr>
<td>1.1.4 Lee-Carter Extensions</td>
<td>4</td>
</tr>
<tr>
<td>1.1.5 Smoothing</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Outline</td>
<td>6</td>
</tr>
<tr>
<td><strong>2 B-splines</strong></td>
<td>7</td>
</tr>
<tr>
<td>2.1 Properties of B-splines</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Recursive Definition</td>
<td>8</td>
</tr>
<tr>
<td>2.3 B-splines for Smoothing</td>
<td>10</td>
</tr>
</tbody>
</table>
3 Generalized Linear Array Model (GLAM) 13
3.1 Model Equations ............................................. 14
3.2 GLAM Inputs .................................................. 14
3.3 Scoring Algorithm ........................................... 15
3.4 Definitions of Matrix Operations ......................... 15
  3.4.1 Kronecker Product ...................................... 15
  3.4.2 Row Tensor (G) .......................................... 16
  3.4.3 H-transform (H) .......................................... 17
  3.4.4 Rotation (R) .............................................. 17
  3.4.5 Rotated H-transform ($\rho$) ............................ 17
3.5 Efficient Evaluation of the Scoring Algorithm ............ 17
  3.5.1 Linear Functions ...................................... 18
  3.5.2 Inner Products ........................................ 18
  3.5.3 Diagonal Functions ..................................... 19
4 Two Dimensional GLAM 20
4.1 Basic Smooth Model ....................................... 21
  4.1.1 Penalized Scoring Algorithm .......................... 22
  4.1.2 Penalty Matrix ......................................... 22
  4.1.3 Efficient Evaluation of the Scoring Algorithm ..... 23
4.2 Additive Shock Model ..................................... 24
  4.2.1 Model Equation ........................................ 24
  4.2.2 Scoring Algorithm ..................................... 24
  4.2.3 Shock Penalty .......................................... 26
5 Lee-Carter Model 27
5.1 Lee-Carter Model ........................................... 27
  5.1.1 Model .................................................... 27
  5.1.2 Data ...................................................... 28
5.2 Estimating the Parameters ................................. 28
  5.2.1 Step 1 ................................................... 29
  5.2.2 Step 2 ................................................... 29
List of Tables

2.1 Data points for B-spline smoothing example . . . . . . . . . . . . . . . . . . 10

6.1 Optimal GLAM parameters with BIC . . . . . . . . . . . . . . . . . . . . . 34

8.1 Life annuity results. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70
List of Figures

2.1 Cubic B-Splines ................................................................. 9
2.2 Data points smoothed with cubic B-splines. .............................. 12
6.1 Raw Swedish mortality surface ............................................ 32
6.2 Mortality surface for smooth GLAM ..................................... 35
6.3 Smooth and shock surfaces for additive GLAM ....................... 36
6.4 Mortality surface for additive GLAM ..................................... 37
6.5 Optimal and adjusted \( k_t \) values for Lee-Carter model .......... 38
6.6 Mortality surface for Lee-Carter model ................................ 39
7.1 Differenced \( k_t \) values through time. ................................. 42
7.2 Fitted and predicted \( k_t \) through time ................................. 44
7.3 Raw and predicted Lee-Carter mortality surface ..................... 45
7.4 GLAM coefficient predictions: random walk with drift ............. 48
7.5 GLAM coefficients predictions: AR(1) with drift .................... 49
7.6 Raw and predicted GLAM mortality surface ........................... 50
7.7 Mortality Predictions Age 20-50 ........................................ 52
7.8 Mortality Predictions Age 60-90 ......................................... 53
7.9 Spread of Mortality Predictions Age 20-50 ............................ 54
7.10 Spread of Mortality Predictions Age 60-90 ........................... 55
8.1 Expected present value surface for a 10-year temporary life annuity ..... 62
8.2 Expected present value surface for a whole life annuity ............. 63
8.3 Expected present value surface for a 10-year guaranteed life annuity .... 64
8.4 Standard deviation surface for a 10-year temporary life annuity .... 66
8.5 Standard deviation surface for a whole life annuity .................. 67
8.6 Standard deviation surface for a 10-year guaranteed life annuity
Chapter 1

Introduction

In December 2003, Swiss Re issued the first mortality catastrophe bond (Cairns et al., 2006). This bond has cash flows linked to the realization of a composite mortality index. In November 2004, the European Investment Bank announced a longevity bond. This is a long term bond designed to be a hedge against a portfolio of annuities. It offers protection against realized mortality improvements exceeding anticipated levels of improvement. The cost of providing these new longevity bonds depends on the future rates of mortality. This is a similar case to pensions and life annuities.

Some work has been done in the early part of this century to assess the effect of mortality projections on life annuities. For example, the papers by Olivieri (2001) and Coppola et al. (2003) consider three future mortality tables, each with a probability of occurring; then, the authors evaluate the expectation and variance for the present value of a life annuity portfolio. Instead of limiting the examination to a few mortality scenarios, Brouhns et al. (2002b) use a Poisson log-bilinear model for mortality rates; then, Monte Carlo methods are used to find the density function for an annuity present value. What is made clear in all three of these papers, is that mortality projections are a necessary part of assessing the total risk of a life annuity.

With the recent macro-economic environment of low inflation and low interest rates, understanding longevity risk has become a more important part of assessing all of these products (Richards et al., 2006). Much work is being put into developing a quality mortality model, providing a best guess for future rates and more importantly an appropriate measure of uncertainty.

Although we may not be certain about the future of mortality rates, we can examine
trends in the past. The curve of death has changed shape over the last few decades; the concentration of deaths around the mode is increasing (rectangularization), and the mode is moving toward the older ages (expansion) (Pitacco et al., 2009, Section 4.1). In addition, mortality rates for each age have been decreasing through time with few exceptions, but there is some debate over the existence of a biological limit (Wong-Fupuy and Haberman, 2004). Some argue this limit exists, because life expectancy seems to follow the law of diminishing returns. But this slowing down is expected, because most improvements in developed countries are occurring in the advanced ages, which do not impact life expectancy as much as improvements in young ages (Wong-Fupuy and Haberman, 2004). The maximum attained age, on the other hand, keeps increasing with no sign of slowing down (Wong-Fupuy and Haberman, 2004).

1.1 Methods of Extrapolating

Models for extrapolating mortality rates could be based on biomedical processes, econometric relationships, or past mortality trends (Wong-Fupuy and Haberman, 2004). We will focus on trend models, using information from the past to hopefully predict the future.

1.1.1 Traditional Models

Traditional models treat the observed rates as the true rates, and extrapolations lack any statistical features, such as interval estimates (Pitacco et al., 2009). For instance, actuaries tend to multiply current rates by a reduction factor, and rarely supply a measure of uncertainty (Wong-Fupuy and Haberman, 2004). Chapter 4 of Pitacco et al. (2009) details some other traditional projection methods including exponential and polynomial extrapolation formulas, age shifting of the mortality table, and parametric methods. Another option the authors discuss is to interpolate between the current mortality values and some sort of theoretical limiting table. These methods only provide a best guess about the future; they do not measure the risk of mortality rates being different from those expected.

1.1.2 Scenarios

One way to assess mortality risk is to consider several scenarios for the future mortality rates (Pitacco et al., 2009, Section 1.5.4). Official projections by governmental agencies often
CHAPTER 1. INTRODUCTION

examine scenarios, incorporating expert opinion and an analysis by cause of death, however these projections have been shown to systematically underestimate mortality improvements (Wong-Fupuy and Haberman, 2004).

Scenarios are a very transparent way to provide some information about future outcomes. As a simple example, consider a set of three scenarios for the future rates of mortality: the most likely scenario, an optimistic scenario, and a pessimistic scenario. Then, conditional on each of these scenarios, the resulting distribution of the annuity present value, or any other quantity of interest, can be calculated. But, there are more than just these three scenarios to consider; there is a continuous range of available scenarios, and some scenarios are more likely than others. Therefore, a better option might be to find an unconditional distribution for the annuity price, incorporating the probabilities of each scenario.

1.1.3 Lee-Carter Model

If on the other hand, we think of the past rates as realizations from random variables, projections can rely on good statistical assumptions, and meaningful measures of uncertainty can be provided. In Lee and Carter (1992), the authors develop a stochastic model for projecting mortality rates, now referred to as the Lee-Carter model. The model has been used in a variety of actuarial and demographic applications including stochastic forecasts of the US social security finances and population projections for Scandinavia and the G7 countries (Li et al., 2009).

The model treats past rates as the observed outcomes of random variables, which we try to estimate in the future (Pitacco et al., 2009). Age and calendar year related parameters are derived from the first set of vectors using a singular value decomposition of the data matrix. There is only one time-related parameter, which can be modeled with an ARIMA time series model to project future rates. It is a pure extrapolation of past trends, not using any information about medical advances or environmental changes (Brouhns et al., 2002a). Methods of fitting and projecting the Lee-Carter model will be explored in Chapter 5.

The model, although popular, is not without criticism. One advantage is that the functional form of the mortality surface is rigidly defined, preventing undesirable behaviour such as mortality crossings for different ages. However, the restrictiveness of the model has also been seen to produce confidence intervals that are too narrow, possibly underestimating the risk of extreme outcomes (Li et al., 2009). The Lee-Carter model requires homoskedasticity, assuming that the errors are normally distributed around zero with constant variance. This
CHAPTER 1. INTRODUCTION

seems to be an unrealistic assumption (Brouhns et al., 2002a), because mortality has been shown to be more variable in the very young and very old (Cairns et al., 2009). This is most likely because the very old have fluctuations in the underlying rate, and the very young have a very small number of deaths on which to base the rate.

Furthermore, the Lee-Carter model is criticized for not including a cohort effect, for which there is strong evidence. In the eight stochastic mortality models examined by Cairns et al. (2008), the standardized residuals reveal distinct bands of non-independence in models that lack a cohort parameter. Cohorts often have similar life habits like smoking and diet, and the fetal and early life conditions experienced by a cohort can increase the incidence of chronic health conditions (Richards et al., 2006). For example, those born in England and Wales between 1925-1935 are sometimes referred to as a ‘golden cohort’ due to improved mortality rates experienced throughout their lifetimes (Cairns et al., 2009). The authors note that these improvements could be explained by a shortage of unhealthy food in the 1940s and early 1950s, along with the introduction of the National Health Service. To account for the cohort effect, Cairns et al. (2009) discuss some alternative multifactor mortality models to the Lee-Carter model, including the Renshaw-Haberman cohort model and the Cairns-Blake-Dowd model with a cohort effect, both developed in 2006. Other Age-Period-Cohort models have been proposed (Pitacco et al., 2009, Section 4.8.1, Chapter 6).

1.1.4 Lee-Carter Extensions

There are a number of models that build on the Lee-Carter framework in an attempt to improve it. For instance, Brouhns et al. (2002a) embed the Lee-Carter method in a Poisson regression model, switching from a classical linear model to a generalized linear model. This model is called a Poisson log-bilinear model (Pitacco et al., 2009). The number of deaths follows a Poisson distribution, a natural choice for modeling count data, and the mean parameter of the Poisson distribution is described by essentially the same equation as the Lee-Carter model. The parameters are determined by the maximum likelihood method. Li et al. (2009) extend this method even further, considering that individuals in the same age-period cell are usually not homogeneous due to unobservable characteristics. For mathematical tractability, the number of deaths is modeled by a Poisson-Gamma mixture, or equivalently a Negative Binomial distribution, to account for individual heterogeneity. Compared to the Lee-Carter model, the forecasts are nearly identical, but the confidence intervals are wider.
CHAPTER 1. INTRODUCTION

Also, the heterogeneous model generally has an improved goodness of fit, and retains the appealing features of the original model.

Renshaw and Haberman (2003) propose an extension to both the Lee-Carter and Poisson log-bilinear models. In the original parameterization of these two models, there are two parameters per age and one parameter per year. Renshaw and Haberman (2003) propose adding an extra age-year parameter product, such that there would be two time-related parameters to project mortality. The Lee-Carter model uses only the first vector set from singular value decomposition, but this two-factor Lee-Carter model would require the second vector set as well. The authors test the original one-factor models and the two-factor extensions on three test cases, but warn against drawing general conclusions. The paper by Cairns et al. (2006) introduces a two-factor model to better fit older ages; the authors state that the model they examine in this article could be considered a special case of the two-factor Lee-Carter model fitted by Renshaw and Haberman (2003).

1.1.5 Smoothing

An alternate approach to these models is to treat the age-period mortality data as a noisy surface to be smoothed. This approach avoids making strong assumptions about the functional form of the surface. For example, Currie et al. (2004) assume a generalized linear model (GLM) with Poisson errors. In the linear part, the authors use penalized basis spline (P-spline) methodology, introduced by Eilers and Marx (1996), to smooth over both the age and calendar year. To fit the model, the penalized likelihood is maximized, and projections are accomplished by treating future values as missing data. Richards et al. (2006) use these same smoothing and projection methods on both age-period and age-cohort mortality surfaces for comparison.

The difficulty with these projections is that they depend highly on the order of penalty chosen, so some judgment is required by the user. For their UK mortality data, Currie et al. (2004) choose a penalty of order 2, explaining that a penalty of order 1 is not consistent with the data, and a penalty of order 3 gives implausible values for projections longer than 50 years.

The smoothing problem can become very slow to implement if there are a large number of data points, or if this two-dimensional problem is increased to include more dimensions. Currie et al. (2006) address this problem in a general way by developing efficient algorithms to fit a generalized linear array model (GLAM). A GLAM is a GLM fitted to array data,
where the design matrix can be written in Kronecker product form, and all the data is kept in array form during calculation.

The smoothing problem addressed by Currie et al. (2004) can be considered as a two-dimensional GLAM. Therefore, Kirkby and Currie (2009) use GLAM methodology to fit the model suggested in Currie et al. (2004) to Swedish mortality data. The paper starts with the smooth model, but then adds a surface of period shocks for additional complexity.

1.2 Outline

Although there are several mortality models available, we will focus on two models in particular: the Lee-Carter model and the Generalized Linear Array Model. Using the Lee-Carter model is standard in academia and industry, so this model will be used as the comparison benchmark for the newly developed GLAM. By fitting both models to the same mortality data, developing projections, and pricing annuities, our goal is to expose the strengths and weaknesses of the two models.

We will proceed in the following way. First, some background on basis splines will be provided in Chapter 2. Then, some details about the general GLAM model and the methods used by Kirkby and Currie (2009) will be given in Chapters 3 and 4. As a means of comparison, we will also fit the classical Lee-Carter model, described in Chapter 5. Chapters 6 and 7 show the resulting fit and describe the prediction process for both models. Finally, we will examine an application to annuities in Chapter 8. Chapter 9 concludes by summarizing key results, commenting on the computation and projection of the GLAM and Lee-Carter models, and suggesting further work.
Chapter 2

B-splines

Before we introduce B-splines, we need to talk about what splines are in general. A spline is a function defined by piecewise polynomials, and these polynomials are joined at locations called knots. To be precise, if \( s(x) \) denotes the spline function and \( p_0, p_1, ..., p_m \) the polynomials, then

\[
s(x) = \begin{cases} 
  p_0(x) & k_0 \leq x < k_1 \\
  p_1(x) & k_1 \leq x < k_2 \\
  ... \\
  p_m(x) & k_m \leq x < k_{m+1} 
\end{cases},
\]

(2.1)

where \( k_0 \) and \( k_{m+1} \) are called outer knots, and \( k_1, ..., k_m \) are inner knots. Splines are named according to the maximum degree of the polynomials that define them. A spline of degree 0 is a step function, degree 1 is a linear spline, degree 2 is a quadratic spline, and degree 3 is a cubic spline. At any knot location \( k_h \), the smoothness of the spline can be defined by how many derivatives of the polynomials \( p_{h-1} \) and \( p_h \) are identical (Pitacco et al., 2009).

Basis splines, or B-splines, are a special variety of spline with some nice mathematical properties. For a complete exploration of B-splines, including the general recursive definition, see De Boor (1978) and Dierckx (1993); also, Eilers and Marx (1996) has an excellent summary. Both the degree and the knot locations for the B-spline are chosen in advance, and then the B-splines are calculated recursively from those of a lower degree. They are very attractive as base functions for univariate regression; for example, a linear combination of cubic B-splines gives a smooth curve (Eilers and Marx, 1996). Therefore, B-splines are often used as a regression basis for smoothing.
2.1 Properties of B-splines

B-splines can be characterized by the degree of their polynomial pieces. For example, B-splines of degree \( k \) will always be composed of \( k + 1 \) polynomial pieces that are joined at \( k \) inner knots. At the joining points, derivatives up to order \( k - 1 \) are continuous. Also, the B-spline will be positive on the domain spanned by \( k + 2 \) knots and zero elsewhere.

As an example, the top pane of Figure 2.1 shows a cubic B-spline, a B-spline where each piece is of degree 3. Knots were chosen to be the integers 1 through 5. Since the B-spline is of degree 3, it is composed of 4 polynomial pieces joined at the 3 inner knots. At the joining points, the first and second derivatives are continuous, so we get a smooth curve. Finally, the B-spline is positive over the range of 5 knots and zero elsewhere.

When we have more than \( k + 2 \) knots, a set of linearly independent B-splines will be created. Then, for any given \( x \) location, the set will have at most \( k + 1 \) B-splines that are non-zero. For example, see the bottom pane of Figure 2.1. In this figure, there are at most 4 non-zero B-splines for any given value of \( x \). Note also that between 4 and 7 on the x-axis, the B-splines will always sum to 1. This result holds true in general. If we number the knots sequentially from 1 to \( n \), smallest to largest, then the B-splines between knots \( k + 1 \) and \( n - k \) will always sum to 1.

2.2 Recursive Definition

The B-splines in Figure 2.1 we determined by a recursive definition. Let \( B_j(x; k) \) denote the value at \( x \) of the \( j^{th} \) B-spline with degree \( k \). Then, \( B_j(x; k) \) will be defined recursively on \( k \). For the purposes of this thesis, we will restrict our attention to splines with equally spaced knot locations. This restriction, although unnecessary, simplifies the formulas that follow.

Suppose we have \( n \) equally spaced knots such that the space between knots is \( dx = (x_{\text{max}} - x_{\text{min}})/n \). Also, let \( p = (x - x_{\text{min}})/dx \) be a relative measure of the location of \( x \). Then, we have the following recursion formula:

\[
B_j(x; k) = \frac{k + p - j + 1}{k} B_{j-1}(x; k - 1) + \frac{j - p}{k} B_j(x; k - 1),
\]

with starting point

\[
B_j(x; 0) = \begin{cases} 
1, & (j - 1)dx < x - x_{\text{min}} \leq j dx \\
0, & \text{all other } j 
\end{cases}.
\]
A Cubic B–Spline With Equally Spaced Knots

Overlapping Cubic B–Splines With Equally Spaced Knots

Figure 2.1: Cubic B-Splines
Note that when \( j < 0 \) or \( j > n \), \( B_j(x;0) \) is always 0.

These formulas are simple to program; for example, the statistical software R contains functions to calculate B-spline values. In the splines package, the function `splineDesign` (or the older version `spline.des`) calculates the B-splines at any \( x \), when supplied with a vector of knot locations and the degree of the splines.

### 2.3 B-splines for Smoothing

B-splines can be used for the smoothing of data. To illustrate this process, we will fit a smooth curve through the data points in Table 2.1, where \( x \) is the input vector, and \( y \) is the response vector. See Figure 2.2 for a plot of these 16 points.

The basic approach involves three steps. First, set up a B-spline basis according to a chosen degree and knot locations. Next, express each element of the input vector \( x \) in terms of this basis. Finally, regress these newly expressed inputs on the observed response \( y \). By re-expressing the predictors in terms of the B-spline basis, we create an interdependence among the points, resulting in a smooth fit. These steps will be detailed in the example that follows. Although this example only smooths over one dimension, the technique can be extended to two or more dimensions.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>1.0</td>
<td>5.2</td>
<td>3.6</td>
<td>6.4</td>
<td>6.2</td>
</tr>
<tr>
<td>4.2</td>
<td>1.6</td>
<td>5.4</td>
<td>2.9</td>
<td>6.6</td>
<td>5.8</td>
</tr>
<tr>
<td>4.4</td>
<td>1.3</td>
<td>5.6</td>
<td>3.1</td>
<td>6.8</td>
<td>6.0</td>
</tr>
<tr>
<td>4.6</td>
<td>2.0</td>
<td>5.8</td>
<td>2.8</td>
<td>7.0</td>
<td>2.0</td>
</tr>
<tr>
<td>4.8</td>
<td>3.3</td>
<td>6.0</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>4.5</td>
<td>6.2</td>
<td>5.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Data points for B-spline smoothing example.

First, we need a B-spline basis, which is a set of B-splines of degree \( k \). The knots that define the B-splines must cover the range of \( x \), with \( k \) extra knots on either side. The extra knots are required to avoid strange behaviour, because the B-spline basis is only useful over the range that the B-splines sum to 1. There is no correct answer for the total number of knots to use. The key is to strike a balance between overfitting with too many knots and underfitting with not enough. For our example, we wish to fit cubic B-splines, so \( k = 3 \). Knots are chosen at each integer from 1 to 10. This covers the range of the data from 4 to
7, with 3 extra knots on either side. The B-spline basis for this example consists of the 6 splines pictured in the bottom pane of Figure 2.1.

Next, we must make the B-spline matrix. This matrix restates each $x$ value from the data set in terms of the B-spline basis. If we let $B$ be the B-spline matrix, then the $i^{th}$ row of $B$ will correspond to a value $x_i$ from the data set, and the $j^{th}$ column will correspond to the $j^{th}$ B-spline. Then, the $(i, j)^{th}$ entry in $B$ is defined by

$$B(i, j) = B_j(x_i; k), \quad (2.4)$$

where $B_j(x_i; k)$ is the height of the B-spline given by (2.2) and (2.3). Recall that in our example $k=3$. Because there are 16 data points and 6 B-splines in our basis, $B$ will be of size $16 \times 6$.

The last step is to perform an ordinary least squares regression of the response vector $y, 16 \times 1$, on the matrix $B, 16 \times 6$,

$$y = B\theta, \quad (2.5)$$

where $\theta, 6 \times 1$, is a vector of coefficients. These coefficients are like multiplicative scalings for the B-splines. The scaled B-splines, when added together, give a smooth curve that approximates the data points. To be clear, if we let $f(x)$ be the smooth curve, then

$$f(x) = [B_1(x; 3) B_2(x; 3) \ldots B_6(x; 3)] \theta. \quad (2.6)$$

The resulting curve $f(x)$ can be seen in Figure 2.2.
Figure 2.2: Data points smoothed with cubic B-splines.
Chapter 3

Generalized Linear Array Model (GLAM)

A generalized linear array model (GLAM) is a special case of a generalized linear model (GLM) when the data is in array form, and the design matrix can be written as a Kronecker product. GLAMs can be applied to factorial designs, contingency tables, and the smoothing of data on multidimensional grids. The novelty of GLAMs is in the computational approach to estimation, originally proposed by Currie et al. (2006). During estimation, the data array and the coefficient array are left in their array forms instead of collapsing them into vectors. Also, the design matrix is never fully constructed, remaining in its Kronecker product form. Then, the necessary calculations are evaluated sequentially through nested matrix operations. For multidimensional, possibly large data sets, this approach allows high-speed and low storage calculation of parameter estimates and standard errors: calculations that were previously infeasible with standard GLM methodology.

In this section, we will first review the model equations used in GLMs, which are also applicable to GLAMs. Next, the input data and the scoring algorithm will be defined. Then, after defining some key matrix operations, we will review the necessary steps for efficiently fitting a GLAM.
3.1 Model Equations

In the following equations, let the columns of matrix $X$ be the independent variables, and the column vector $y$ be the dependent variable; each row of $X$ and $y$ represents an observation. Also, let the expected value of $y$ be the vector $\mu$.

As in GLMs (see Agresti, 2002), there are two components to a GLAM: a random component and a systematic component. The random component of a GLM assumes independent observations $y = (y_1, \ldots, y_n)$ come from a distribution, usually assumed to be in the natural exponential family. A distribution in this family would require a probability density function of the form

$$f(y_i; \phi_i) = a(\phi_i) b(y_i) \exp\left[ y_i Q(\phi_i) \right],$$

(3.1)

where $\phi_i$ is the distribution parameter, $a$ and $b$ are functions of only $\phi_i$ and $y_i$ respectively, and $Q(\phi_i)$ is called the natural parameter. Also, the GLM assumes a systematic component of the form

$$g(\mathbb{E}(y)) = g(\mu) = X\theta,$$

(3.2)

where $\theta$ is a parameter vector. Equation (3.2) creates a linear map of the explanatory variables $X$ to the expected observations $\mu$ by way of a link function $g(\cdot)$. A link function of the form $g(\mu) = \mu$ is called an identity link, and a function of the form $g(\mu) = \log(\mu)$ is called a log link.

Choosing both a distribution for the random component and a form for the link function defines what kind of model we are using. For example, when a normally distributed random component is combined with an identity link, we get ordinary linear regression. Also, a Poisson random component with a log link is called a loglinear model.

Keep in mind that the difference between GLMs and GLAMs is the way the computations are done. All of the results will be the same, equatable through rearrangement, but the GLAM methodology will arrive at them quicker.

3.2 GLAM Inputs

GLAMs require data to be in a specific form. The response data $Y$ must be a $d$-dimensional array of size $n_1 \times n_2 \times \ldots \times n_d$. Note that the corresponding vector arrangement $y$ would be of length $n = n_1 n_2 \ldots n_d$. In addition, explanatory variables must be arranged in a design
matrix of the form $X = X_d \otimes \ldots \otimes X_1$, where $\otimes$ denotes the Kronecker product (defined in Section 3.4.1). $X_i$ is an $n_i \times c_i$ matrix corresponding to the $i^{th}$ marginal variable.

### 3.3 Scoring Algorithm

The GLAM coefficient vector $\theta$ is found using the same equation as a standard GLM. This is achieved by repeatedly evaluating the scoring algorithm

$$X'\tilde{W}_\delta X\hat{\theta} = X'\tilde{W}_\delta \tilde{z},$$

(3.3)

where $W_\delta$ is a diagonal matrix of weights. $\hat{\theta}$ is the current estimate of $\theta$, and the tilde over $W_\delta$ and $z$ indicates the quantity should be calculated using the current approximate solution $\tilde{\theta}$. This is the same form as the solution to the normal equations using weighted least squares. The diagonal elements of $W_\delta$ are

$$w_{ii}^{-1} = (\partial g(\mu_i)/\partial \mu_i)^2 \text{var}(Y_i),$$

(3.4)

and the vector $z$ is

$$z = X\theta + W_\delta^{-1}(y - \mu).$$

(3.5)

Evaluating $(X'\tilde{W}_\delta X)^{-1}X'\tilde{W}_\delta \tilde{z}$ (a rearrangement of Equation (3.3)) gives the improved approximation $\hat{\theta}$. Reevaluation continues until a certain tolerance level is achieved, when the difference between $\tilde{\theta}$ and $\hat{\theta}$ becomes very small.

### 3.4 Definitions of Matrix Operations

The operations defined below are necessary for solving a GLAM in $d$-dimensions. Once these basic operations are covered, we can start going through the details of calculating the matrix products in the scoring algorithm.

#### 3.4.1 Kronecker Product

A Kronecker product is a multiplicative matrix operation that combines two matrices of arbitrary dimension. Starting with a small example, and letting $\otimes$ denote the Kronecker product,
Note that each block of the resulting matrix is the product of one element of $A$ with the entire matrix $B$. To generalize, let matrix $A$ be of size $n_1 \times c_1$ and matrix $B$ be of size $n_2 \times c_2$. Then, the Kronecker product of $A$ and $B$

$$A \otimes B = \begin{bmatrix}
\begin{array}{cccc}
a_{11}b_{11} & \cdots & a_{11}b_{1c_2} & \cdots & a_{1c_1}b_{11} & \cdots & a_{1c_1}b_{1c_2} \\
a_{11}b_{21} & \cdots & a_{11}b_{2c_2} & \cdots & a_{1c_1}b_{21} & \cdots & a_{1c_1}b_{2c_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
a_{11}b_{n_{12}} & \cdots & a_{11}b_{n_{2c_2}} & \cdots & a_{1c_1}b_{n_{12}} & \cdots & a_{1c_1}b_{n_{2c_2}} \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
a_{n_{11}}b_{11} & \cdots & a_{n_{11}}b_{1c_2} & \cdots & a_{n_{1c_1}}b_{11} & \cdots & a_{n_{1c_1}}b_{1c_2} \\
a_{n_{11}}b_{21} & \cdots & a_{n_{11}}b_{2c_2} & \cdots & a_{n_{1c_1}}b_{21} & \cdots & a_{n_{1c_1}}b_{2c_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
a_{n_{11}}b_{n_{12}} & \cdots & a_{n_{11}}b_{n_{2c_2}} & \cdots & a_{n_{1c_1}}b_{n_{12}} & \cdots & a_{n_{1c_1}}b_{n_{2c_2}}
\end{array}
\end{bmatrix} \quad (3.6)$$

becomes a block matrix of size $n_1 n_2 \times c_1 c_2$.

### 3.4.2 Row Tensor ($G$)

Let $X$ be a $n_1 \times c_1$ matrix. Also, let $1_c$ be a $c \times 1$ vector of ones. Then, the row tensor of $X$ is defined as

$$G(X) = (X \otimes 1_{c_1}') * (1_{c_1}' \otimes X), \quad (3.7)$$

where $*$ denotes element-by-element multiplication. Furthermore, let $Y$ be a matrix of size $n_2 \times c_2$. The row tensor of $X$ and $Y$ is defined as

$$G_2(X, Y) = (X \otimes 1_{c_2}') * (1_{c_1}' \otimes Y). \quad (3.8)$$

Note that $G_2(X, X) = G(X)$, meaning that (3.7) is a special case of (3.8).
3.4.3 H-transform (H)

The H-transform is a general form of pre-multiplication for arrays. Let

- $A$ be a d-dimensional array of size $c_1 \times c_2 \times ... \times c_d$
- $X$ be a matrix of size $n_1 \times c_1$
- $A^*$ be the matrix of size $c_1 \times c_2c_3...c_d$, a matrix version of $A$, where dimensions 2 through $d$ have been flattened.

Then, the H-transform of $A$ by $X$, denoted $H(X, A)$, is the product $XA^*$, $n_1 \times c_2c_3...c_d$, when the flattened dimensions are reinstated to achieve a d-dimensional array of size $n_1 \times c_2 \times ... \times c_d$.

In one dimension, $A$ is a column vector, and in two dimensions, $A$ is a matrix. In both of these cases, $A^*$ is exactly $A$, so the conventional pre-multiplication applies.

3.4.4 Rotation (R)

Rotation is a general form of transpose for arrays. Again, let $A$ be a d-dimensional array of size $c_1 \times c_2 \times ... \times c_d$. Then, the rotation of $A$, denoted $R(A)$, permutes the indices to achieve an array of size $c_2 \times c_3 \times ... \times c_d \times c_1$.

3.4.5 Rotated H-transform ($\rho$)

The rotated H-transform is a nested operation combining the two definitions above. Let array $A$ and matrix $X$ be defined as above. Then, the rotated H-transform of $A$ by $X$ is

$$\rho(X, A) = R(H(X, A)),$$

with dimension $c_2 \times ... \times c_d \times n_1$.

3.5 Efficient Evaluation of the Scoring Algorithm

When evaluating the scoring algorithm to solve a GLAM, we need to calculate linear functions such as $X\theta$ and inner products such as $X'\hat{W}_\delta X$. Normally, the matrix multiplications would simply be carried out. But, when $X$ gets large, the number of operations becomes
plentiful and very time consuming. Thankfully, we can exploit the Kronecker product form of $X$ to perform the matrix multiplications efficiently.

The necessary operations can be grouped into three categories: linear functions, inner products, and diagonal functions. In each of these categories, there is a main equation. The left hand side of this equation is the desired quantity in K-form (Kronecker product form). This is the original formulation of the model where direct calculation would be time consuming. The right hand side is a nested representation in A-form (array form). It operates on each dimension in turn, taking advantage of the Kronecker form of the design matrix. We will need the matrix operations defined in Section 3.4. Although the results on either side of the equation will have different dimensions, they have the same elements. Therefore, the right hand side can be made equivalent to the left hand side through rearrangement and redimensioning, which is a very efficient operation.

3.5.1 Linear Functions

Let $\Theta$ be a coefficient array of size $c_1 \times \ldots \times c_d$ such that $\theta = \text{vec}(\Theta)$, $c_1c_2\ldots c_d$, where the $\text{vec}$ operator collapses the dimensions of an array. A linear function such as $X\theta = (X_d \otimes \ldots \otimes X_1)\ theta$ can be evaluated as

$$X\theta, n_1 \ldots n_d \times 1 \equiv \rho[X_d, \ldots, \rho[X_2, \rho[X_1, \Theta]]], n_1 \times \ldots \times n_d, \quad (3.10)$$

where $\rho$ is the rotated H-transform (3.9), and '$\equiv$' means that the two sides contain a rearrangement of the same elements.

3.5.2 Inner Products

Let $W$ be an array of size $n_1 \times \ldots \times n_d$ such that $\text{diag}(W_\delta) = \text{vec}(W)$. An inner product like $X'W_\delta X$ is then calculated by rearranging its elements as

$$X'W_\delta X, c_1 \ldots c_d \times c_1 \ldots c_d \equiv \rho[G(X_d)', \ldots, \rho[G(X_2)', \rho[G(X_1)', W]]], c_1^2 \times \ldots \times c_d^2, \quad (3.11)$$

where $\rho$ is the rotated H-transform (3.9), $G$ is the row tensor function (3.7).
3.5.3 Diagonal Functions

Let the square matrix \( S_m = (X' \hat{W}_\delta X)^{-1}, \) \( c_1 c_2 \ldots c_d \times c_1 c_2 \ldots c_d, \) be a rearrangement of the array \( S, \) \( c_1^2 \times \ldots \times c_d^2. \) For details of this rearrangement, see Currie et al. (2006). Since \( var(\hat{X} \theta) = X S_m X', \) the variances for the predicted values \( \hat{Y} \) are

\[
\text{diag}(var(\hat{X} \theta)), n_1 \ldots n_d \times 1 \equiv \rho[G(X_d), ..., \rho[G(X_2), \rho[G(X_1), S]...], n_1 \times \ldots \times n_d. \tag{3.12}
\]
Chapter 4

Two Dimensional GLAM

Having introduced generalized linear array models in Chapter 3, we will now examine a specific application: modelling Swedish mortality data. Note that this chapter follows and reiterates the work done by Kirkby and Currie (2009). In the mortality data, we have two predictive factors: age and year, and two sets of counts: death and exposure. Dividing deaths by the corresponding exposure gives a rough mortality surface, where each combination of age and year has an associated probability of death. Assuming that this rough surface is a smooth probability surface with added noise, multidimensional smoothing is an appropriate goal.

Multidimensional smoothing of this two dimensional array data can be accomplished with a GLAM that has a model matrix of B-splines. The idea is to smooth the mortality surface over each predictive factor. Suppose our data matrix consists of \( n_y \) calendar years and \( n_a \) ages. Also, we choose to use a basis of B-splines order \( k \) with \( c_y + k + 1 \) knots in the year direction and \( c_a + k + 1 \) knots in the age direction. We have defined the number of knots in this way, because \( k + 2 \) knots give one B-spline of order \( k \), and in general, \( c_y + k + 1 \) knots result in \( c_y \) B-splines of order \( k \). Therefore, the B-spline matrix \( B_y, n_y \times c_y \), will smooth in the time direction, and \( B_a, n_a \times c_a \), will smooth in the age direction. The design matrix \( B \) (taking the place of \( X \) in Chapter 3) will be defined as the Kronecker product

\[
B = B_y \otimes B_a, n_y n_a \times c_y c_a. \tag{4.1}
\]

We will use the GLAM methodology with this design matrix to fit a Poisson loglinear model, a GLM commonly used for count data.
4.1 Basic Smooth Model

As discussed in Chapter 2, the B-spline basis, and hence the B-spline matrix, depends on the degree of the polynomials that define them and the location of the knots. One way of controlling the smoothness of the fitted model is to adjust the frequency of knots; more knots allow more variability in the fitted model, and less knots result in something smoother. Instead, Eilers and Marx (1996) suggested that one could choose a relatively large number of knots, but apply a roughness penalty to the regression coefficients, reducing the log-likelihood if there are large fluctuations in these parameters. This is called the method of penalized splines or P-splines. With this method, the smoothness of the fit relies on a weight parameter of the penalty term, and not the frequency of the knots. To choose the number of knots, Currie et al. (2004) suggest that for equally spaced data, one knot every four or five observations up to a maximum of 40 knots is often sufficient; the authors also suggest to use cubic splines and a quadratic penalty term. The basic smooth model, presented in this section, is a GLAM that fits a smooth mortality surface using the method of P-splines. In later sections, we will add additional levels of complexity to this basic model.

First, some definitions are in order. The matrices $E$ and $D$, both $n_a \times n_y$, represent the data collected, with ages in rows and calendar years in columns. The $(x, t)^{th}$ entry of the exposure matrix $E$ contains the total person-years lived in the age interval $[x, x+1)$ during calendar year $t$. The death matrix $D$ contains the number of deaths aged $x$ in calendar year $t$. Then, the model is

$$\log(\mathbb{E}(vec(D))) = \log(\mu) = \log(vec(E)) + B\theta. \tag{4.2}$$

Here, $\mathbb{E}$ represents the expected value, $vec$ rearranges $E$ and $D$ to be vectors size $n_a n_y \times 1$, $B$ is the Kronecker product of B-spline matrices (4.1), and $\theta$, $c_a c_y \times 1$, is a vector of coefficients. The observed deaths for each age and year will follow a Poisson distribution:

$$vec(D) \sim \text{Pois}(\mu). \tag{4.3}$$

This model is called a Poisson loglinear model, due to its Poisson distributed random component and log link function. More specifically, due to the $\log(vec(E))$ term, it is called a Poisson Rate Regression or a Poisson regression with offsets (Agresti, 2002). Note that one could analyze the data by cohort year instead of calendar year; Richards et al. (2006) make a comparison of the two approaches.
4.1.1 Penalized Scoring Algorithm

The penalized scoring algorithm will be defined in much the same way as the scoring algorithm in Section 3.3. The only difference is the introduction of a matrix $P$ to penalize roughness in the coefficients and consequently the fitted surface. So, the best estimate for the vector of coefficients $\theta$ is achieved by repeatedly evaluating the scoring algorithm

$$(B'\tilde{W}_\delta B + P)\hat{\theta} = B'\tilde{W}_\delta \tilde{z},$$  \hspace{1cm}(4.4)$$

where $W_\delta$, $n_an_y \times n_an_y$, is a diagonal matrix of weights. The diagonal elements of $W_\delta$ are

$$w_{ii} = \mu_i, \ i = 1,2,...,n_yn_y,$$  \hspace{1cm}(4.5)$$

and the vector $\tilde{z}$, $n_an_y \times 1$, is

$$\tilde{z} = B\tilde{\theta} + \tilde{W}_\delta^{-1}(vec(D) - \tilde{\mu}).$$  \hspace{1cm}(4.6)$$

Finally, the penalty matrix is chosen to be of the form

$$P = \lambda D_2' D_2, c_ay \times c_ay$$  \hspace{1cm}(4.7)$$

where $D_2$, $(c_ay - 2) \times c_ay$, is a second order difference matrix and $\lambda$ is a constant. Although we have chosen to use a second order difference matrix, any order may be chosen at the discretion of the user. This is the method of P-splines, introduced by Eilers and Marx (1996), where penalties are imposed on the differences between adjacent coefficients.

4.1.2 Penalty Matrix

We now discuss the components of the penalty matrix in more detail. First, the second order difference matrix $D_2$ takes on a very specific form, the construction of which is best explained through example. Suppose we want to construct a $D_2$ matrix with 5 columns. First, we would start with an identity matrix of size 5. This identity matrix is first differenced by taking each row i+1 and subtracting row i, which results in the following matrix

$$
\begin{bmatrix}
-1 & 1 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & -1 & 1 \\
\end{bmatrix}.
$$  \hspace{1cm}(4.8)$$
By differencing the matrix again, we end up with the desired $D_2$ matrix
\[
\begin{bmatrix}
1 & -2 & 1 & 0 & 0 \\
0 & 1 & -2 & 1 & 0 \\
0 & 0 & 1 & -2 & 1
\end{bmatrix}.
\] (4.9)

This construction method can be applied to make a $D_2$ matrix of any size. For our penalty matrix $P$, $D_2$ must have the same number of columns as the length of $\theta$.

The second component of the penalty matrix is the constant $\lambda$. By changing $\lambda$ we can control the balance of fit and smoothness. If we let $\lambda = 0$, the penalty disappears, and we are left with a basic B-spline regression. We have a good fit, but the regression surface may not be very smooth, because we have used a relatively high frequency of knots. For the interested reader, Figure 13 of Richards et al. (2006) plots a comparison of a B-spline fit and a P-spline fit to age 65 UK mortality data by year. On the other hand, if we let $\lambda = \infty$, we have a linear regression. The surface has the maximum amount of smoothness, but may not be a good fit. The key is to choose a reasonable $\lambda$ to balance the two desirable criteria.

### 4.1.3 Efficient Evaluation of the Scoring Algorithm

As in Section 3.5, to evaluate the scoring algorithm, we wish to obtain the desired K-form quantities by working with their corresponding A-forms for efficient calculation. The general formulas simplify nicely in the 2-dimensional case.

Let the B-spline age matrix $B_a$ be of size $n_a \times c_a$, and the B-spline year matrix $B_y$ be of size $n_y \times c_y$. Recall that $B = B_y \otimes B_a$, $\theta = \text{vec}(\Theta)$, and $\text{diag}(W_\delta) = \text{vec}(W)$. Then, the key equivalences we need to know are
\[
B\theta, n_an_y \times 1 \equiv B_a\Theta B_y', n_a \times n_y
\]
(4.10)
\[
B'W_\delta B, c_ac_y \times c_ac_y \equiv G(B_a)'WG(B_y), c_a^2 \times c_y^2
\]
(4.11)
\[
\text{diag}(\text{var}(B\hat{\theta})) = \text{diag}(BS_mB'), n_an_y \times 1 \equiv G(B_a)SG(B_y), n_a \times n_y;
\]
(4.12)
the two dimensional equivalents of Equations (3.10), (3.11), and (3.12). Also necessary is to redefine the penalty matrix to exploit the Kronecker form of the model matrix
\[
P = \lambda_a I_{c_y} \otimes D_a'D_a + \lambda_y D_y'D_y \otimes I_{c_a}, c_ac_y \times c_ac_y,
\]
(4.13)
where $\lambda_a$ and $\lambda_y$ are constants, $I_{c_a}$ and $I_{c_y}$ are square identity matrices, and $D_a$ and $D_y$ are second order difference matrices with $c_a$ and $c_y$ columns respectively. We now have everything we need to fit the basic smooth model.
4.2 Additive Shock Model

Sometimes current events such as war or disease can cause temporary shocks to the mortality surface. For example, in 1918 the Spanish influenza epidemic caused an unusually large, temporary spike in Swedish mortality. Therefore, we can improve the basic smooth model discussed in Section 4.1 by considering an additive GLAM. The additive shock model has two components, a smooth surface and a shock surface, which add together to give the total fitted surface. We will only consider period shocks, shocks to a given year, as these make the most intuitive sense.

4.2.1 Model Equation

The shock model requires only slight modifications to the smooth model. First, the main model equation from the smooth model (4.2) now has an additional term $\tilde{B}\tilde{\theta}$

$$
\log(E(\text{vec}(D))) = \log(\mu) = \log(\text{vec}(E)) + B\theta + \tilde{B}\tilde{\theta},
$$

where

$$
\tilde{B} = I_{n_y} \otimes \tilde{B}_a, n_a n_y \times c_s n_y.
$$

The B-spline matrix $\tilde{B}_a, n_a \times c_s$, smooths the period shocks in the age direction, and $\tilde{\theta}$, $c_s n_y \times 1$, is a vector of shock surface coefficients. As in the smooth model, the quantities above are not calculated directly, but through more efficient methods. Letting $\tilde{\theta} = \text{vec}(\tilde{\Theta})$, the linear equation (4.10) becomes

$$
B\theta + \tilde{B}\tilde{\theta}, n_a n_y \times 1 \equiv B_a\Theta B'_y + \tilde{B}_a\tilde{\Theta}, n_a \times n_y.
$$

As the name implies, the additive shock model simply adds on an extra term to account for the period shocks.

4.2.2 Scoring Algorithm

The next difference from the smooth model is found in the scoring algorithm. It is of the same form as Equation (4.4), but the following replacements are made:

$$
B = \begin{bmatrix} B & \tilde{B} \end{bmatrix}, \quad P = \begin{bmatrix} P & 0 \\ 0 & \tilde{P} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta \\ \tilde{\theta} \end{bmatrix},
$$

(4.17)
and the shock penalty matrix is
\[
\tilde{P} = \lambda_s I_{n_y} \otimes I_{c_y},
\]  
(4.18)
where \(\lambda_s\) is a constant, and \(I_{n_y}\) and \(I_{c_y}\) are square identity matrices. When the substitutions from (4.17) are made into (4.4), the scoring algorithm becomes:
\[
\begin{bmatrix}
B'\tilde{W}_\delta \tilde{B} + P & B'\tilde{W}_\delta \tilde{B} \\
\tilde{B}'\tilde{W}_\delta \tilde{B} & \tilde{B}'\tilde{W}_\delta \tilde{B} + \tilde{P}
\end{bmatrix}
\begin{bmatrix}
\tilde{\theta} \\
\tilde{\theta}
\end{bmatrix}
= 
\begin{bmatrix}
B \\
\tilde{B}
\end{bmatrix}
\tilde{W}_\delta \tilde{z},
\]  
(4.19)
where \(\tilde{z}\) is
\[
\tilde{z} = B\hat{\theta} + \tilde{B}\tilde{\theta} + \tilde{W}_\delta^{-1}(\text{vec}(D) - \hat{\mu}),
\]  
(4.20)
the result from substituting (4.17) into (4.6). Although the right hand side of the equation can be calculated with techniques already discussed, the pre-multiplying matrix on the left hand side
\[
\begin{bmatrix}
B'\tilde{W}_\delta \tilde{B} + P & B'\tilde{W}_\delta \tilde{B} \\
\tilde{B}'\tilde{W}_\delta \tilde{B} & \tilde{B}'\tilde{W}_\delta \tilde{B} + \tilde{P}
\end{bmatrix}
\]  
(4.21)
requires some extra attention.

We are familiar with the upper left entry, \(B'\tilde{W}_\delta \tilde{B} + P\), because this term was calculated efficiently for the smooth model using Equation (4.11). The other three entries use the new shock terms. Note that the upper right entry is simply the transpose of the lower left one. Efficient calculation of these matrix entries follows the same general pattern. For example, the inner product equation for the lower left entry is calculated using
\[
\tilde{B}'\tilde{W}_\delta \tilde{B}, c_y n_y \times c_a c_y \equiv G_2(B_y, I_{n_y})' W G_2(B_a, \tilde{B}_a), c_y n_y \times c_a c_a,
\]  
(4.22)
where \(G_2\) is the row tensor operator defined in Equation (3.8). Although (4.22) applies specifically to the lower left entry, the other matrix entries can be calculated in a similar way, noting that \(G(B, B) = G(B)\). Note that we are using the code developed by Kirkby and Currie (2009), which uses a few tricks to simplify these matrix entries even further and provide additional computational savings. These tricks, although we will not discuss them here, are detailed in the appendix of Kirkby and Currie (2009). Finally, the inverse of (4.21) can be calculated using
\[
\begin{bmatrix}
A & B \\
B' & C
\end{bmatrix}^{-1} = 
\begin{bmatrix}
X & -XBC^{-1} \\
-C^{-1}B'X & C^{-1} + C^{-1}B'XBC^{-1}
\end{bmatrix},
\]  
(4.23)
where \(X = (A - BC^{-1}B')^{-1}\), a short cut to find the inverse of a partition matrix when \(C\) is block-diagonal.
4.2.3 Shock Penalty

When the shock penalty matrix $\tilde{P}$ was defined in Equation (4.18), we did not specify the form of the shrinkage parameter $\lambda_s$. This is because the form of $\lambda_s$ determines the type of shock model we are using. If we let $\lambda_s$ be a constant $\lambda_s$, then we have a \textit{simple shock model}. This model shrinks each shock toward the smooth surface in the same way, regardless of the shock size. Like Kirkby and Currie (2009), we prefer a model that keeps the large shocks, and smooths out the small ones. Therefore, a scaled shock model is proposed.

A scaled shock model shrinks each shock by a different amount, allowing a systematic way of keeping the large shocks and shrinking the small ones. $\lambda_s$ will be a vector whose values are determined in relation to shock size. We define the shock size for each year to be the mean absolute value of the column of $\tilde{B}_a\tilde{\Theta}$ associated with that year. The mean is taken to get an average level of the shock in each year, and the absolute value to consider negative shocks just as severe. Note that the simple shock model must be fit first to know what the columns of $\tilde{B}_a\tilde{\Theta}$ should be.

Having defined shock size, we can now talk about each entry in $\lambda_s$. The largest shock in the Swedish data is for the year 1918, so we will use this as a comparison point for all the other shocks. So we let $\lambda_s$ be a vector whose $i^{th}$ entry is

$$\lambda_s \left( \frac{\text{shock size for 1918}}{\text{shock size for year } i} \right)^{\alpha}$$

where $\lambda_s$ is a constant that adjusts the overall level of scaling. If we let $\alpha = 0$ the model reduces to the simple shock model. If $\alpha = 1$, we have what is called a \textit{simple scaled shock model}. If $\alpha$ is left as a free parameter to optimize over, we get an \textit{optimal scaled shock model}. 

Chapter 5

Lee-Carter Model

5.1 Lee-Carter Model

The Lee-Carter model, a now popular method of modeling mortality, was first introduced by Lee and Carter (1992). Their idea was to pick a model with only one time-related parameter to forecast mortality rates for all ages. After examining a variety of models (such as the Brass logit model with fixed Beta, Coale-Demeny Model West Female, etc.) the authors settled upon what we know now as the Lee-Carter model.

The Lee-Carter model has a few key advantages. One advantage is its parsimonious nature; there are relatively few parameters to determine. Having few parameters implies restrictions, but these lead to another advantage: the model preserves the current age pattern in future mortality. It is generally assumed, in adult ages, that an older person should have a higher level of mortality than a younger person (although this may be violated by data from small populations). The exception to this monotonicity is a hump around age 20: a temporary elevation in the mortality rates due to higher accidental death rates in this age group. The small number of parameters prevent mortality crossovers that may occur in other stochastic models (Li et al., 2009).

5.1.1 Model

Let $m_{x,t}$ be the central death rate for someone age $x$ at time $t$. Then,

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t},$$

(5.1)

where
CHAPTER 5. LEE-CARTER MODEL

$a_x$: age specific parameters that indicate the average level of mortality

$b_x$: age specific parameters that indicate the responsiveness of each age to the time parameter. Some ages decline in mortality quicker than others.

$k_t$: time specific parameters that are an index of the level of mortality

$\varepsilon_{x,t}$: error terms with mean 0 and variance $\sigma^2$

Unfortunately, if we were to estimate the parameters above, we would find that $a_x$, $b_x$, and $k_t$ are only determined up to a linear transformation. Therefore, we require two additional restrictions for a unique solution: $\sum x b_x = 1$ and $\sum t k_t = 0$. These restrictions were chosen in particular, because they force $a_x$ to be the averages over time of $\ln(m_{x,t})$.

5.1.2 Data

Suppose we have the same death and exposure data described in Chapters 3 and 4. Let $d_{x,t}$ represent the number of deaths for those age $x$ in year $t$, and $e_{x,t}$ be the total number of person-years lived in the interval $[x, x+1)$ during calendar year $t$. Then, the matrices $D$ and $E$ contain as elements $d_{x,t}$ and $e_{x,t}$ respectively, with age in rows and time in columns. The central death rates $m_{x,t}$ are elements in matrix $m$

$$m = D/E,$$

where division is performed element by element. Often we would rather work with the probability of dying in a single year, denoted $q_{x,t}$, for someone age $x$ in year $t$. Assuming a constant force of mortality over each age and calendar year, $m_{x,t} = \mu_{x,t}$ (Bowers et al., 1997); therefore,

$$q_{x,t} = 1 - e^{-\mu_{x,t}} = 1 - e^{-m_{x,t}}.$$

This equivalence will become important in Chapter 8 for pricing annuities.

5.2 Estimating the Parameters

Estimation of the parameters in the Lee-Carter model is a two step process. First, we find the set of $a_x$, $b_x$, and $k_t$ that give a least squares solution. Then, we reestimate $k_t$ such that, given the exposure, the implied number of deaths equal the actual number of deaths observed.
5.2.1 Step 1

In step 1, we wish to find the least squares estimates of $a_x$, $b_x$, and $k_t$. The estimation cannot be completed with ordinary regression methods, because there are no given regressors on the right hand side. But, a solution can be found using singular value decomposition (SVD).

First, we must estimate $a_x$. We know that each $a_x$ should be the average over time of $ln(m_{x,t})$. Therefore, the vector $a$ ($r \times 1$) will be the row averages of $ln(m)$. Next, we must prepare for SVD by subtracting $a_x$ from $ln(m_{x,t})$. Since the same $a_x$ will be subtracted from $ln(m_{x,t})$ regardless of $t$, we should construct a matrix $a$, the same size as $ln(m)$, with every column being the vector $a$. Now we can apply SVD to the matrix $\tilde{m} = ln(m) - a$ to find the estimates for $b_x$ and $k_t$.

Singular Value Decomposition is a special factorization of a matrix $\tilde{m}$ ($r \times c$) into matrices $U$ ($r \times r$), $V$ ($c \times r$), and diagonal matrix $D$ ($r \times r$), such that $\tilde{m} = UDV'$. The SVD of a matrix can be found by most statistical programs (e.g. the `svd` command in R). Then, once the normalization conditions have been taken into account, the first column of $U$ will be the estimates for $b_x$, and the first column of $V$ will be the estimates for $k_t$. To be exact, let $u_1$ be the sum of the elements in the first column of $U$, or the sum of $U(x, 1)$ over all values of $x$. Then,

$$b_x = U(x, 1)/u_1$$

(5.4)

and

$$k_t = V(t, 1) \ast u_1 \ast D(1, 1).$$

(5.5)

By dividing by $u_1$, we have forced the elements of the first column of $U$ to sum to 1, and therefore $\sum_x b_x = 1$. Also, since each column of $V$ always sums to 0, multiplying it by a constant will result in $\sum_t k_t = 0$. Now, all the estimates have been determined.

5.2.2 Step 2

Although the estimates in Step 1 satisfy the least squares property, it is most often the case that the implied number of deaths in each year do not equal the actual number of deaths in the data. Since we work from the $q_x$ values, equal weighting is placed on all age groups, even though the older ages contribute many more deaths. Thus, Step 2 works to correct this problem.
We keep the estimates for $a_x$ and $b_x$ from Step 1, but we adjust the $k_t$ values through an iterative search. For each value of $t$, we adjust the value of $k_t$ until both sides are equal in the following equation:

$$\sum_x d_{x,t} = \sum_x e_{x,t} e^{a_x + b_x k_t}.$$ (5.6)

The new estimates for $k_t$ will be only slightly different from the old estimates. Now the implied number of deaths equal the actual number of deaths in each year.
Chapter 6

Model Fitting

In this chapter, we will model the Swedish mortality data discussed briefly at the beginning of Chapter 4. First, we will fit a GLAM with optimally scaled shocks as was done in Kirkby and Currie (2009). Then, a Lee-Carter model will be fitted. The two models will be compared.

6.1 Data

Swedish death and exposure figures (matrices $D$ and $E$) were retrieved from the Human Mortality Database at http://www.mortality.org. Both sets of figures span calendar years 1751-2007 and ages 0-110+. The data used in this thesis can be retrieved from the thesis section at http://www.stat.sfu.ca/.

The documentation on the website mentions that the quality of data prior to 1860 is questionable. Therefore, for the purposes of this fit, we will restrict the data to years 1900-2000. We will also limit the examination to ages 10-90, to avoid the very young which behave a little differently from the rest of the curve and the very old which have limited exposure. These are the age and time limits imposed by Kirkby and Currie (2009), although the authors do not explicitly discuss their reasons for truncation. Figure 6.1 shows the raw mortality surface, where the height of the surface is $\log(D/E)$.
Figure 6.1: Swedish mortality data. Each value of log(death/exposure) is plotted as a function of age and calendar year. The large spike in mortality in 1918 is due to the Spanish influenza epidemic.
CHAPTER 6. MODEL FITTING

6.2 GLAM Fitting

6.2.1 B-spline Matrices

The first step to fitting the additive GLAM described in Chapter 4, is to construct the B-spline matrices $B_a$, $B_y$, and $B_s$. These matrices rely on the choice of two things: the order the B-splines and the locations of the knots. Kirkby and Currie (2009) chose to use cubic B-splines for all three matrices. For knot locations, $B_a$ and $B_y$ have a knot placed every 5 years, and $B_s$ has a knot placed every 15 years. Recall that since these B-splines are cubic, three extra knots are required before and after the range of the data. Therefore, the knot vector for $B_a$ is $[-5, 0, 5, 10, ..., 90, 95, 100, 105]$, the knot vector for $B_y$ is $[1885, 1890, 1895, 1900, ..., 2000, 2005, 2010, 2015]$, and the knot vector for $B_s$ is $[-35, -20, -5, 10, ..., 100, 115, 130, 145]$.

6.2.2 Model Selection

As mentioned in Section 4.1.2, we choose a second order difference matrix to be the penalty matrix. The associated constants $\lambda_a$, $\lambda_y$, and $\lambda_s$ have yet to be determined. Given any set of lambdas, we can fit the model, and then evaluate the Bayesian Information Criterion (BIC) (Schwarz, 1978). The smaller the BIC, the better the model fits the data. BIC is defined as:

$$BIC = \text{Deviance} + k \log(n) = 2 \log \Lambda + k \log(n), \quad (6.1)$$

where $\Lambda$ is the likelihood ratio, $k$ is the effective dimension of the model, and $n$ is the sample size. For the Swedish mortality data, $n$ is the number of year and age combinations ($81*101=8181$), $k$ can be calculated as the trace of the hat matrix

$$H = B(B'\hat{W}_\delta B + P)^{-1}B'\hat{W}_\delta, \quad (6.2)$$

and the likelihood ratio is

$$\Lambda = \text{sum}(D \log(\frac{D}{\mu})). \quad (6.3)$$

The optimal set of lambdas are found by minimizing the BIC. Since standard minimization methods do not reliably yield an answer, a grid search is used to find this minimum.
6.2.3 Results

Table 6.1 lists the optimal set of lambdas for the Smooth, Simple Shock, and the Scaled Shock models with various $\alpha$ levels. Note that the simple shock model with $\alpha = 0$ must be fitted before the models with a scaled shock component can be fitted.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Smooth Model</th>
<th>Simple Shocks</th>
<th>Scaled Shocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N/A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\lambda_a$</td>
<td>10.24</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda_y$</td>
<td>6.50</td>
<td>2150</td>
<td>1300</td>
</tr>
<tr>
<td>$\lambda_s$</td>
<td>$\infty$</td>
<td>800</td>
<td>250</td>
</tr>
<tr>
<td>BIC</td>
<td>23485</td>
<td>13451</td>
<td>12621</td>
</tr>
</tbody>
</table>

Table 6.1: Optimal parameters and corresponding BIC for various candidate models fitted to Swedish mortality data. Parameters for the shock models were found with grid search. These results, which we have verified, are from Kirkby and Currie (2009). Since the scaled shock model with $\alpha = 3.5$ has the smallest BIC, this is the optimal model.

We would like to make a few comments before analyzing the results of the table. First, a general interpretation of the parameters can be made. The larger the values of the lambdas and $\alpha$, the more smoothing that is taking place in the fitted surfaces. Choosing a set of very large parameters will make the surface very smooth, but then the fit will be poor. Choosing a set of small parameters makes the surface rough and the residuals small, but the large differences in the coefficients are severely penalized in the penalty matrix $P$. Therefore, the parameters that strike a healthy balance will have the smallest BIC. The second comment concerns the roundness of the numbers in Table 6.1. It may seem to the reader that the four parameters are not of very high precision. For example, the true minimum of $\lambda_y$ with $\alpha = 3$ is probably not exactly 1600. However, applying any extra effort to be more exact in these parameters will not have a noticeable impact on the BIC or the fit.

We now start our discussion of the fitted models. First, consider the simplest model: the smooth GLAM. The optimal parameters for this model are in the first column of Table 6.1, and the fitted surface is pictured in Figure 6.2. This surface attempts to model the entire data set, so it will produce humps when there are temporary shocks to the surface. This means that shocks can obscure the underlying trend.
Figure 6.2: Fitted surface for the smooth GLAM to Swedish mortality data.
Therefore, it makes sense to consider an additive GLAM, where the fitted surface is the sum of a smooth surface and a shock surface. Although this extra complexity requires more parameters, the BIC does decrease by nearly 50%, implying that the gain in fit overtakes the desire for parsimony. We can see, from Table 6.1, that the best model is the optimal scaled shock model with \( \alpha = 3.5 \), because it has the lowest BIC. The smooth and shock portions of this model are pictured in Figure 6.3. When added together, they produce the total fitted surface in Figure 6.4. Looking at these plots, it is evident that the large shocks have been retained, while the small shocks have been smoothed out.

Figure 6.3: Smooth surface (top) and shock surface (bottom) of optimal scaled shock model (\( \alpha = 3.5 \)) fitted to Swedish mortality data with additive GLAM methodology. The largest spike in mortality in 1918 is due to the Spanish influenza epidemic.
Smooth + Shocks

Figure 6.4: Total surface of optimal scaled shock model ($\alpha = 3.5$) fitted to Swedish mortality data with additive GLAM methodology. This surface is the sum of the smooth surface and the shock surface in Figure 6.3.
6.3 Lee-Carter Fitting

As a point of comparison, we fit the Lee-Carter model to the same Swedish mortality data (Figure 6.1). The steps for fitting this model are covered in Chapter 5. In short, the first step is to find optimal values for $a_x$, $b_x$, and the $k_t$ with singular value decomposition (SVD). The $k_t$ values from the first step appear in the top left panel of Figure 6.5. The lower left panel of the same figure shows that there are discrepancies between the observed number of deaths and the predicted number of deaths in each year. Therefore, the second step keeps the estimated values for $a_x$ and $b_x$, but adjusts $k_t$ to force the observed and expected deaths in each year to match. The adjusted $k_t$ values are pictured in the upper right panel of Figure 6.5.

![Graphs showing two-step $k_t$ fitting](image)

Figure 6.5: Two-step $k_t$ fitting. The optimal $k_t$ values found with SVD are pictured in the upper left, and the adjusted $k_t$ values are pictured in the upper right.

Using the adjusted $k_t$ from Step 2 with $a_x$ and $b_x$, we can derive the fitted surface in Figure 6.6 with the help of Equations (5.1) and (5.3).
Figure 6.6: Lee-Carter model fitted mortality surface to Swedish mortality data.
6.4 Comparing the Two Models

Comparing the GLAM surface in Figure 6.4 and the Lee-Carter surface in Figure 6.6, we can comment on a few differences. First, the GLAM is much more dynamic. Since we decomposed the surface into a smooth part and a shock part, we can tease apart unusual mortality events like pandemics and war to focus on the general trend. This is not possible with the Lee-Carter model, because there is no such division. Also, the GLAM surface is free to change in age related mortality structure over time. The Lee-Carter model on the other hand has to keep the same general mortality profile through time, because $a_x$ and $b_x$ are fixed. This restriction could be seen as an advantage, if for prediction, we want to retain the current age structure and avoid crossings: locations where younger ages have higher mortality than older ages. But, with restriction comes limitation; if the data exhibited any major changes, then the Lee-Carter model would be unable to accommodate them.

Another difference lies in inter-age and inter-year estimation. Because of the way the GLAM is constructed, estimations of the surface height between years and between ages is a built-in feature. This is because the surface is defined continuously according to the B-spline basis. The Lee-Carter surface on the other hand is defined discretely. There are no $k_t$ values between years, or $a_x$ and $b_x$ values between ages, so there is no existing provision for applying a smooth progression. This could be an issue if, for example, the data was only available every 10 years, and estimations for every year were needed.

The advantage of the Lee-Carter model is simplicity. There are fewer parameters, fitting the surface is an easy two-step procedure with minimal calculations, and projections through time are easily accomplished by fitting a time series model to the values of $k_t$. If older ages always have greater mortality than younger ages in the fitted model, we can be guaranteed that this ordering is preserved in the projections. But, for the GLAM, a method of projection is not obvious. A lack of restriction could cause predicted values to behave strangely. A method of prediction for the GLAM will be developed in Chapter 7.
Chapter 7

Predicting Mortality

Thus far, we have used the GLAM and Lee-Carter models to describe mortality patterns of the past. But, the ultimate goal is not analyzing what has passed, but what is yet to come. Prediction can be important for many reasons including the setting of public policies, anticipating the size of future populations, and pricing mortality dependent contracts. The Lee-Carter model uses a time series method of prediction. On the other hand, the existing method of GLAM prediction treats future values as missing data, and lets the model fill in the blanks (Currie et al., 2004). Although the order of penalty does not greatly alter the fit in the range of the data, predictions can be noticeably affected. Therefore, it is not the data that determines the penalty order, but a personal expectation of how the future mortality should behave. As an alternative, we propose using time series methods to predict the GLAM coefficients, and consequently the mortality surface. In the following sections, we will forecast Swedish mortality with the Lee-Carter model, and develop a time series method of prediction for the GLAM.

7.1 Lee-Carter Prediction

Prediction with the Lee-Carter model is accomplished by fitting a time series to the values of \( k_t \). Since the values themselves are not stationary, we difference them to get the changes in \( k_t \) from one period to the next, pictured in Figure 7.1. These differenced values do appear to be stationary, with some exceptions due to large shocks.

Various orders of ARMA models are then fitted to the differenced values, and we use Akaike’s information criterion (AIC) (Akaike, 1973) to evaluate them. Although including
an AR(1) term or an MA(1) term does give lower AIC values than a model without, the differences are so small that a model with only random noise was chosen for the differenced values. In other words, we choose an ARIMA(0,1,0) model for $k_t$, which is a random walk with drift.

![Graph showing $k_t$ differences through time.](image)

**Figure 7.1:** First-differenced $k_t$ values against time.

Lee and Carter (1992) take similar steps to those above, applying Box Jenkins model identification procedures to US mortality data. In their paper, they choose a random walk with drift as well, concluding that, “A similar model with an ar(1) term added was marginally superior, but we preferred the (0,1,0) model on the grounds of parsimony...all gave essentially the same point forecasts and differed only slightly on the confidence bands” (Lee and Carter, 1992). The findings of Lee and Carter are very similar to our own, providing extra confidence in our results.

Another issue to address is the extreme value of $k_t$ in 1918, due to the Spanish influenza epidemic. Lee and Carter (1992) encountered this problem as well, deciding to remove its influence with a dummy variable, thereby considering it an unusual event. If we let $flu$ be
1 only in 1918 and 0 elsewhere, we arrive at the following fit for Swedish data:

\[
k_t = k_{t-1} - 1.6156 + 31.7639 \times flu + \varepsilon_t
\]

\[
(7.1) \quad (0.5230) \quad (3.6983)
\]

\[
\varepsilon_t \sim N(0, 27.35).
\]

The standard errors for the estimated coefficients appear in parentheses below the main equation. From Equation (7.1), we see that \(k_t\) is drifting downward at an average rate of -1.6156 per year. A plot of the fitted and the predicted \(k_t\) values appears in Figure 7.2. The slope of the line after the year 2000 is -1.6156.

Suppose we wish to simulate future values of \(k_t\) according to the time series model in (7.1). The first \(k_t\) will be the most recently observed value. Then, by randomly generating errors from a normal distribution with mean 0 and standard deviation 27.35, we can determine a possible path for future values of \(k_t\). Each generated path can be combined with \(a_x\), \(b_x\), and Equations (5.1)-(5.3) to solve for the corresponding predicted levels of mortality \(m = D/E\). By repeating this simulation process many times, we can estimate the distribution of future mortality.

If we let the path of \(k_t\) follow the predicted values exactly, we can find the predicted mortality surface. The raw mortality surface, with values predicted by the Lee-Carter model, is plotted in Figure 7.3. One noticeable feature of this surface is the rather large discrepancy between the most recently observed mortality rate and next year’s predictions in younger ages. This is due to the various restrictions imposed by the Lee-Carter model, and the large weight being placed on fitting older ages to make the total predicted deaths in a year match the observed number.
Figure 7.2: Fitted and predicted $k_t$ through time.
Figure 7.3: Raw mortality surface before year 2000, and the predicted Lee-Carter mortality surface thereafter.
CHAPTER 7. PREDICTING MORTALITY

7.2 GLAM Prediction

The chosen GLAM model that best fits the mortality surface was an additive shock model. Therefore, the fitted surface is the sum of a smooth surface and a surface of shocks. When we carry out prediction, we will ignore the shock surface, and concentrate on predicting the smooth surface. Health advances take time to implement, but an illness or natural disaster is sharp and temporary. Therefore, it seems safe to assume that any large shocks to the mortality surface will be positive. These positive shocks will cause people to die earlier than expected, and thus, reduce the cost of providing the annuity. So, by ignoring the shocks, our estimates will be conservative. On the other hand, a life insurance policy would be adversely affected by positive shocks, making their inclusion much more important to risk assessment.

The GLAM is a more complicated model than the Lee-Carter model to extrapolate through time. The smooth surface has a matrix of coefficients $\Theta$ of size $19 \times 23$, where the rows correspond to age and the columns to time. Each coefficient in the matrix corresponds to a B-spline in the age direction and a B-spline in the time direction, so we label the coefficients by the age and time at which they peak. Therefore, the 19 rows will represent every 5 years from ages 5 to 95, and the 23 rows will represent every 5 years from 1895 to 2005. The basis of B-splines can be easily extended through time, so the only hard part is predicting each of the 19 age related coefficients.

Since there are only 23 points in the series for each of the 19 coefficients, fitting a multidimensional time series becomes problematic. There are too many parameters to determine and not enough time points to provide the necessary information. Fortunately, the 19 coefficients are highly correlated, so we simplify the problem by arbitrarily choosing only three key age coefficients to predict: ages 30, 50, and 80. Then, we will assume that the other age coefficients can be described as a linear combination of these three key ages. This means that we fit a multiple regression for each of the age coefficients, using the coefficients for ages 30, 50, and 80 as predictors. Now, we only have to predict three coefficients through time, and use the multiple regression results to find all the other ages.

We have simplified our problem down to three coefficients, so the next step is to pick an appropriate time series model based on observed values. First, a random walk with drift is fitted. Let $\Theta(x, t)$ be the coefficient for age $x$ in year $t$, and $\varepsilon_t$ be the white noise vector
following a multivariate normal distribution. Then, the fitted equation is

\[
\begin{bmatrix}
\Theta(30, t) \\
\Theta(50, t) \\
\Theta(80, t)
\end{bmatrix}
= \begin{bmatrix}
\Theta(30, t - 1) \\
\Theta(50, t - 1) \\
\Theta(80, t - 1)
\end{bmatrix}
- \begin{bmatrix}
1.187e-17 \\
4.237e-18 \\
1.515e-18
\end{bmatrix}
+ \begin{bmatrix}
\varepsilon_{1t} \\
\varepsilon_{2t} \\
\varepsilon_{3t}
\end{bmatrix}
\] (7.2)

\[
\begin{bmatrix}
\varepsilon_{1t} \\
\varepsilon_{2t} \\
\varepsilon_{3t}
\end{bmatrix}
\sim N\left(\begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}, \begin{bmatrix}0.00673 & 0.00263 & 0.00090 \\ 0.00263 & 0.00263 & 0.00114 \\ 0.00090 & 0.00115 & 0.00134\end{bmatrix}\right)
\]

Resulting predictions are shown in Figure 7.4.

Next, an AR(1) model with drift is fitted, pictured in Figure 7.5. Referring to Figures 7.4 and 7.5, the random walk performs well, while the AR(1) model gives unreasonable behaviour. We see that the coefficients predicted by the AR(1) model are in some instances increasing, which implies an increase in mortality rates for the long term. This is counter-intuitive, because aside from the occasional shock, mortality has demonstrated unwavering improvement through time. Also, there are age crossings: instances in which mortality rates for younger ages are larger than for older ages. This is unrealistic as well, as we expect older ages to have a larger probability of dying. Although we only show the results for two simple models, we found that more complicated, higher order models lead to more extreme problems. Therefore, we choose the random walk with drift given by Equation (7.2) to model the three GLAM coefficients.

The last step is to translate these predicted coefficients into a predicted surface. Using the multiple regression results on past data, all the other age coefficients are determined. Then, \(B_\alpha \Theta B_\gamma^T\) will give the fitted surface in Figure 7.6. The only difference is that \(\Theta\) and \(B_y\) have more columns to accommodate the future years. For example, since we have data up to the year 2000, predictions to 2020 will require 4 extra columns: one column for each additional set of 5 years.

In addition to the predicted surface, we could use (7.2) to simulate predicted paths for the coefficients, and thus possible mortality surfaces. We keep the most recent values for the age 30, 50, and 80 coefficients, and randomly draw errors from the multivariate normal distribution described in (7.2). Then, the time series equation determines future values for these three coefficients. The missing age coefficients are estimated through the multiple regression results on the past data. Finally, the coefficients are translated into a mortality
Figure 7.4: Observed coefficients are shown up to 2005 with solid lines, and predicted coefficients are shown thereafter with dotted lines. Age 30, 50, and 80 coefficients are predicted with a multivariate time series model: a random walk with drift. Each of the other coefficients are determined through a multivariate regression on the age 30, 50, and 80 coefficients.
Figure 7.5: Observed coefficients are shown up to 2005 with solid lines, and predicted coefficients are shown thereafter with dotted lines. Age 30, 50, and 80 coefficients are predicted with a multivariate time series model: an AR(1) with drift. Each of the other coefficients are determined through a multivariate regression on the age 30, 50, and 80 coefficients.
Figure 7.6: Raw and predicted GLAM mortality surface.
surface. Repeating this simulation process many times will provide an approximation for the distribution of future mortality.

### 7.3 Comparing Lee-Carter and GLAM

With 3-dimensional plots, it can be difficult to directly compare the surfaces produced by the GLAM and Lee-Carter models. Therefore, Figures 7.7 and 7.8 show the paths of the two models for every 10 years of age. With these cross sections, we are more able to make comparisons between the two models.

First, examine the two fitted models before the year 2000. A noticeable difference is that the Lee-Carter fit in red is more erratic than the smooth GLAM fit in blue. Also, the Lee-Carter model does not fit the data well in recent years. For ages 20, 30, 40, and 50, the model severely underestimates the level of mortality. Because there are more deaths in the older ages, more weight gets placed on fitting those ages properly (see (5.6)). Therefore, due to certain inflexibilities in the model, the younger ages are forced into a poor fit.

After year 2000, we ran 1000 simulations of the future parameter path for each model, each time resulting in a different projected mortality. In Figures 7.7 and 7.8, the blue and red lines after year 2000 show selected percentiles: the points at which 10%, 30%, 50%, 70%, and 90% of the simulated data are smaller in a given year. These percentiles give us an idea of the variability in the projections. For example, the difference between the 90th and 10th percentiles will give a measure of spread, pictured in Figures 7.9 and 7.10. We know that both models depend on parameters that follow a random walk with drift, which has an ever increasing spread. However, the equations that translate these parameters into mortality levels limit the projected mortality spread, such that it will not widen indefinitely. Looking at Figures 7.9 and 7.10, we see that the spread temporarily increases to a maximum sometime before the year 2100 and then decreases with time. For all ages, the Lee-Carter model starts with a larger spread than the GLAM. Then at some point, the lines cross, and the GLAM spread becomes larger than the Lee-Carter spread; this cross happens sooner for younger ages, and later for older ages.

The GLAM appears to fit the data better. The past fit is much smoother, which we assume captures the trend better. Also, the future projections begin near the last observed value for all ages. The Lee-Carter model, on the other hand, does a poor job at modeling
Figure 7.7: In each plot, raw mortality data are shown as black circles. The blue and red lines show the GLAM and Lee-Carter models respectively: the fitted values up to year 2000 and selected percentiles of 1000 simulations afterward. The green dashed lines show the expected values according to each model, which are almost identical to the 50th percentile.
Figure 7.8: In each plot, raw mortality data are shown as black circles. The blue and red lines show the fitted GLAM and Lee-Carter models respectively up to year 2000. Afterward, the blue and red lines show selected percentiles of 1000 simulations. The green dashed lines show the expected values according to each model, which are almost identical to the 50th percentile.
Figure 7.9: In each plot, the difference between the 90th and 10th percentiles of 1000 simulations is shown for the GLAM and Lee-Carter models. This is an interval width that will contain 80% of the simulated points, and is therefore a measure of spread for the predictions.
Differences Between the 90th and 10th Percentiles

Figure 7.10: In each plot, the difference between the 90th and 10th percentiles of 1000 simulations is shown for the GLAM and Lee-Carter models. This is an interval width that will contain 80% of the simulated points, and is therefore a measure of spread for the predictions.
recent years for young ages. But, since the mortality rates are so small anyway, the discrepancy might not make much of a difference, depending on the application. This issue will be addressed in Chapter 8, when we use these predicted mortality rates to price some life annuities.
Chapter 8

Pricing Life Annuities

Now that we have developed a way to project the future mortality rates, with both GLAM and Lee-Carter models, our aim is to apply these projections to the pricing of life annuities. In this chapter, we will price, at various ages and times, three types of life annuities: a whole life, a 10-year temporary, and a 10-year guaranteed. The prices of these and their associated variances should give us a better idea about the performance of the two mortality models.

8.1 Basic Definitions

An annuity certain is a promise to make $1 payments at the end of every year for $k$ years. The present value of an annuity certain at interest rate $i$ is defined as

$$a_{\overline{k}|i} = \sum_{j=1}^{k} v^j = \frac{1 - v^k}{i},$$

where $v = (1 + i)^{-1}$.

If instead of certain payments, the payments are made only if a person now age $x$ is alive to receive them, then the promise is called a life annuity. Three standard types of life annuities are whole life, n-year temporary, and n-year guaranteed. We can describe these three annuity types in terms of their present-value random variable $Y$. All the definitions to follow are given by Bowers et al. (1997).
Suppose the integer-valued random variable $K$ is the curtate-future-lifetime of someone age $x$. Then, the present-value random variable $Y$ for an $n$-year temporary life annuity is
\begin{equation}
Y = \begin{cases} 
  a_{K\mid i}, & 0 \leq K < n \\
  a_n, & K \geq n 
\end{cases},
\end{equation}
for a whole life annuity is
\begin{equation}
Y = a_{K\mid i}, \quad K \geq 0,
\end{equation}
and for an $n$-year guaranteed life annuity is
\begin{equation}
Y = \begin{cases} 
  a_{n\mid i}, & 0 \leq K < n \\
  a_{K\mid i}, & K \geq n 
\end{cases}.
\end{equation}
To evaluate the moments of $Y$, we need the probability function of $K$:
\begin{equation}
P(K = k) = k p_x q_x + k, \quad k = 0, 1, 2, \ldots,
\end{equation}
where $k p_x$ is the probability of someone age $x$ living to age $x + k$, and $q_x + k$ is the probability that someone age $x + k$ will not be alive at age $x + k + 1$. These probabilities can be made to depend on the year the annuity starts, as mortality is improving with time.

Having defined all the possible outcomes and the associated probabilities, the moments of $Y$ are simple to determine. For example, the expected present value of an $n$-year temporary life annuity for someone age $x$ combines (8.2) and (8.5) as
\begin{equation}
E(Y) = a_{x\mid i} = \sum_{k=0}^{n-1} a_{k\mid i} k p_x q_x + k + \sum_{k=1}^{n} v_k k p_x.
\end{equation}
We are also interested in variance of the present value, so we require the second moment
\begin{equation}
E(Y^2) = \sum_{k=0}^{n-1} a_{k\mid i}^2 k p_x q_x + k + \sum_{k=1}^{n} v_k^2 k p_x.
\end{equation}
The moments for whole life and $n$-year guaranteed are similarly calculated. The first two moments of a whole life annuity present value (8.3) are
\begin{equation}
E(Y) = a_x = \sum_{k=0}^{\infty} a_{k\mid i} k p_x q_x + k = \sum_{k=1}^{\infty} v_k k p_x
\end{equation}
\begin{equation}
E(Y^2) = \sum_{k=0}^{\infty} a_{k\mid i}^2 k p_x q_x + k,
\end{equation}
and the first two moments of an n-year guaranteed life annuity present value (8.4) are

\[
E(Y) = a_{x-n|n} + a_{x-n+1|n+1} + \sum_{k=n}^{\infty} a_{k|n} \ k p_x q_{x+k} = a_{x|n} + \sum_{k=n+1}^{\infty} v^k k p_x
\]  
(8.10)

\[
E(Y^2) = a_{x-n|n}^2 + \sum_{k=n}^{\infty} a_{k|n}^2 \ k p_x q_{x+k}.
\]  
(8.11)

For all three types, the variance of the present value is

\[
V(Y) = E(Y^2) - E(Y)^2.
\]  
(8.12)

### 8.2 Incorporating Changing Mortality

Chapters 6 and 7 developed mortality models that rely on stochastic parameters. The GLAM and Lee-Carter projections depend on one or more key parameters that follow a random walk with drift. So, we can simulate a path for these parameters and get a corresponding set of mortality projections. There are many different paths, and therefore many possible projections; this makes the matrix of mortality rates a random variable. Equations (8.8) and (8.9), although a good start, assume the probabilities of life and death are fixed. Using a few statistical identities, we can expand upon these equations to find the expected value and variance when mortality is stochastic.

Let \( Q \) be the matrix of mortality rates for all ages and all times, including future times. Ages are in rows and years are in columns. Therefore, the mortality probabilities that an individual experiences in their lifetime follow a diagonal path down and to the right. This is because with each year of increase in age also comes an increase in one year of time. We know the past rates, so they are fixed, but the future rates are random. This makes the matrix \( Q \) a random variable.

Let \( a \) be the matrix of life annuity present values that depends directly on \( Q \), indexed by each starting age and year in the same way as \( Q \). We assume that every person in the population will adhere to these mortality rates. Then, the expected value of \( a \) will be

\[
E(a) = E(E(a|Q)),
\]  
(8.13)

where \( E \) is the expectation operator. Equation (8.13) can be evaluated with the following approach. We will first use the mortality models to repeatedly generate realizations of \( Q \).
For every realization, we can calculate the corresponding $E(a|Q)$ matrix using (8.8) for each entry. Then, as the number of realizations gets large, a simple average of the realized values of $E(a|Q)$ will approach the true value of $E(a)$.

The variance of the present value has two sources of variability: the mortality rates a person experiences and when that person actually dies. If $V$ is the variance operator, then

$$V(a) = E(V(a|Q)) + V(E(a|Q)).$$

(8.14)

The first term in the sum represents the average risk in not knowing an individual’s year of death, and the second term is the risk associated with not knowing the future rates of mortality. If $Q$ were fixed, $E(a|Q)$ would always be the same, so $V(E(a|Q))$ would go to zero.

Equation (8.14) is evaluated in a way similar to Equation (8.13). First we generate several realizations of $Q$, and use them to find the entries in $E(a|Q)$ with (8.8) and $V(a|Q)$ with (8.12). The first term in (8.14) will be the average of the $V(a|Q)$ realizations, and the second term will be the variance of the $E(a|Q)$ realizations.

The question remaining is how many realizations to use. To test this, $E(a)$ and $V(a)$ were recalculated and stored with each additional set of 1,000 runs, up to a total of 10,000 runs. By doing this, we can track how the entries in these matrices change when more runs are added, to see if the extra work is necessary. To measure these changes, the absolute difference and the absolute percent difference for each entry in these matrices are calculated when an extra 1,000 runs are added. As a crude summary of the entire matrix, these two measures can each be summed over all of the entries.

We ran a total of 10,000 simulations in blocks of 1,000. It is immediately clear from these four summary measures that 2,000 runs is not enough. The entries in $E(a)$ and $V(a)$ are still changing too much from 2,000 runs to 3,000 runs, making the summary measures large. After 5,000 runs however, $E(a)$ never has an entry that differs by more than 2% or more than 0.05 for an additional 1,000 runs. Likewise, $V(a)$ never has an entry that differs by more than 2% or by 0.002 for an additional 1,000 runs. It also seems that the sum of absolute differences and the sum of absolute percent differences have stopped making large improvements. Therefore, 5,000 realizations is probably enough, but we will use the full 10,000 to be safe.
8.3 Results

In this section, we apply the methodology described above. The expected present values and associated variances are found for three types of life annuities: whole life, 10-year temporary, and 10-year guaranteed. Annuities starting in every year from 1932 to 2030, and at every age from 20 to 80, are evaluated at a constant interest rate of $i = 4\%$. We choose to start in 1932, because the calculation of any annuity values before this year will only depend on observed mortality rates; therefore, it provides no information about the performance of our mortality projections. Also, we stop at age 80 because the limiting age of the table is at age 91; a 10-year temporary annuity sold to someone age 82 or older would have no possibility of receiving all 10 payments. Age 80 is therefore a natural stopping place. If a limiting age of 91 is not satisfactory, one could extend the projected life table above age 91 with methods suggested by Coale and Guo (1989). The resulting surfaces under both the Lee-Carter model and the GLAM are pictured in Figures 8.1-8.6.

8.3.1 Expected Present Value

Referring to Figures 8.1-8.3, the expected present value surfaces for all three annuity types are declining with age. The older contract holders are closer to the limiting age of the life table, so the maximum number of payments is declining. Also, there is a smaller chance of being alive to receive those payments.

The rate of decline in present value over age is more constant for the whole life and 10-year guaranteed than for the 10-year temporary surface. The temporary present-value surface is very flat in the young ages, because chances are high that all 10 payments will be made. Note that the present value of a 10-year temporary annuity at age 20 is very close to the present value of a 10-year annuity certain. But, when the purchase age approaches the 50 to 60 range, death probabilities start increasing very rapidly; the probability of payments being made over each of the next 10 years starts dropping, and so the decline in present value is very steep.

Another common feature is that all three present-value surfaces are generally increasing with calendar year. Mortality is improving, so people are expected to live longer and receive more payments. The surfaces are rougher in early years, because their present values are mostly based on the raw experienced mortality. Moving forward in time, the expected
Expected Present Value of 10–year Temporary Life Annuity

Lee–Carter

GLAM

Figure 8.1: Expected present value surface for a 10-year temporary life annuity at an interest rate of 4%.
CHAPTER 8. PRICING LIFE ANNUITIES

Expected Present Value of Whole Life Annuity

Lee–Carter

GLAM

Figure 8.2: Expected present value surface for whole life annuity at an interest rate of 4%.
Expected Present Value of 10–year Guaranteed Life Annuity

Lee–Carter  
GLAM

Figure 8.3: Expected present value surface for 10-year guaranteed life annuity at an interest rate of 4%.
present value depends more and more on an averaging over projected mortality realizations. This process of averaging over many possibilities will make the surface smooth.

Now, comparing the surfaces to each other, one can see that a guaranteed life annuity is always more expensive than the other two types of annuities. The whole life and the guaranteed life annuities provide the same benefit pattern for life, but the first 10 payments of the guaranteed contract are paid regardless of the year of death. This makes the present value much higher for older ages where death probabilities are high, but has little effect on the young ages where death probabilities are low. The temporary annuity is of course the least expensive of the three, because payments are only made for the first 10 years. This type of annuity is much less expensive than the whole life annuity at young ages. At old ages, the present values are similar, because the probability of making more than 10 payments is very small.

8.3.2 Standard Deviation of the Present Value

Now, consider the standard deviation surfaces in Figures 8.4-8.6. The surfaces for the whole life and the 10-year guaranteed annuities have the same general shape: low for the youngest and oldest ages, and peaking between ages 50-60.

Young people will have a low standard deviation, because they are expected to die many years from now. The bulk of the present value is made up of almost certain payments in the short term with a small amount of discounting. The most uncertain payments will occur so far in the future that their discounted value is a very small proportion of the total present value. Old ages have a small standard deviation, because those people will die sooner rather than later; there are not many future payments left to consider, and so the range of possible present values cannot be very wide.

So, almost all of the time, young ages are expected to receive a lot of payments and old ages are expected to receive very few payments. It is for a person now around age 60 that the probability of dying at a given age is almost uniform. A person around this age has an almost equal chance of dying 10 years from now or 30 years from now. This will cause the standard deviation to be high.

The standard deviation of the present value for a temporary life annuity is much smaller than the other two annuity types. The maximum for the temporary surface is around 2.7 at age 80, which coincides with the value given by the whole life surface. The standard deviation minimum is almost 0 at age 20, because it is very rare to make anything but 10
Figure 8.4: Standard deviation surface of present value for 10-year temporary life annuity at an interest rate of 4%.
Figure 8.5: Standard deviation surface of present value for whole life annuity at an interest rate of 4%.
Figure 8.6: Standard deviation surface of present value for 10-year guaranteed life annuity at an interest rate of 4%.
payments. So, as age increases, the uncertainty about the number of payments increases, resulting in a larger standard deviation.

### 8.3.3 Some Representative Numbers

As it is sometimes difficult to assess data values from a 3-D graph, a representative selection of the expectation and variance of the present value appears in Table 8.1. Values for the Lee-Carter model and GLAM are displayed over a range of purchase years, and for ages 25, 45, and 65. Each of the three life annuity types are represented.

First, focus on the whole life annuity table. In particular, consider someone age 25 in 1990; this person wants to purchase a whole life annuity 40 years from now, at age 65, for retirement income. The 1990 present value of this contract under the GLAM is 10.77. The person could assume that the present value would be the same 40 years from now. However, this assumption would be inadvisable. By 2030, the present value is expected to rise to 11.86 due to mortality improvements, resulting in a 10.13% increase. The change from 1970 to 2010 is even greater at 18.11%. To ignore mortality improvement would significantly underestimate the cost of an annuity in the future.

The whole life annuity table also provides a breakdown between the two sources of variability that comprise the total standard deviation: $\text{Var}(\mathbb{E}(a|Q))$ and $\mathbb{E}(\text{Var}(a|Q))$ (see Equation (8.14)). Note that $\text{Var}(\mathbb{E}(a|Q))$ is very small compared to $\mathbb{E}(\text{Var}(a|Q))$. This is good, because $\mathbb{E}(\text{Var}(a|Q))$ goes to zero with a very large portfolio of similar contracts. However, an insurance company does not usually have enough identical contracts for this to be the case; approximately 10,000 contracts are usually needed to make the $\mathbb{E}(\text{Var}(a|Q))$ term negligible. As an example, consider an age 25 whole life contract under GLAM in 2030. This contract has an expected present value of 22.03 with a standard deviation of 1.88. With 100 contracts, the standard deviation will drop to 0.21, and with 1000 contracts, it drops to 0.12. The limiting standard deviation is 0.10. Similarly, a 65 year old with the same characteristics has an expected present value of 11.86 with standard deviation 4.06 for a single contract. For 100, 1000, and infinite identical contracts, the standard deviation becomes 0.47, 0.26, and 0.23 respectively. So, the standard deviation is reduced significantly with only 100 contracts, but even with 1000 contracts, the standard deviation can still be noticeably different from the limiting case.
### 10-Year Temporary Life Annuity

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(a)$</td>
<td>25</td>
<td>8.07</td>
<td>8.06</td>
<td>8.07</td>
<td>8.08</td>
<td>8.10</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>7.92</td>
<td>7.92</td>
<td>7.93</td>
<td>7.93</td>
<td>7.97</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>6.88</td>
<td>6.88</td>
<td>6.98</td>
<td>6.97</td>
<td>7.16</td>
</tr>
<tr>
<td>$SD(a)$</td>
<td>25</td>
<td>0.47</td>
<td>0.51</td>
<td>0.49</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>0.97</td>
<td>0.98</td>
<td>0.95</td>
<td>0.94</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>2.21</td>
<td>2.21</td>
<td>2.15</td>
<td>2.17</td>
<td>2.02</td>
</tr>
</tbody>
</table>

### Whole Life Annuity

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>45</td>
<td>16.88</td>
<td>16.90</td>
<td>17.20</td>
<td>17.35</td>
<td>17.53</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>9.73</td>
<td>9.73</td>
<td>10.12</td>
<td>10.67</td>
<td>10.77</td>
</tr>
<tr>
<td>$SD(a)$</td>
<td>25</td>
<td>2.89</td>
<td>2.97</td>
<td>2.76</td>
<td>2.80</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>4.25</td>
<td>4.28</td>
<td>4.07</td>
<td>4.12</td>
<td>3.78</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>4.44</td>
<td>4.44</td>
<td>4.48</td>
<td>4.49</td>
<td>4.37</td>
</tr>
</tbody>
</table>

### 10-Year Guaranteed Life Annuity

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>45</td>
<td>17.08</td>
<td>17.09</td>
<td>17.38</td>
<td>17.52</td>
<td>17.67</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>10.96</td>
<td>10.96</td>
<td>11.25</td>
<td>11.26</td>
<td>11.73</td>
</tr>
<tr>
<td>$SD(a)$</td>
<td>25</td>
<td>2.65</td>
<td>2.70</td>
<td>2.48</td>
<td>2.52</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>3.69</td>
<td>3.72</td>
<td>3.52</td>
<td>3.57</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>2.80</td>
<td>2.80</td>
<td>2.89</td>
<td>2.88</td>
<td>2.89</td>
</tr>
</tbody>
</table>

### Table 8.1: Various life annuity related values by age and year for three types of contracts. These values are based on 10,000 simulations of the mortality path under the Lee-Carter (L-C) and GLAM models at an interest rate of 4%.
Having discussed the whole life annuity table in detail, we can now compare it to the other two contract types. The temporary and guaranteed life annuities give an idea of what happens to the expected value and standard deviation when the benefit pattern is altered. First, the temporary life annuity only has a maximum of 10 payments, so it is inexpensive, and the standard deviations are much smaller than the whole life annuity. In fact, when the total number of identical contracts is very high, the largest GLAM standard deviation is $0.055$ ($\sqrt{0.003}$) for age 65 in 2030, compared to 0.23 for the same whole life contract. Also note that the increase in the GLAM present value over a 40 year period is much smaller. Between 1990 and 2030, the age 65 temporary annuity present value only increases by 4.71%, compared to the 10.13% increase seen for the whole life.

Now, referring to the guaranteed life annuity table, we see that this contract is more expensive and less variable than the whole life. However, the present values and standard deviations follow a similar pattern for both contracts. The provision for 10 years of guaranteed payments lends some stability to the annuity value, increasing the present value but also reducing the standard deviation. The percentage increase in present value over 40 years is not as large as we saw with the whole life annuities. For example, from 1990 to 2030, the age 65 GLAM present value increases by 6.43% for the guaranteed life annuity, compared to 10.13% for the corresponding whole life annuity. Also, the GLAM standard deviations for the guaranteed life annuity are smaller than for the whole life. But, for a large number of contracts, the standard deviations for the two annuity types in the table become very close, with a maximum discrepancy of $0.05$ ($\sqrt{0.054} - \sqrt{0.033}$) for age 65 in 2030.

The final point to consider is how the results differ between the Lee-Carter model and the GLAM. The whole life annuity table exhibits the greatest differences between the two models. Large differences mainly occur in older ages. For example, at age 65, the GLAM predicts a present value in year 2000 that is 0.38 larger than the one given by the Lee-Carter model. This difference gets larger as the years progress: 0.44 in 2010 and 0.47 in 2030. In addition, the GLAM standard deviations are larger than those of the Lee-Carter for ages 25 and 45, especially in years 2000, 2010, and 2030. The differences in standard deviation for age 25 in those years are 0.41, 0.43, and 0.35 respectively. For age 65, on the other hand, the Lee-Carter standard deviations for years 2000, 2010, and 2030 exceed those estimated from the GLAM; the difference is at most 0.14 in 2030.
Chapter 9

Conclusion

Over the course of this project, we have explored Generalized Linear Array Models, and how they can be used to smooth a noisy mortality surface. The GLAM methodology is efficient and requires less storage than traditional GLM methods. Following the work of Kirkby and Currie (2009), we fit several models; a scaled shock model was selected, because it had the lowest BIC, indicating the best fit. We also fitted a Lee-Carter model for a measure of comparison, because it is a standard in academia and industry.

9.1 Comparing the Fit

The Lee-Carter model offers simplicity, but the GLAM has flexibility, resulting in a better fit to the raw data. In particular, the GLAM surface fits young ages better in recent years. Also, separately modeling the shocks reflects our view of the mortality surface as a smooth trend that occasionally gets jolted by temporary current events such as war and disease. The Lee-Carter model tries to capture the time trend with a single parameter, and so the fitted model will waver with these shocks. Also, with the GLAM, we are better able to account for extreme events like the influenza epidemic of 1918, instead of simply removing it with a dummy variable.

9.2 Prediction

The next step is to predict with the two models. The Lee-Carter model has a well developed method of forecasting that fits a time series model to the calendar year parameter.
We wished to do something similar with the GLAM. Choosing only three of the surface coefficients to predict in the calendar year direction, a multivariate time series model was fit to these representative values. Then, multivariate regression fills in the missing coefficients, and the coefficient matrix is easily translated into a mortality surface that includes predictions.

With the time series models in place, we find the prediction intervals as well. By simulating errors in the time series model, from a normal distribution for the Lee-Carter model and a multivariate normal distribution for the GLAM, parameters are projected and translated into possible outcomes for future mortality. Through simulation, we can find prediction intervals for mortality rates at every age.

Comparing the two models, the GLAM predictions seem more reasonable than the Lee-Carter predictions. The steep decline in mortality predicted by the Lee-Carter model for young ages does not seem consistent with recent data. The GLAM however, appears to adhere to recent trends in all ages. Also, the Lee-Carter model always starts with wider prediction intervals than the GLAM, and older ages maintain this relation far into the future. Perhaps these intervals are wider than necessary for annuity pricing, because the Lee-Carter model accounts for the positive mortality shocks in its prediction intervals.

9.3 Annuities

The simulations above were also used to derive expected values and variances for annuity present values, not possible through analytical methods. Of course, these expected values and variances rely on the correctness of the model. We have not considered how uncertainty in the model parameters might affect these results, nor have we considered violations of certain model assumptions that could show up in residual plots.

The expected annuity present values for the Lee-Carter model and the GLAM were often very close to one another. But, when the expected values differed by more than about 0.03 in either direction, it was always the GLAM that had the larger present value. The GLAM is more conservative in this sense. Also, compared to the Lee-Carter model, standard deviations for the GLAM were generally larger for the young age groups and smaller for the old age groups.
9.4 Future Work

There is much room here for future work. For instance, we could include another source of variability in the prediction intervals using parametric bootstrapping (Brouhns et al., 2002b). The current method only considers the part of the variability resulting from the time series prediction, ignoring the variation in the time series parameter estimates themselves (model error). Also, it would be useful to compare predictions based on subsets of the data. We used the whole data set from calendar years 1900 to 2000, but it might make more sense to only use data from the last 50 years. Alternatively, we could test the predictive ability of the model by only fitting to years 1900 to 1980, and testing the accuracy of the forecasts from years 1981 to 2000. In addition, our examination only categorized the data by age and year, disregarding heterogeneity within the age-year cells. Instead, we could split the data by sex, examining the performance of these models for non-aggregated data.

It would also be interesting to explore the existing method of GLAM prediction that treats future values as missing values. Then, a comparison of these predictions to our time series predictions could be made. Safety loading could be added, and the modeling of reserves could be carried out. Finally, we have only modeled the mortality risk, but we have ignored the interest rate risk. Instead of fixing the interest rate, we could model it with time series methods, painting a better picture of the total risk.
Bibliography


