Quantifying Biometric Life Insurance Risks With Non-Parametric Smoothing Methods

Julien Tomas

Quantifying Biometric Life Insurance Risks

With Non-Parametric Smoothing Methods

Julien Tomas

Quantifying Biometric Life Insurance Risks

With Non-Parametric Smoothing Methods

UNIVERSITEIT VAN AMSTERDAM
Amsterdam School of Economics

This work was carried out in the Amsterdam School of Economics Research Institute.

ISBN: 978 90 9027305 1

Typeset by LaTeX.

Quantifying Biometric Life Insurance Risks

With Non-Parametric Smoothing Methods

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 18 Januari 2013, te 10:00 uur

door

Julien Tomas

geboren te Valence, Frankrijk.

Promotor:	Prof. dr. R. Kaas (Universiteit van Amsterdam)
Co-promotor:	Prof. dr. F. Planchet (Université Lyon 1 - ISFA)
Overige leden:	Dr. K. Antonio (Universiteit van Amsterdam)
	Prof. dr. A. Charpentier (Université du Québec à Montréal)
	Prof. dr. M.J. Goovaerts (Universiteit van Amsterdam)
	Prof. dr. A. De Schepper(Universiteit Antwerpen)
	Prof. dr. M.H. Vellekoop (Universiteit van Amsterdam)

Faculteit Economie en Bedrijfskunde

To my grandfather, Antonio
To my parents, Elisabeth and Jean-Paul

Avec amour

Preface

The days of solitary research are long gone, and this research could not have been possible without the help and support of a great number of people.

This thesis is the result of a fruitful and comprehensive cooperation with my co-promotor Frédéric Planchet whom I would like to thank first and foremost. His deep multidisciplinary knowledge combined with immense energy and enthusiasm, friendly attitude and endless patience were the main ingredients of my academic development. Furthermore, I would like to thank my promotor Rok Kaas for making this project possible and encouraging my research. Rob's pleasant cooperation, helpful remarks and sustained precision have increased the quality of this thesis. I am also very grateful to the committee, Katrien Antonio, Arthur Charpentier, Marc Goovaerts, Ann De Schepper and Michel Vellekoop for inspecting my thesis.

I would like to thank in particular Pascal Schoenmaekers who introduced me into the world of life insurance, in 2009. Thanks for making my stay at the Financial Solutions Life and Health divisional unit in Munich Reinsurance Company into an inspiring and a productive experience that laid the basis of this thesis.

I also thank the University of Amsterdam for providing their generous financial support.

It has been a great pleasure to work at the Amsterdam School of Economics Research Institute. I would like to express my special gratitude to the staff members, Ida Delponte, Kees Nieuwland and Andries Jansen for their excellent assistance. I am grateful to all my (former) colleagues for creating a pleasant and constructive atmosphere, in particular Willem Jan Willemse, Roger Laeven, André Klein, Angela van Heerwaarden and Henk Wolthuis. Thanks to my office-mates, Zhenzhen, Jan and Frank.

I would like to thank my friends, Guillaume, Jamaal, François, Hélène, Philipe, Marianne. Their encouragements have always be a source of power and they show me that there is a lot of beauty in the world, even in just small and simple things.

I would like also to thank Stephane and Nicolas for giving to my stay in Amsterdam a typical French taste, especially when playing our numerous *jeux de boules* along the Amstel.

I would like to gratefully thank Sun for her forbearance and for being always close to me while I have spent the last years of this dissertation.

And the last but most important acknowledgment. I am extremely grateful to my grandfather, Antonio and my mother and father, Elisabeth and Jean-Paul. They have always been ready to show their understanding and trust for the interests that I pursue, to the greatest extent they can have. I am indebted to them. Their unreserved love and support for these years far from their home is what makes this dissertation valuable. Despite the actual distance and the separation, I have always felt them next to me.

Julien Tomas
Amsterdam, December 2012

Contents

List of Publications

- J.Tomas (2011). A local likelihood approach to univariate graduation of mortality. *Bulletin Français d'Actuariat*, **11**(22), 105-153.
- J.Tomas (2012a). Univariate graduation of mortality by local polynomial regression. *Bulletin Français d'Actuariat*, **12**(23), 5-58.
- J.Tomas (2012b). Essays on boundaries effects and practical considerations for graduation of mortality by local likelihood models. *Insurance and Risk Management*, forthcoming.
- J.Tomas and F.Planchet (2012a). Multidimensional smoothing by adaptive local kernel-weighted log-likelihood with application to long-term care insurance. *ISFA - Laboratoire SAF Working paper - 2012.8 - Submitted to Insurance: Mathematics & Economics.*
- J.Tomas and F.Planchet (2012b). Essays on the construction and validation of specific prospective mortality tables. *ISFA - Laboratoire SAF Working paper.*

Chapter 1

Context and motivations

1.1 Context

Outside of the world of property or liability insurance, life insurance occupies a separate place that it deserves in more ways than one. It emerges as an atypical island teeming with singularities. We can report for example a legal environment of its own, dedicated accounting rules, a specific technical approach, and more generally, principles of functioning that diverge from the foundational philosophy of other branches. In a life insurance contract, the concepts of injury, repair or compensation remain absent in the contractual terms. The guarantees are fixed and freely consented in advance at the time of subscription. Benefits are paid without reference to a financial damage sustained or caused. This positioning also leads to the idea that one can give a value to life and this heretic idea was not easy to admit.

In the following, we present the heuristic evolution of the analysis of mortality. We discuss briefly the mathematical developments and mental changes toward viewing death as a proper subject of human and mathematical investigation and not the concern of god alone. With few exceptions it was mathematicians and astronomers who built the mortality table that deserves to be considered as one of the crowning achievements of the scientific revolution.

1.1.1 The origins of life tables and population dynamics studies

> «*From these Considerations I have formed the Adjoyned Table, whose Uses are manifold, and give a more just Idea of the State and Condition of Manking, than nay thing yet extant that I know of. It exhibits the Number of People in the City of Breslaw of all Ages, from the Birth to extream Old Age, and thereby shews the chances of mortality at all Ages, and likewise how to make a certain Estimate of the value of Annuities for Lives, which*

> *hitherto has been only done by an imaginary valuation: Also the Chances that there are that a Person of any Age proposed does live to any other Age given; with many more, as I shall hereafter shew.»*.
>
> Halley (1693, p.600)

The idea that one can give a value to life runs up through the history against ethical, religious and political considerations leading to prohibit this life insurance, viewed as intrinsically immoral, *malum omen non est providendum.*
In the late Middle Ages, the traditional christian conception of death forbids speculation about it, and thus the idea that there may be laws - other than god - that can explain it. This christian view of a *divine order* - which was that a man died by the will of god who offered the paradise as a reward or hell as a damnation - seemed to respond to an older belief for which death was following physical and deterministic laws. As recalled by Charpentier (2007) the first civilization of Mesopotamia believed in the concept of climacteric age, meaning a critical year marked by fatal accidents in which astrologists claim that considerable alterations appear in the body that leads to illness and death. The climacteric ages are multiples of seven or nine where the danger of death is much larger than the others. This idea, born from astrologists, is found as well in Europe and Japan, and among philosophers and mathematicians like Gottfried Wilhem Leibniz, see Rohrbasser and Véron (1998, p.32). Briefly, the idea that there are physical *laws* for the death or accidents, although contested by the christians, is relatively old.

Insurances linked to life expectancy requires the existence of tables. However at the beginning, such tables have appeared to answer other needs. The idea has sprouted in Rome. In the early 3rd century, the jurist Ulpian (Dometius Ulpianus), perhaps to be considered as the father of actuaries, devised a table for the legal conversion of a life annuity to an annuity certain and identified that the values of annuities should be based on the age of the beneficiaries. But it was much later that these tables were created. To build a table, one needed a census to know the distribution of a population by age (with reliable years of birth). If some brilliant mathematicians have done much for the conceptualization of probabilities, we must remember that is a merchant, John Grant and his friend William Petty, one of the founders of the Royal Society in London, who first conceived the notion of a mortality table. However, Le Bras (2000) asks who between John Grant, and William Petty has first conceived this notion? The question would be insignificant but for a philosophical issue about the role of demography. Le Bras (2000) explains that John Grant represents the *plebeian* who works with a scientific method away from the oligarchy. While William Petty is close to the political power, he has succeeded in the oligarchy instead of following a modest and detached existence such as expected from scientists. In other words, by the choice of its founding hero, demography is defined either as a pure science or as an instrument at the service of a state, because we should remember that since the 17th century, the population represents the wealth of nations and the power of the states. William Petty understood that this *new science* referred to a political project and not the converse.

The political origin of the life table is English, but the economical origin appeared in the Netherlands. Johan de Witt, in 1671, implemented a method rather pragmatic and empirical to calculate the annuities. His method allowed many mathematicians to address the issue by introducing probabilities on the duration of human life.

If the notion of life expectancy has arisen for the first time in 1746 in the work of Antoine Deparcieux, "Essai sur les probabilités de la durée de la vie humaine" see Charpentier (2007), Véron and Rohrbasser (2000, p.11) note that calculations done by Lodewijk Huygens appear in his mail in 1669 with his brother Christiaan where he estimated that his brother will live until the age of 56 and half, and him only until age 55.

At Breslau (belonging to the Habsburg empire, now in Poland and called Wroclaw), registers of births and deaths according to gender and age had been kept since the end of the 16th century. Hald (1990) recalls that a prominent evangelical pastor and scientist, Caspar Neumann, used the list, in 1687 and the following years, in his attempts to fight popular superstitions about the influence on health of the phases of the moon and the climacteric ages.

Neumann sent his results to Leibniz, who in 1689 informed Henry Justell, secretary of the Royal Society in London, of Neumann researches, see Dupâquier (1985). Justell therefore wrote to Neumann who responded by sending his observations for each of the years 1687-1691. The Society asked Edmond Halley to analyze the data and Halley (1693) presented a table with the number of people living in an age group. From this material, some figures of modern science hypothesized the first age patterns of adult mortality and deduced the associated life tables, i.e, the corresponding survivors.

In 1740, Nicolaas Struyck pointed out that the value of annuities should be calculated from life tables based on observations (as done by Halley) and not from hypotheses (as done by de Witt) see Hald (1990, p.395). However, he considered the construction of Halley's table as unsatisfactory because Halley had access to the number of deaths only and not to the corresponding number of living. He wished to provide a reliable life table for annuitants. His observations comprise 794 male and 876 female annuitants who bought their annuities in Amsterdam in 1672-1674 and 1686-1689. For each five-year group, he tabulates the number of annuitants entering at a given age and the number of survivors at any later age. Assuming that mortality at a given age did not change over time, he summed the number exposed to risk and the number of deaths for each age group. He calculated the rates of mortality from which he derived a table that corresponds to the form still used today. He stressed that the mortality of females was smaller than that of males and presented the first life tables for males and females separately.

Perhaps the first statistical results to be taken seriously were the Northampton tables of 1780, devised by Richard Price. He worked from parish registers in Northampton, and produced corresponding tables. Price's tables were not very conservative for the annuities. In 1808 the British government, hard-pressed by war and inflation, decided to issue annuities based on Price's Tables. Hence it lost millions of pounds because people lived longer than was implied by the table, see Hacking (1975, p113-114). But the first table that became the usual standard of British and American insurance companies for nearly a century is the table known as the Carlisle table, built in 1815 by Joshua Milne on the basis of statistics from parishes in Carlisle.

This element of the panoply of the perfect actuary is so essential today that it is sometimes hard to imagine that it has only more than two centuries of existence. In fact the invention was not so simple, as we have seen. It is the result of the meeting of two favorable events. The first is the scientific invention of probabilities. The second, much more down to earth, is the growing need for actuaries to refine their calculation of annuities. Thus, from 1662 to 1766, from Grant and Petty to Depracieux and Milne, through Leibniz and the Huygens brothers, Halley and Struyk, actuaries on one side and mathematicians and astronomers on the other tackled the same questions about the duration of life, each bringing his stone to the edifice, and finally built in 100 years, after many hesitations, the mortality table.

Figure 1.1 compares the survival functions at birth issued from the different tables. We note that the survival curves move towards a rectangular shape. We use the term rectangularization to describe this feature: the more the time passed, the more the probability of death becomes flat at younger ages (one died rarely before 60 years), then much more brutal one the end. The point of maximum downward slope of the survival curve progressively moves toward the very old ages. This feature is called the expansion of the survival function, see Pitacco *et al.* (2009, p.53).

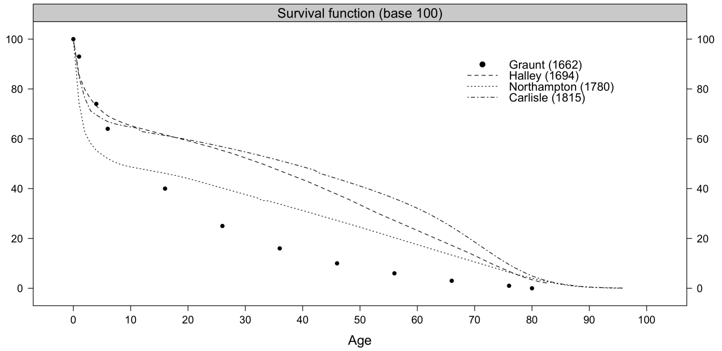


Figure 1.1: *Survival functions at birth issued from the different tables. Source: Hald (1990), Halley (1693), Gompertz (1825) and Gompertz (1871)*

Around 1870, demographers particularly in Germany felt the need for a simple chart to present population dynamics, especially in view of establishing life table formulas. This chart is known as the *Lexis Diagram*, but it is a misnomer according to Vandeschrick (2001).

To be useful, this chart must allow for location on one plane of three coordinates used to classify deaths and survivors, namely: the date, the age and the moment of birth.

Briefly, there were three solutions for this problem: In 1869, Gustav Zeuner worked out a first solution. In 1870, Otto Brasche proposed a second one with networks of parallels; his version is the most currently used now. In 1874, Karl Becker proposed a third one. In 1875, Wilhelm Lexis took back Zeuner's diagram and just added networks of parallels. In spite of all this, the name *Lexis Diagram* is now used universally.

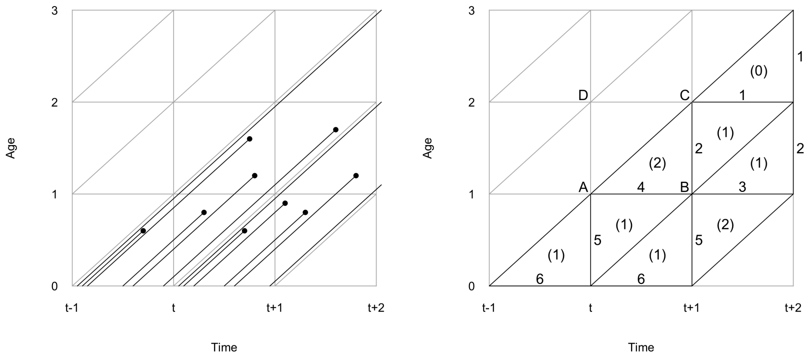


Figure 1.2: *Left panel: Lexis diagram containing life-times for birth cohorts of $t-1$ and t. Each individual is presented as a line in a time-age plane, and points denote the death for a given individual. Right panel: Lexis diagram containing counts of events pertaining to birth cohorts of $t-1$ to t.*

Figure 1.2, left panel, shows a simplified version of a Lexis diagram. In this diagram, an individual life history is drawn as a line segment with slope 1. This line starts on the horizontal axis at the time of birth and ends at the time of death. The value on the vertical axis is the individual's age. Hence a life-time starts at zero (birth) and ends at the age of death. In this way data are properly represented according to the three demographic coordinates. The individual life-time can be regrouped and hence the *Lexis Diagram* also allows a summary of aggregated death and population data by age, period and cohort. For instance, in Figure 1.2, right panel, from the birth cohort of six births during period t: (1) death in t and five survivors to the beginning of the following period $t+1$; (2) deaths at age 0 in $t+1$ and three survivors to age 1; (1) death to the cohort at age 1 during $t+1$ and two survivors to the beginning of the period $t+2$. The *Lexis Diagram* has become a standard tool for summarizing population dynamics.

1.1.2 Measures of mortality: Notation

Probabilities of survival and death

This section makes precise the notation used in this dissertation to quantity the biometric life insurance risks. We refer to Pitacco *et al.* (2009) for more details. The age at which a person will die is obviously unknown. At most we can evaluate, for a particular population, the risk of death in a given time interval. Death is then viewed as an event whose occurrence is probabilistic in nature and it is natural to resort to a mathematical framework and probabilities calculus to describe the life time of individuals.

We consider a person aged x, and denote by T_x the random variable representing his/her remaining lifetime. In actuarial notation, probabilities like $\mathbb{P}[T_x > h]$ and $\mathbb{P}[h < T_x \leq h+k]$ are usually involved. When a life table is available, these probabilities can be immediately derived from the life table itself, provided that the ages and durations are integers.

In life insurance mathematics, a specific notation is commonly used for the probabilities of survival or death. The notation for the survival probability is as follows,

$$ {}_hp_x = \mathbb{P}[T_x > h], \text{where } h \text{ is an integer.} \tag{1.1}$$

In particular ${}_1p_x$ is simply denoted p_x. Trivially ${}_0p_x = 1$.

The notation for the probability of death is as follows,

$$ {}_{h|k}q_x = \mathbb{P}[h < T_x \leq h+k]. \tag{1.2}$$

If $h = 0$, the notation ${}_kq_x$ is used. In particular, when $h = 0$ and $k = 1$, the symbol q_x is commonly adopted. Clearly, ${}_0q_x = 0$.

Note that in all symbols, the right-hand side subscript denotes the age being considered. Conversely, the left-hand side subscript refers to the duration, whose meaning depends on the specific probability addressed. The purpose of measuring the life span or conversely the mortality is to enable inferences to be drawn about the likelihood of death occurring within a specific population during a specific period of time. It is natural, therefore, for the basic measure to be expressed in proportional terms. The denominator (of which the numerator is the relevant number of deaths) is commonly referred to as *population at risk* or *exposed to risk.* To be specific, let us assume that we are given the number of deaths recorded, d_x, and the number of individuals initially exposed to the risk of death, l_x, all aged x last birthday, and that our experience, for simplicity, is limited to this single age x, where $x = 1, 2, \ldots, n$. The observed estimate of the one-year probability of death is denoted by q_x,

$$ q_x = 1 - \frac{l_{x+1}}{l_x} = \frac{d_x}{l_x}. \tag{1.3}$$

In Figure 1.2, let Z_{AD} the number of life-lines crossing segments AD and Y_{ABCD} the number of deaths in the square $ABCD$, then equation (1.3) is Y_{ABCD}/Z_{AD}. For the observed annual survival probability, we have

$$ p_x = 1 - q_x. $$

In general for the observed survival probability, we have,

$$_hp_x = p_x\ p_{x+1} \cdots p_{x+h-1} = \frac{l_{x+h}}{l_x},$$

while for the observed probability of dying,

$$_kq_x = 1 -_k p_x = 1 - \frac{l_x + k}{l_x},$$

and

$$_{h|k}q_x =_h p_x\ _kq_{x+h} = \frac{l_{x+h} - l_{x+h+k}}{l_x}.$$

Survival function

Suppose that we have to evaluate the probability of survival and of dying when age and times are real numbers. Tools other than the life table are then needed. We move now to an age-continuous context.
We call $S(t)$ the survival function and define it for $t \geq 0$ as follows,

$$S(t) = \mathbb{P}[T_0 > t],$$

where T_0 denotes the random lifetime for a newborn. Considering the probability (1.1), we have

$$\mathbb{P}[T_x > h] = \mathbb{P}[T_0 > x + h | T_0 >] = \frac{\mathbb{P}[T_0 > x + h]}{\mathbb{P}[T_0 > x]},$$

and thus

$$_hp_x = \frac{S(x+h)}{S(x)}.$$

For the probability (1.2), we obtain

$$_{h|k}q_x = \frac{S(x+h) - S(x+h+k)}{S(x)},$$

and in particular,

$$_kq_x = \frac{S(x) - S(x+k)}{S(x)}.$$

Turning back to the mortality table, we note that since l_x is the expected number of people alive at age x out of a cohort initially composed of l_0 individuals, we have

$$l_x = l_0\mathbb{P}[T_0 > x],$$

and in terms of the survival function, $l_x = l_0 S(x)$, provided that all individuals have the same age-pattern of mortality described by $S(x)$. Thus, the l_x's are proportional to the values which the survival function takes on integer ages x, and so the mortality table can be interpreted as a tabulation of the survival function, see Pitacco *et al.* (2009, p.52).

Forces of mortality

We consider the probability of an individual age x of dying before age $x+t$ (with x and t real numbers), namely ${}_tq_x$. The force of mortality (or mortality intensity) is defined as

$$\varphi_x = \lim_{t\to 0} \frac{\mathbb{P}[T_x \le t]}{t} = \lim_{t\to 0} \frac{{}_tq_x}{t},$$

hence it represents the instantaneous rate of mortality at a given age x. In terms of the survival function,

$$\varphi_x = \frac{-\frac{d}{dx}S(x)}{S(x)} = -\frac{d}{dx}\ln S(x),$$

so

$$S(x) = \exp\left(-\int_0^x \varphi_u \, du\right).$$

Central death rates

The behavior of the force of mortality in the interval $(x, x+1)$ can be summarized by the central death rate at age x,

$$m_x = \frac{\int_0^1 S(x+u)\varphi_{x+u}du}{\int_0^1 S(x+u)du} = \frac{S(x)-S(x+1)}{\int_0^1 S(x+u)du}. \tag{1.4}$$

The integral $\int_0^1 S(x+u)du$ can be approximated using the trapezoidal rule. In Figure 1.2, let Z_{AD} and Z_{BC}, the number of life-lines crossing segments AD and BC respectively, and Y_{ABCD} the number of deaths in the square $ABCD$, then the central death rate is approximated by $Y_{ABCD}/((Z_{AD}+Z_{BC})/2)$, and

$$m_x = \frac{S(x)-S(x+1)}{(S(x)+S(x+1))/2}.$$

With the assumption of constant force of mortality - frequently adopted in actuarial science calculations - which assumes $\varphi_{x+t} = \varphi_x$ for $0 \le t < 1$, we obtain, from (1.4),

$$m_x = \varphi_x.$$

1.1.3 Portraying mortality over age and over age and time

Portraying mortality over age

Figure 1.3 displays the one-year transformed crude probabilities of death (year 2008), logit scale, for ages $x = 0, 1, \ldots, 98$ and each gender for the dutch population provided by the Human Mortality Database (2012). The Human Mortality Database (HMD) was initiated by the Department of Demography at the University of California Berkeley, USA, and the Max Planck Institute for Demographic Research, Rostock, Germany. This international project provides detailed mortality and population data that can be accessed online for research purposes.

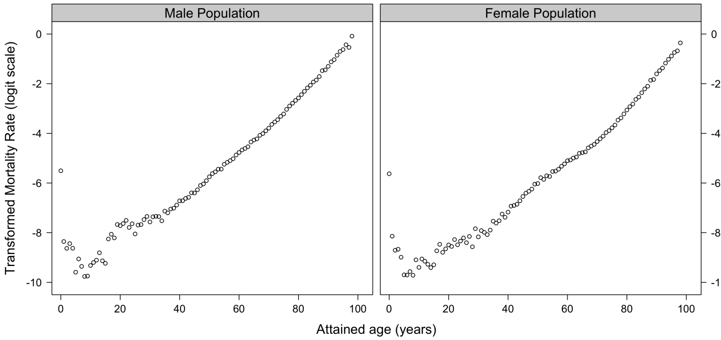


Figure 1.3: *Transformed crude one-year probabilities of death, logit scale, for Dutch Male (left panel) and Dutch females (right panel) in* 2008*. Source: HMD.*

From Figure 1.3, we recognize the typical shape of a mortality curve. Mortality is highest at the extremes of age. Once the newborn infant has survived the hazard of the first days of life, the rate of mortality falls rapidly. Most of the deaths after the first days are due to exogenous causes, mainly infections and until recent times when this component has shrunk to very small proportions, the rate was a sensitive index of social conditions and of public health progress. During childhood the risk of death is very small, being very largely confined to that of the occasional lethal infection, which modern treatments have made extremely rare, and severe accidental injuries to which child risk recklessness or lack of adult care sometimes leads. In adolescence, the impact and strain of industrial and urban life bring a rise in mortality. These and other factors, inherent in the social and economic environment and individual ways of life, reacting upon constitutional weakness, lead to a continuing increase in the risk of death as age advances. At later ages, the wearing out of the human frame rather than inimical qualities of the environment becomes the dominant cause of mortality, see Benjamin and Pollard (1980).
We show in Figure 1.3 the difference in the patterns of mortality for the two genders. The death rates for females are lower than those for males at all ages. (Before 1890 there was an excess in the death rate of females at adolescence and early adult ages mainly associated with the heavier mortality from tuberculosis in girls). Briefly, the higher mortality of males may be explained in medical terms as follows, see Benjamin and Pollard (1980) for more details.

In infancy and early childhood, boys are generally more vulnerable to some birth hazards (prematurity, malformation, birth injury), to infection (possibly as a result of some biological factors) and to injuries (possibly as a result of more vigorous and venturesome activities). These are the principal causes of death at those ages.

In early and middle adult life, the principal causes of death are accidents and violence, heart diseases and cancers. The higher risk for accidents must be regarded as occupational in the broader sense of including, as compared with females, more outdoor movement in traffic for instance, as well as greater industrial hazards.

At more advanced ages, the process of physical deterioration and lessening resistance to disease associated with general wear and tear appear to proceed faster in men. Age for age, cerebral hemorrhages, arterial diseases, cancers (especially of the lungs) and bronchitis take a heavier toll of males than females. Some, at least, of this excess mortality has been self inflicted by cigarette smoking. The contemporary increase in industrial countries of mortality cancer of the lung and coronary arterial disease (especially for men) has been exercising considerable influence on the shape of the curve of death rates with age.

Portraying mortality over age and time

Figures 1.4 and 1.5 display the mortality surfaces and level plots for the Dutch males and females respectively. We see that the surface is subjected to period shocks corresponding to wars, epidemics, summer heat waves, and so on. It is apparent that dramatic changes in mortality have occurred over the 20th century, as illustrated by the downward trends and variations in shape.

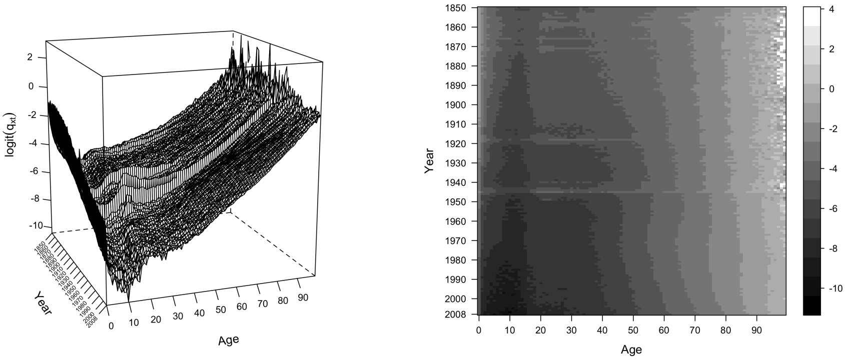


Figure 1.4: *Surface and level plot of the observed one-year probabilities of death, logit scale, for Dutch males, period* 1850-2008. *Source: HMD.*

Figures 1.6 and 1.7 depict the observed annual probabilities of death, for some selected periods. The mortality has decreased for both sexes and all ages without interruption, primarily due to the control of infectious diseases.

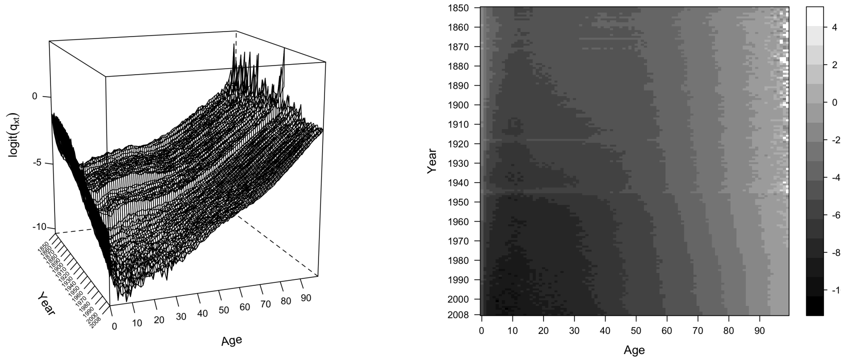


Figure 1.5: *Surface and level plot of the observed one-year probabilities of death, logit scale, for Dutch females, period* 1850-2008*. Source: HMD.*

This reduction is stronger for the young ages. The decrease over time at ages 20-30 for the females reflects the rapid decline in childbearing mortality. However, the hump in mortality around ages 18-25 has become increasingly important especially for the young males. The increase of life expectancy has continued to the late 20th century with the decline in mortality at the highest ages, mainly due to the reduction of mortality from cardiovascular diseases.

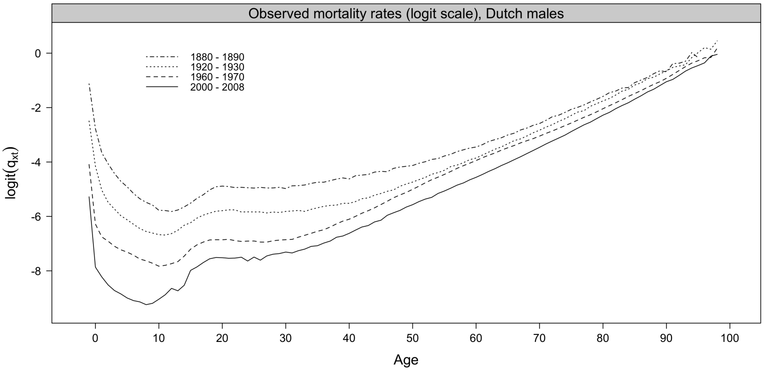


Figure 1.6: *One-year probabilities of death, logit scale, for Dutch males from period life tables* 1880-1890*,* 1920-1930*,* 1960-1970 *and* 2000-2008*. Source: HMD.*

The trend in the observed annual probabilities of death are displayed in Figures 1.8 and 1.9, for Dutch males and females respectively. When we examine Figure 1.8, we see different behavior for age-specific probabilities of death affecting Dutch males. At age 20, a rapid reduction in mortality took place after a peak in the early 1940s due the World War II. However,

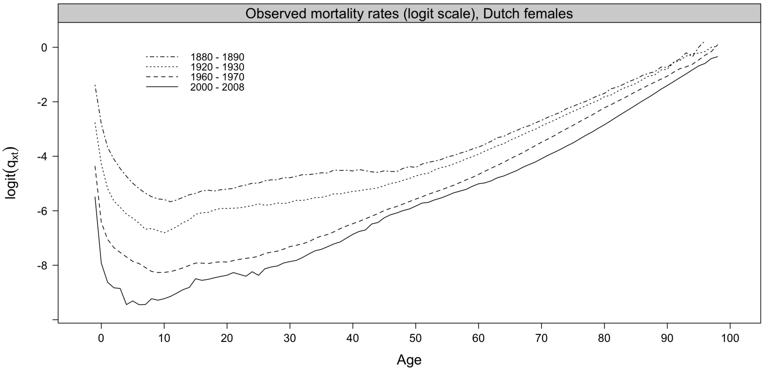


Figure 1.7: *One-year probabilities of death, logit scale, for Dutch females from period life tables* 1880 − 1890*,* 1920 − 1930*,* 1960 − 1970 *and* 2000 − 2008*. Source: HMD.*

since the 1950s, only modest improvements have occurred. This is typical for ages around the accident hump, as explained in Pitacco *et al.* (2009, p.98); male mortality has not really decreased since the 1970s. We even observe an increase of the mortality. This unfavorable evolution is due to the increase of traffic accidents particularly acute in the 1960s. Between 1980 and the mid-1990s, the apparition of AIDS had a negative influence on the reduction of mortality. At age 40, the same decrease is present after the World War II, followed by a much slower reduction in mortality after 1960. The decrease after 1970 is more marked than at age 20. At age 60, the mortality rates have declined rapidly after 1970, whereas the decreasing during 1850-1970 was more moderate. At age 80, this decrease appears after 1990.

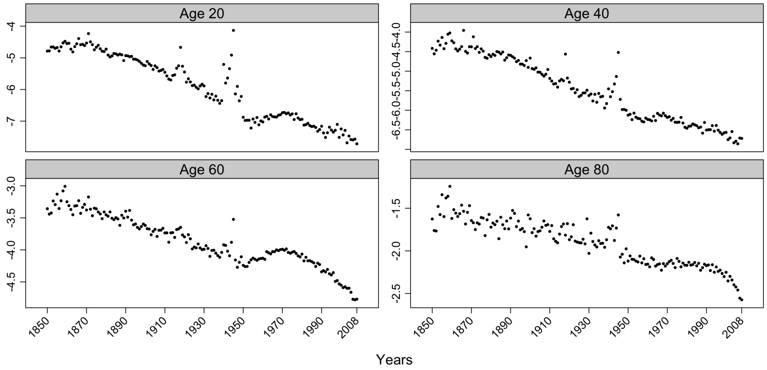


Figure 1.8: *Trend in the observed probabilities of death, logit scale, for Dutch males at ages* 20*,* 40*,* 60 *and* 80*, period* 1850-2008*. Source: HMD.*

The analysis for the Dutch females is similar to the one for the male population for age 20 and 40, but with several differences. At age 20, a structural

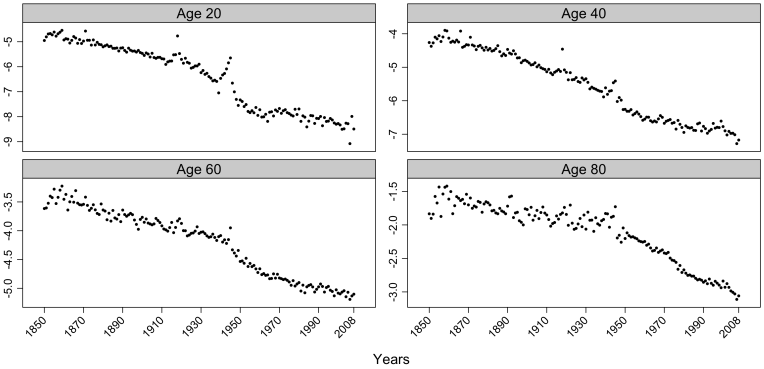


Figure 1.9: *Trend in the observed probabilities of death, logit scale, for Dutch females at ages* 20*,* 40*,* 60 *and* 80*, period* 1850-2008*. Source: HMD.*

break seems to have occurred, with a relatively high level of mortality before the second world war and a much lower one after 1950. Then after the mid-1950s modest improvements are visible. At age 40, the decline is more pronounced after 1960 than for the male population. At age 60, the rate of decrease is more regular. At age 80, after 1950, the trend in the reduction of mortality has tended to accelerate.

Until 1980, females have benefited more from the reduction of mortality than males, and the gap in life expectancy has widened significantly between the genders. Nevertheless, in the last three decades, the gap has stabilized and begun to decline. This reduction is essentially due to an acceleration in the improvement among the males and some slowing of the improvement among females under age 60. At the later ages, on the other hand, improvement continued to be more rapid for females than males. Although cancer mortality is falling for both men and women, cancer is now the leading cause of death, overtaking cardiovascular diseases, for which mortality has considerably reduced, see Meslé (2006). Future improvement will depend on success in the control of cancer and neuron-degenerative diseases.

1.1.4 The irregularities in the progression of the observed rates

The symbol q_x represents the one-year observed probability of death for a particular population at age x. It lies above or below the true underlying value. From Figures 1.4 and 1.5, the roughness of the surface indicates volatility. In estimating mortality, the actuary knows that the past experience from which the observed mortality rates and the life table have been derived will never be exactly reproduced in the future. Thus a certain random element of fluctuation will be inherent in the observations and the smaller the group, the greater will be the relative random errors in the deaths and the less reliable will be the resulting estimates.

These deviations from the true underlying rates may be assumed to be random and to fluctuate from age to age both in size and sign. These irregularities in the progression of the observed rates of mortality could be reduced by increasing the number l_x of persons observed. If the number of individuals in the group had been considerably larger, the set of observed probabilities, q_x, would have displayed a much more regular progression with x. In the limit, it would have exhibited a smooth progression, as explained in Copas and Haberman (1983).

The idea of a group of persons attaining age x and being gradually reduced in numbers, until they are all dead, in such a way that the rates of mortality at successive ages form a smooth series is a purely theoretical conception. It is nevertheless a very useful conception, as recalls Alistair (1989), which forms the basis of the theory of life contingencies and has been shown by long use to be suitable for solving most actuarial problems in life insurance. This is not to suggest that measurement can be allowed to be inexact. On the contrary, as Benjamin and Pollard (1980) mention, if judgment has to be introduced in any final estimation, it is likely to be sounder when on the basis of adequate analysis of past experience.

Provided these errors are random in nature, they may be reduced by increasing the size of the sample and thereby extending the scope of the investigation. A simpler, cheaper and more practicable alternative is often to use graduation to partly remove these random errors. Thus, by graduating the mortality rates, we aim to concentrate on the underlying mortality pattern (high mortality at birth, low infant mortality, accident hump, senescence effect) avoiding the erratic departures from it.
Various approaches to graduation can be adopted. In particular, two broad categories can be recognized:

i. Parametric approaches, involving the use of mortality laws; Hannerz (2001) defines a mortality law as a mathematical expression that describes mortality as a function of age.

ii. Non-parametric approaches.

1.2 Motivations

1.2.1 Getting out of a procrustean bed of fixed parametrization: From parametric to smooth models

Assume n pairs of observations $\{(x_i, q_i)\}_{i=1}^{n}$ have been collected, then the regression relationship can be modeled as

$$q_i = f(x_i) + u_i, \qquad i = 1, 2, \ldots, n;$$

with the unknown regression function f and an error term u_i, representing random errors in the observations or variability from sources not included in the x_i.

The aim of a regression analysis is to produce a reasonable analysis of the unknown response function f. This task of approximating the mean function can be done essentially in two ways. The quite often used *parametric* approach is to assume that the mean curve f has some pre-specified functional form, for instance, a line with unknown slope and intercept. As an alternative one could try to estimate f *non-parametrically* without reference to a specific form.

The first approach to analyze a regression relationship is called parametric since it assumes that the functional form (for example, Thiele law, Perks laws, Gompertz-Makeham class of models, and so on) is fully described by a finite set of parameters. A tacit assumption of the parametric approach though is that the curve can be represented in terms of the parametric model or that, at least, it is believed that the approximation bias of the best parametric fit is a negligible quantity. Such laws simplify the calculation of mortality functions and allow to extrapolate at the highest ages for instance. But to be useful, they have to reproduce closely the data. According to Alistair (1989) it is now thought that it is unlikely that a law can be found that represents the mortality rate over a large range of ages, although some complicated expressions have been used in the attempt.

By contrast, non-parametric modeling of regression relationship does not project the observed data into a Procrustean bed of a fixed parametrization. A preselected parametric model might be too restricted or too low-dimensional to fit unexpected features, whereas the non-parametric approach offers a flexible tool in analyzing unknown regression relationship. The term *non-parametric* thus refers to the flexible functional form of the regression curve. Like parametric methods, they too are liable to give biased estimates, but in such a way that it is possible to balance an increase in bias with a decrease in sampling variation.

The question of which approach should be taken in data analysis was a key issue in a bitter fight between Pearson and Fisher in the 1920's, as recalls Härdle (1990). Fisher pointed out that the non-parametric approach gave generally poor efficiency whereas Pearson was more concerned about the specification question. Both points of view are interesting in their own right. Pearson pointed out that the price we have to pay for pure parametric fitting is the possibly of gross misspecification resulting in too high model bias. On the other hand, Fisher was concerned about a too pure consideration of parameter-free models which may result in more variable estimates, especially for small sample size.

1.2.2 Natura non agit per saltum: The basic idea of smoothing

We have previously seen that the crude rates can be seen as a sample from a larger population of lives and thus they contain some random fluctuations.

If we believed that the true rates were independent, then the crude rates would be our final estimate of the true underlying mortality rates. However, a common prior opinion about the form of the true rates is that each true rate of mortality is closely related to its neighbors, that is the observations q_j in the neighborhood of the target point q_i should contain information about the value of f at x_i. Gavin *et al.* (1993) explain that this relationship is expressed by the belief that the true rates progress smoothly from one age to the next.

Benjamin and Pollard (1980) recall the saying, *Natura non agit per saltum*, which expresses the fact that natural forces operate gradually and their effects become apparent continuously and not in sudden jumps. It follows that the data for several ages x_j on either side of age x_i can be used to augment the basic information we have at age x_i, and an improved estimate of q_i can be obtained by smoothing the individual estimates.

So the next step is to graduate the crude rates in order to remove any random fluctuation. This procedure of approximation of the mean response curve $f()$ is commonly called *smoothing.* Hence, the mortality is not summarized by a small number of parameters, but described by the n annual probabilities of dying. It may be considered as a compromise between faith in the data and reduced roughness caused by the noise. In the actuarial literature, the process of smoothing a mortality table was known as graduating the data, the little hills and valleys of the rough data were to be graded into smoothness, just as in building a road over rough terrain.

The concept of smoothness has been used in the previous paragraphs without actually being defined. We deliberately avoid a detailed presentation here. The interested reader can have a look at Bizley (1958) and Diewert and Wales (2006). We all have an intuitive idea about what we mean by smooth, as for instance the road building analogy.
Formal mathematical analysis may state the smoothness condition as a bound on derivatives of f. Bizley (1958) observes that smoothness is intimately concerned with predictability, and proposes the following definition of smoothness: a continuous curve is smooth at the points for which the absolute value of the rate of change of curvature with respect to distance measured along the curve is small. For Benjamin and Pollard (1980), the Bizley's requirements of small change of curvature turns out to be equivalent in the mortality context to requiring that third-order differences are small.

1.2.3 Smoothers and parameters selection

Smoothing alone, however, is not graduation. Graduated rates must be representative of the underlying data. The two qualities, *smoothness* and *goodness of fit*, tend to conflict, in the sense that smoothness may not be improved beyond a certain point without some sacrifice of goodness of fit and vice versa. Thus, a graduation will often turn out to be a compromise between optimal fit and optimal smoothness.

To be useful, the method should allow the graduator some latitude in choosing the relative emphasis to place smoothness and fit.

Special attention has to be paid to the fact that smoothers, by definition, average over observations with different mean values. The amount of averaging is controlled by a weight sequence which is tuned by a smoothing parameter, denoted λ. This smoothing parameter regulates the size of the neighborhood around the target point x_i.
A local average over a too large neighborhood would cast away the good with the bad. In this situation an extremely *over-smooth* curve would be produced, resulting in a wrong estimate $\widehat{f}$. On the other hand, defining the smoothing parameter so that it corresponds to a very small neighborhood would not sift the chaff from the wheat. Only a small number of observations would contribute non-negligibly to the estimate $\widehat{f}(x_i)$ at x_i making it very rough and wiggly. In this case the variability of $\widehat{f}(x_i)$ would be inflated. Finding the choice of smoothing parameter that balances the trade-off between *over-smoothing* and *under-smoothing* is called the smoothing parameters selection problem. To give insight into the smoothing parameters selection problem, consider Figure 1.10 below.

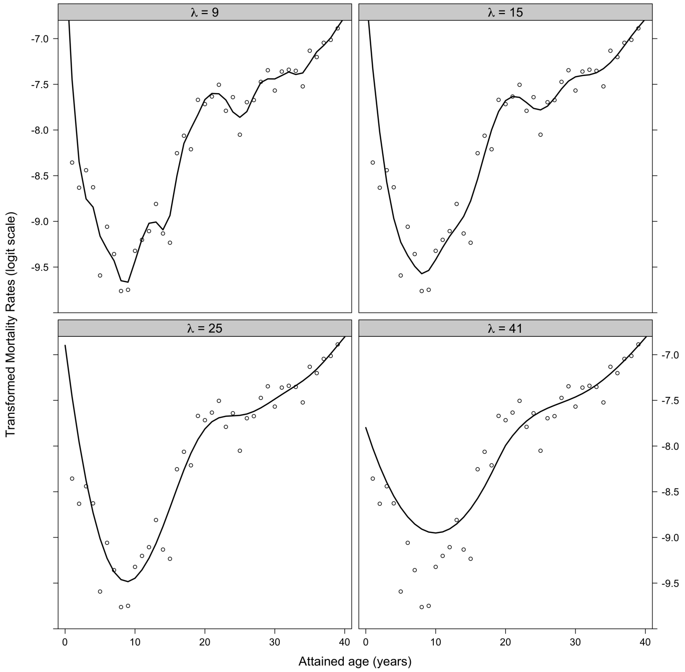


Figure 1.10: *Estimated curve and transformed crude mortality rates (dots), logit scale, for Dutch Male 2008. Source: HMD.*

The curves represent non-parametric estimates of the mortality rates. The more wiggly curve has been computed using a local polynomials estimate with a very small neighborhood. By contrast, the flatter curve has been computed using a very large neighborhood. Which smoothing parameter is correct? The question will be discussed in Section 2.5.

1.2.4 Historical review of the development of smoothing approaches

The problem of smoothing sequences of observations is relevant in many branches of sciences. In the following, we review the development of smoothing methods starting in the late 18th to the 21st centuries, leading up to the development of the use of local polynomial regression and afterward local likelihood methods.

Early work

Local regression is a natural extension of parametric fitting, so natural that local regression arose independently at different points in time and in different countries. The setting for this early work was univariate and involved equally spaced x. It was simple enough that good-performing smoothers could be developed and were computationally feasible by hand calculation. Also, most of the early work arose in actuarial studies, as remark Cleveland and Loader (1996). Mortality and sickness rates were smoothed as a function of age.

Haberman (1996, p.40) reports that smoothing was used as early as 1765 by the Swiss mathematician and physicist Johann Lambert. He was born in Mülhausen, now Mulhouse in Alsace, France; then an exclave of Switzerland. Daw (1980, p.347) explains in his 1765's work (volume 1) that he graduated the value l_x, at decennial ages, which he had calculated from the deaths recorded in the London Bills of Mortality for 1753-1758. He does not read off the graduated values of l_x at all ages from his graph, but gives two methods of graduation and/or interpolation. The first was a graphical method for introducing *osculating parabolas* between two points. The second was a method of fitting a polynomial of fifth degree to represent a section of the curve which was then able to *hang together* with the corresponding polynomials for the immediately preceding and succeeding sections of the curve. This methodology is effectively what has come to be known as *osculatory interpolation*, and was re-invented more than 100 years later by Thomas Sprague.

John Finlaison, subsequently first president of the Institute of Actuaries in January 1823, started preparing the mortality data that were to provide the first life table consisting of graduated observations at individual ages. His 1829 work is described by Seal (1982, p.89). His formula is based on overlapping piecewise linear arcs extending over nine successive values, with eight of the nine being used in the next arc, and thus represents the first published example of a graduation by the adjusted-average method. This piecewise approach to smoothing was extended in 1866 by the Italian meteorologist and astronomer Giovanni Schiaparelli who assumed a cubic polynomial to extend to a stretch of consecutive observed values.

In the same year (1866), Wesley Woolhouse presented a detailed exposition of graduation of mortality rates using summation formulae, stressing the conceptual differences between graduation and interpolation. He considered the case where the fourth differences of the corrections $v_x = \widehat{q}_x - q_x$ to an observed series of rates had small values and proposed to minimize $\sum v_x^2$ in terms of $\Delta^4 v_x$ and thus obtain estimates of v_x and hence $\widehat{q}_x$. Seal (1982, p.93) demonstrates that the equations for $\widehat{q}_x$ are equivalent to those which arise from fitting piecewise cubic polynomials by least squares to equidistant observations.

The use of symmetrical moving weighted average formulae to smooth equally spaced observations of a function of one variable, which generalized Woolhouse's summation formulae, was systematically investigated in a series of papers by the American statistician Erastus De Forest, as reports Haberman (1996, p.41). De Forest's principal innovation was to introduce optimality criteria into the problem of estimating the coefficients.

In 1887, Thomas Sprague's paper on the graphic method of graduation appeared. This paper rediscovered (following Lambert) osculatory interpolation showing how formulae could be devised to ensure continuity of the first derivatives of overlapping interpolation curves. Osculatory interpolation was used as a method of graduation for the English life table in the early nineteenth century.

A new style of summation graduation and its testing had started with Spencer, in 1904 and 1907, and had blossomed in Vaughan's 1933, 1934 and 1935 articles. The method developed by Spencer in his 1904 article had become popular because it was computationally efficient and had good performance. We note three crucial properties. First, the smoother exactly reproduces cubic polynomials as explained in Cleveland and Loader (1996). Second, the smoothing coefficients are a smooth function of length of the bandwidth, and decay smoothly to zero at the ends. Third, the smoothing can be carried out by applying a sequence of smoothers each of which is simple; this was done to facilitate hand computation. Achieving all three of these properties is remarkable.

Whittaker (1923) suggested an alternative method of graduation. This can be regarded as what would now be called a Bayesian approach to graduation, see Taylor (1992). It results in the minimization of the combination of a measure of goodness of fit of the graduation to the observation and a measure of smoothness.

Modern work

We have seen that the methods presented in the introduction are inherited from a long actuarial tradition. However local regression methods received little attention in the statistical literature until the late 1970's.

For Cleveland and Loader (1996), the modern view of smoothing by local regression has origins in the 1950's and 1960's, with kernel methods introduced in the density estimation setting (Rosenblatt (1956), Parzen (1962)) and the regression setting (Watson (1964)). Kernel methods are a special case of local regression; it amounts to choosing the parametric family to consist of constant functions. Kernel methods have found actuarial application by Copas and Haberman (1983), followed by Gavin *et al.* (1993) and Gavin *et al.* (1995).

However, recognizing the weaknesses of a local constant approximation, the more general local regression enjoyed a reincarnation beginning in the late 1970's. It includes the mathematical development of Stone (1977), Stone (1980), and the *lowess* procedure of Cleveland (1979). It provides a number of important insights about the choices of the smoothing parameters. For example it was nearly a given that for most applications the weight function needed to be smooth, that local constant fitting was inadequate, and that smoothers needed to reproduce exactly (and not just asymptotically) at least a quadratic.

Among other features, the local regression method and linear estimation theory trivialize problems that have proven to be major stumbling blocks for more widely studied kernel methods. The kernel estimation literature contains extensive work on bias correction methods: finding modifications that *asymptotically* remove dependence of the bias on the slope, curvature, and so on. Examples include boundary kernels, see Müller (1987), and higher order kernels, see Gasser *et al.* (1985) and Schucany (1989). Local regression methods can then be viewed as an extension of kernel methods and an attempt to extend the theory of kernel methods. This treatment has become popular in the 1990s, for example Hastie and Loader (1993) and to some extent Loader (1999b). The approach has its uses: small bandwidth asymptotic properties of local regression, such as rates of convergence and optimality theory, rely heavily on results for kernel methods. But for practical purposes, the kernel theory is of limited use, since it often provides poor approximations and requires restrictive conditions.

Furthermore, while the early smoothing work was based on an assumption of a near-Gaussian distribution, the modern view extended smoothing to other distributions. Cleveland (1979) developed robust smoothers. Later, Tibshirani and Hastie (1987) took local fitting one step further; in any situation where a dependent variable depends on independent variables, a local likelihood procedure can be carried out. Hence they substantially extended the domain of smoothing to many distributional settings such as logistic regression, and developed general fitting algorithms. The extension to new settings has continued in the 1990's with Fan *et al.* (1998) and Loader (1996).

1.3 Outline of the thesis

In Chapter 2, a non-parametric graduation method is discussed. We introduce local polynomial regression. We discuss the choice of the smoothing parameters and criteria used for models selection. We graduate the data through the choice of the smoothing parameters. The graduation and corresponding confidence intervals are carried out over the entire age range. Tests are used to compare the graduated rates obtained with those obtained by the Whittaker-Henderson smoothing.
Our contribution Tomas (2012a) - that presents extensively local polynomial technique in view of graduating experience data originating from life insurance - can be viewed as the prolongation of the kernel estimation for graduation proposed by Gavin *et al.* (1993). It is completed in this chapter with Tomas (2012b) analyzing the influence of the boundaries on the choice of the smoothing parameters.

In Chapter 3, our aim is to extend the local smoothing technique described in Chapter 2 to model situations where a non Gaussian likelihood is appropriate. We incorporate the concepts of the non-parametric regression technique of local polynomials to localized generalized linear models and local likelihood contexts.
Related work is in Delwarde *et al.* (2004) and Debón *et al.* (2006), but our work examines the statistical properties of the estimators and the choice of the smoothing parameters by classical selectors as well as the plug-in methodology.
The applications cover the graduation of both the probabilities of death and the forces of mortality over the entire age range involving historical data from the Netherlands. In addition we provide a method for constructing pointwise confidence intervals that are not depending on the estimates using the variance stabilizing link. This method allows us to produce confidence intervals in presence of zero-responses.

In Chapters 2 and 3, the weight functions have always had a fixed or global bandwidth. Rather than restricting the smoothing parameters to a fixed value, Chapter 4 discusses more flexible approaches allowing the smoothing parameters to vary across the observations. An application involving individuals subscribing long-term care insurance is presented. We analyze the incidence of mortality as a function of both the age of occurrence of the pathology and the duration of the care. We distinguish the intersection of confidence intervals rule and local bandwidth correction factors.
Part of our work is an extension of the adaptive kernel methods proposed by Gavin *et al.* (1995) to adaptive local kernel-weighted log-likelihoods techniques. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data. Tests and single indices summarizing the lifetime probability distribution are used to compare the graduated series to p-splines smoothing and local likelihood models.

Chapter 5 illustrates the construction and validation of portfolio specific prospective mortality tables. We are interested in the variation of mortality with attained age and calendar year. From portfolios of several insurance companies we construct, in a first step, a global prospective reference table summarizing the mortality experience of these portfolios. We investigate the divergences in the mortality surfaces generated by a number of previously proposed models. We focus on the model risk and, to a lesser extent, on the risk of expert judgment related to the choice of the external references used. We use non-parametric method, namely local kernel-weighted log-likelihood and semi-parametric relational models, to graduate and extrapolate the surfaces. The extrapolation of the smoothed surface, obtained by local likelihood methods, is performed by identifying the mortality components and their importance over time using functional principal component analysis. Then time series methods are used to extrapolate the time-varying parameter, while semi-parametric relational models have the advantage of integrated estimation and forecasting. Tests and indices summarizing the lifetime probability distribution are used to measure the impact of model choices.

The mortality of the entire population is not specific to any subpopulation. The second step of our approach is then to build entity specific prospective mortality tables by adjusting the reference table validated in the first step to the mortality of each portfolio. A Poisson generalized linear model including interactions with age and calendar year gives a solution to this problem.

Chapter 2

Local regression methods

This chapter is based on Tomas (2012a), Univariate graduation of mortality by local polynomial regression, *Bulletin Français d'Actuariat*, **12**(23), 5-58; and on Tomas (2012b), Essays on boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models, *Insurance and Risk Management*, forthcoming.

2.1 Introduction

This chapter discusses a non-parametric graduation method. We introduce local polynomial regression. We discuss the choice of the smoothing parameters and criteria used for models selection. The statistical properties of the estimators are covered. We graduate the data through the choice of the smoothing parameters. The graduation and corresponding confidence intervals are carried over the entire age range. Tests are used to compare the graduated rates obtained with those obtained by the Whittaker-Henderson smoothing.

The motivation for local regression is that it is easy to understand and to interpret; because of its simplicity it can be tailored to work for many different distributional assumptions; it adapts well to bias problem at boundaries and in regions of high curvature; it does not require smoothness and regularity conditions required by other methods such as boundary kernels; and so on, see Hastie and Loader (1993) for a detailed presentation of the strengths of local regression. Separately, none of these provides a strong reason to favor local regression over other smoothing methods such as smoothing splines, regression splines with knot selection, wavelets, and various modified kernel methods. Rather, it is the combination of these issues that combine to make local regression attractive.

This chapter begins by presenting, in Section 2.2, a general theory of local polynomial regression, showing that this method falls into the class of linear smoothers. The weighting system of the smoothers is discussed in Section 2.3 including the weighting system shapes, the smooth weighted diagram and specific treatments at the boundaries. Section 2.4 develops important properties, including bias and variance, which allow us in Section 2.5 to develop methods for statistical inference, model diagnostics and choice of the smoothing parameters. We emphasize results that have immediate practical consequences. To illustrate the discussion, we present an application based on historical data from the Netherlands in Section 2.6. Section 2.7 provides comparisons with the Whittaker-Henderson framework. Finally, Section 2.8 summarizes the conclusion drawn in this chapter.

The main merits of the material presented in this Chapter are twofold. Firstly, we present extensively local polynomial techniques in view of graduating experience data originating from life insurance. It can be viewed as the prolongation of the kernel estimation for graduation proposed by Gavin *et al.* (1993). It is completed in this chapter with Tomas (2012b) analyzing how the boundaries influence the choice of the smoothing parameters. Secondly, the approach allows relatively easy implementation of the techniques as well as different boundaries corrections in standard statistical software such as R, R Development Core Team (2012).

2.1.1 Premises

> *«If nature were kind enough to make all regression surfaces well approximated by low-order polynomials or other simple parametric functions, there would be no need for the local-fitting methodology. Unfortunately, nature is frequently not so accommodating».*
>
> William S. Cleveland in Cleveland *et al.* (1988, p.88)

The underlying model for local regression is

$$q_i = f(x_i) + u_i, \qquad i = 1, 2, \ldots, n. \tag{2.1}$$

The distribution of the q_i, including the mean, $f(x_i)$, is unknown. However, the u_i are assumed to be independently, identically distributed normal random variables, with zero mean and a finite variance.

In practice we must first model the data, which means making certain assumptions about f and other aspects of the distribution of the q_i. For example, one common distributional assumption is that the q_i have a constant variance. We need to ensure that these assumptions are reflected in the data and, if not, to make appropriate adjustments.

For f, it is supposed that the function can be well approximated locally by a member of a parametric class, frequently taken to be polynomials of a certain degree. We refer to this as parametric localization. Thus, in carrying out local regression we use a parametric family just as in global parametric fitting, but we ask only that the family fit locally and not globally. Parametric localization is the fundamental aspect that distinguishes local regression from other smoothing methods such as smoothing splines, see Silverman (1985); or wavelets, see Donoho and Johnstone (1994); although the notion is implicit in these methods in a variety of ways.

For clarity we distinguish the fitting point with the suffix i to the data points with suffix j. Then, the estimation of f that arises from the above modeling is simple. For each fitting point x_i, we define a neighborhood in the design space of the independent variables. The size λ of the neighborhood is an adjustable parameter that determines how local the fitting is; it is analogous to the length of the moving average in the time series case, and as the neighborhood size increases the estimate becomes smoother.
Within this neighborhood, we assume f is approximated by some member of the chosen parametric family. For example the family might be quadratic polynomials. Then, estimate the parameters from observations in the neighborhood; the local fit at x_i is the fitted function evaluated at x_j. Almost always, we will want to incorporate a weight function, $W(.)$, that gives more weight to the x_j close to x_i and less weight to those that are further.

In short, to use local regression, we must choose the weight function, the bandwidth, the parametric family, and the fitting criterion. The first three choices depend on assumptions we make about the behavior of f. The fourth choice depends on the assumptions we make about other aspects of the distribution of the q_i. In other words, as with parametric fitting, we are modeling the data.

2.1.2 Transforming the data

Before model (2.1) is applied, a key part of any data analysis is to consider transforming the data into a more tractable form that reflects the strengths of the model or that more clearly reveals the structure of the data. In parametric graduation, for example, it may be easier to transform the data and work with a linear model than to graduate the raw rates. The same philosophy applies in non-parametric graduation. If the transformed crude rates broadly follow a straight line, then this may lead to reduced bias over much of the age range, if the data are also evenly spaced. In the following part, we consider transforming the crude rates before graduating and then back-transforming to obtain our estimate of the true rates. The transformation considered satisfies the model,

$$g(q_i) = q_i + r_i \ , \quad \text{for } i = 1, 2, \ldots, n,$$

where the function g denotes the transformation and the residuals r_i are assumed to be independent, identically distributed random variables, with zero mean and a constant, finite variance. Hence the graduation process is carried out on a transformed scale and model (2.1) becomes

$$g(q_i) = \psi(x_i) + \epsilon_i \ , \quad \text{for } i = 1, 2, \ldots, n, \tag{2.2}$$

where the ϵ_i are independent, identically distributed normal random variables with mean 0 and finite variance σ^2. Once it is completed, the transformation is reversed to obtain the graduated rates on the original scale. A commonly used transformation in binary analysis is the logit transformation. For our applications,

$$g(q_i) = \log\left(\frac{q_i}{1-q_i}\right).$$

By smoothing on a logistic scale and then back-transforming, we are guaranteed that the predicted values stay in an appropriate scale, $0 \leq \widehat{q}_i \leq 1$. Gavin *et al.* (1995, p.177-178) provide the motivation that this transformation also reflects the fact that small changes, when the mortality rate is near zero, are as important as larger changes, when the mortality rate is much higher. Note that binary data are often assumed to be independent, but this may not be the case for mortality data due to migration between ages during the period of investigation. This leads to look for smooth relations between neighboring rates by merging information from individuals with similar ages.

Many other transformations are possible (Gompertz, Weibull, $\sin^{-1}(\sqrt{q_i})$ transformation), but their relative merits are beyond the scope of this dissertation. Overall, the choice of transformation remains subjective, and the relative success of a particular transformation seems to depend on the data set. However as Kaas *et al.* (2008, p.232-233) mention it, transformations do not always achieve normality, skewness zero and homoscedasticity at the same time. Moreover an unbiased estimator in the new scale is no longer unbiased when returning to the original scale, which follows from the Jensen's inequality.

For the remaining part, we denote the dependent variable $g(q_i)$ by y_i to ease the notation.

2.2 The local regression estimate

2.2.1 Uni-dimensional case

We assume a model of the form of (2.2),

$$y_i = \psi(x_i) + \epsilon_i \quad , \quad \text{for } i = 1, \ldots, n,$$

where $\psi(x_i)$ is an unknown function and ϵ_i is an error term. The errors ϵ_i are assumed to be independent and identically distributed with mean 0, $\mathbb{E}\left[\epsilon_i\right] = 0$, and have finite variance, $\mathbb{E}\left[\epsilon_i^2\right] = \sigma_i^2 < \infty$.

We now turn to non-parametric estimation of ψ. Globally, no strong assumptions are made about ψ. Locally around a point x_i, we assume that ψ can be well approximated by a member of a simple class of parametric functions. Assume that the function ψ has continuous $(p+1)$st derivative at the point x_i.

For data points x_j in a neighborhood of x_i, we approximate $\psi(x_j)$ via a Taylor expansion by a polynomial of degree p:

$$\begin{aligned}\psi(x_j) &\approx \sum_{p=0}^{P} (\psi^{(p)}(x_i)/p\,!)\,(x_j - x_i)^p \\ &= \psi(x_i) + \psi'(x_i)(x_j - x_i) + \frac{1}{2}\psi''(x_i)(x_j - x_i)^2 \\ &+ \ldots + \frac{1}{p\,!}\psi^{(p)}(x_i)(x_j - x_i)^p \\ &= \sum_{p=0}^{P} \beta_p(x_i)\,(x_j - x_i)^p.\end{aligned} \tag{2.3}$$

We then carry through a weighted polynomial regression:

$$\min \sum_{j=1}^{n} \left(y_j - \sum_{p=0}^{P} \beta_{i,p}(x_j - x_i)^p \right)^2 W\left(\frac{x_j - x_i}{h}\right), \tag{2.4}$$

where $W(.)$ denotes a non-negative weight function depending on the target value x_i and the measurement points x_j, and in addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the sizes of the neighborhood of x_i.

If $\{\widehat{\beta}_p(x_i)\}$ denotes the solution to the above weighted least squares problem (2.4), then it is clear from approximation (2.3) that $p\,!\,\widehat{\beta}_p(x_i)$ estimates $\psi^{(p)}(x_i)$, $p = 0, 1, \ldots, P$. The weighted sum of squares can be written in matrix form as

$$(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X\beta}),$$

with

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 - x_i & (x_1 - x_i)^2 & \ldots & (x_1 - x_i)^P \\ 1 & x_2 - x_i & (x_2 - x_i)^2 & \ldots & (x_2 - x_i)^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_i & (x_n - x_i)^2 & \ldots & (x_n - x_i)^P \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

and $\boldsymbol{W}$ a diagonal matrix with entries $\{w_j\}_{j=1}^n$, such that

$$w_j = \begin{cases} W(|x_j - x_i|/h) & \text{if } |x_j - x_i|/h \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

If $\boldsymbol{WX}$ has full column rank, least squares theory gives the explicit expression for the minimizer

$$\widehat{\boldsymbol{\beta}}(x_i) = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\ \boldsymbol{y}, \tag{2.5}$$

and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_P})$. Hence,

$$\widehat{\beta_0}(x_i) = \widehat{\psi}(x_i) = \boldsymbol{e}_1^T \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\ \boldsymbol{y}. \tag{2.6}$$

Here and throughout, we let $\boldsymbol{e}_v$ denote a column vector of length $P + 1$ having 1 as its vth entry and all other entries equal to zero.

It is important to note that, contrary to ordinary parametric least squares, this estimator varies with x_i, as locally around the target value a polynomial of degree P is fitted by using the familiar technique of least-squares fitting. Thus, local regression is conceptually quite simple. In order to get an estimate for the function $\psi(x_i)$, one has to minimize (2.4) for a grid of target values x_i. For each target value one gets specific parameter estimates $\boldsymbol{\beta}(x_i)$.

2.2.2 Two-dimensional case

Extending the the theory of local regression to multiple predictors is straightforward. We would require a multivariate weight function and multivariate local polynomials. This idea was first considered by Shepard (1968) who realized that a surface based on a weighted average of the values of the data points, where the weighting was a function of the distances to those points, satisfied the problem. However, the interpolation method described in Shepard's article used the weights to determine the height directly. McLain (1974) and later Stone (1982) used a weighting technique with weights depending on the distances of the data points where the weights were used with a least squares fit to find coefficients of a quadratic polynomial to act as an approximation to the surface. Statistical methodology and visualization for multivariate fitting has been developed by Cleveland and Devlin (1988) and the associated *lowess* procedure.

With two predictor variables, the local regression model becomes

$$y_i = \psi(x_{i,1}, x_{i,2}) + \epsilon_i,$$

where $\psi(\cdot; \cdot)$ is an unknown function. Again, the smooth function ψ can be approximated in a neighborhood of a point $x_i = (x_{i,1}, x_{i,2})$ by a local polynomial of degree p.

If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares and their cross-products. For example, a local quadratic approximation is:

$$\psi(x_j) = \psi(x_{j,1}, x_{j,2}) \approx \beta_0(x_i) + \beta_1(x_i)(x_{j,1} - x_{i,1}) + \beta_2(x_i)(x_{j,2} - x_{i,2})$$
$$+\frac{1}{2}\beta_3(x_i)(x_{j,1}-x_{i,1})^2+\beta_4(x_i)(x_{j,1}-x_{i,1})(x_{j,2}-x_{i,2})+\frac{1}{2}\beta_5(x_i)(x_{j,2}-x_{i,2})^2.$$

The weights are defined on the multivariate space. First, we define a distance measure $\rho(x_i, x_j)$ between the observations $x_j = (x_{j,1}, x_{j,2})$ and the fitting point $x_i = (x_{i,1}, x_{i,2})$.

A common choice is the Euclidean distance,

$$\rho(x_i, x_j) = \sqrt{(x_{j,1} - x_{i,1})^2 + (x_{j,2} - x_{i,2})^2}.$$

Secondly, a spherical weight function gives to the observation $x_j = (x_{j,1}, x_{j,2})$ the weight

$$w_j = \begin{cases} W(|\rho(x_i, x_j)| / h) & \text{if } |\rho(x_i, x_j)| / h \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the spherical weight function could be asymmetric, allowing more smoothing in one direction than in another. As in the univariate case, the local coefficients $\beta_p(x_i)$ are estimated by solving the weighted least squares problem (2.4). Following (2.6), the local polynomial estimate is then $\widehat{\beta_0}(x_i) = \widehat{\psi}(x_i)$.

2.3 The weighting system

2.3.1 The weighting system shape

The weighting system of local regression depends on the constellation of smoothing parameters formed by the weight function, the bandwidth and the degree of the polynomial. In addition, it depends on the variance function and on the link function in case of local likelihood models, see Chapter 3. These choices depend on assumptions we make about the behavior of the true curve.

It is well known that between the three smoothing parameters, the weight function has much less influence on the bias and variance trade-off. The choice is not too crucial, at best it changes the visual quality of the regression curve.

We consider a weight function $W(u)$ that has the properties

i. $W(u) > 0$ for $|u| < 1$;

ii. $W(-u) = W(u)$;

iii. $W(u)$ is a non increasing function for $u \geq 0$;

The requirements for $W(u)$ described above are needed for the following reasons: (i) is necessary, of course, since negative weights do not make sense; (ii) is required since there is no reason to treat points to the left of x_i differently from those to the right; (iii) is required for it seems unreasonable to allow a particular point to have less weight than one that is further from x_i. So $W(u)$ is a weight function like those given in Table 2.1.

Weight function	$W(u)$
Uniform or Rectangular	$\frac{1}{2}I(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)I(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)I(\lvert u\rvert \leq 1)$
Quartic (Biweight)	$\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \leq 1)$
Triweight	$\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \leq 1)$
Tricube	$(1-\lvert u^3\rvert)^3 I(\lvert u\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(\frac{1}{2}u^2)$

Table 2.1: *Example of weight functions with $u = |x_j - x_i|/h$.*

Figure 2.1 displays some of the weight functions presented above.

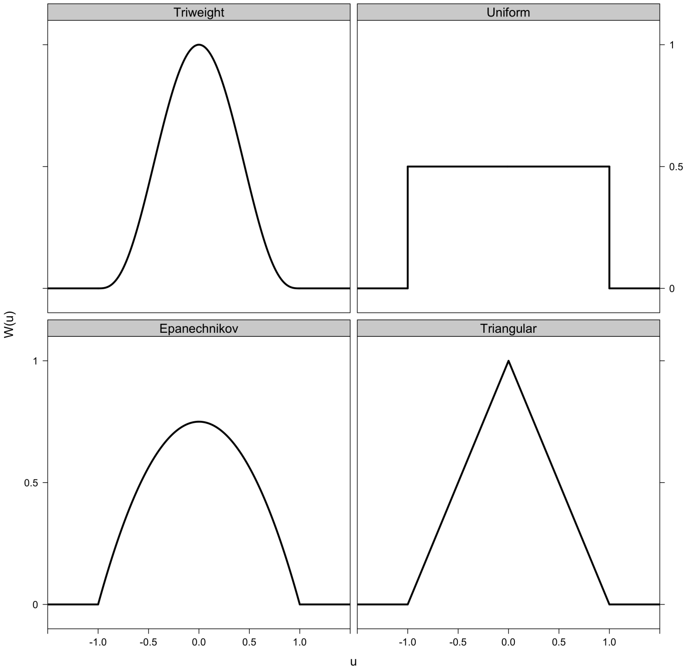
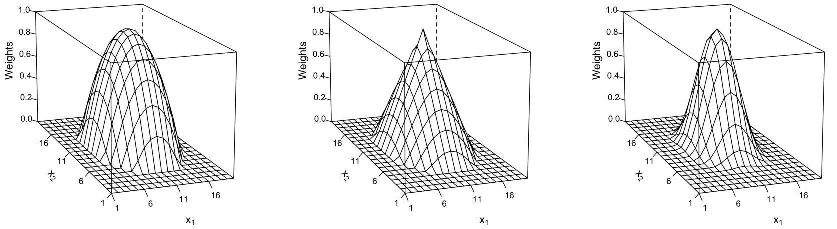


Figure 2.1: *Weighting system shape of some weight functions.*

For a weight function $W(u)$, the weights decrease with increasing distance $|x_j - x_i|$. The window-width or bandwidth λ determines how fast the weights decrease. For small λ, only values in the immediate neighborhood of x_i will be influential; for large λ, values more distant from x_i may also influence the estimate. One alternative is a rectangular weight function, or uniform. With uniform weights, all observations within the window width receive weight $1/2$, those further away receive weight 0, and observations abruptly switch in and out of the smoothing window.

In two dimensions, the weights are defined on the multivariate space. Figure 2.2 shows some of the weight functions displayed above.



(a) *Epanechnikov* **(b)** *Triangular* **(c)** *Triweight*

Figure 2.2: *Weighting system shape in two dimensions with a radius* $h = 7$.

2.3.2 The smooth weight diagram

The form of the local regression estimate is simple in that it is linear in y_i. Because local polynomial regressions solve a least squares problem, $\widehat{\psi}(x_i)$ is a linear estimate. That is, for each x_i there exists some smoothing weights $s_1(x_i), s_2(x_i), \ldots, s_n(x_i)$ such that

$$\widehat{\psi}(x_i) = \sum_{j=1}^{n} s_j(x_i) y_j, \tag{2.7}$$

where the smoothing weights on the observed responses are given by

$$s_j(x_i) = w_j \sum_{p=0}^{P} \beta_p (x_j - x_i)^p. \tag{2.8}$$

This is equivalent to the theorem originally from Henderson (1916) for local cubic fitting and reformulated by Loader (1999b), which provides a characterization of the smoother matrix for local polynomial regression: the smoother matrix for a local polynomial fit of degree P has the form of least squares weights multiplied by a polynomials of degree P. This representation is unique, provided $\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ is non-singular.

Figure 2.3 presents the smooth weights, $\boldsymbol{s}(x_i)$, according to the order of polynomial, for the four weighting system shapes drawn in Figure 2.1.

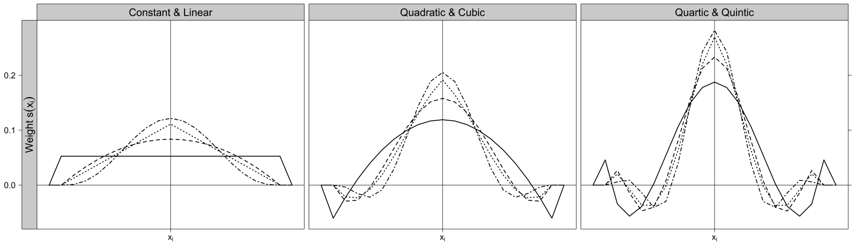


Figure 2.3: *Smooth weights $\boldsymbol{s}(x_i)$, for observation i in the central region, computed with $\lambda = 19$ for rectangular (solid line), triangular (dotted line), epanechnikov (dashed line) and triweight (dotdashed line) weight functions.*

It is obvious that the triweight weight function has the smallest dispersion around the target point x_i while the rectangular weight function implies more smoothing. Note that the fit to a polynomial of even degree gives the same result as that of the next odd degree for values at the central region, see Section 2.4.3 It has also been discussed by Fan and Gijbels (1995a, p.215-218) and Ruppert and Wand (1994).

As we can see in (2.8) the smoother weights $s_j(x_i)$ depend on λ and $\boldsymbol{X}$ in a highly non-linear way. The only linearity we have in equation (2.7) is linearity in $\boldsymbol{y}$. This linear representation (2.7) provides a basis for the theoretical development of local regression estimation. Likewise in a matrix form,

$$\begin{bmatrix} \widehat{\psi}(x_1) \\ \widehat{\psi}(x_2) \\ \vdots \\ \widehat{\psi}(x_n) \end{bmatrix} = \boldsymbol{S} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

where $\boldsymbol{S}$ is the smooth weight diagram, an $n \times n$ matrix

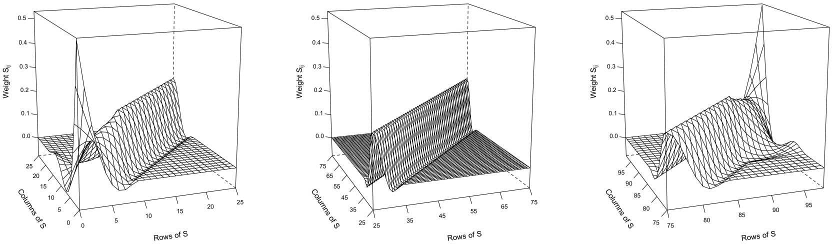
$$\boldsymbol{S} = \begin{bmatrix} s_1(x_1) & s_2(x_1) & \dots & s_n(x_1) \\ s_1(x_2) & s_2(x_2) & \dots & s_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ s_1(x_n) & s_2(x_n) & \dots & s_n(x_n) \end{bmatrix},$$

with rows

$$\boldsymbol{s}(x_i)^T = (s_1(x_i), s_2(x_i), \dots, s_n(x_i)) = \boldsymbol{e}_1^T \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}. \qquad (2.9)$$

The above distributional results are the same as those for parametric least-squares except that for least-squares $\boldsymbol{S}$ is replaced by the so called *hat matrix*, the projection operator onto the space spanned by the fitting variables. $\boldsymbol{S}$ shares with the hat matrix the property that if $\boldsymbol{z}$ is a vector in this space then $\boldsymbol{S}\boldsymbol{z} = \boldsymbol{z}$. In other words, the smooth weight diagram is constant preserving, the rows of $\boldsymbol{S}$ sum to one. The result of this partial analogy with parametric least-squares is that, in a few aspects, distributional results for local regression are the same as those for least-squares and, in most other aspects, statistical quantities for local regression that are defined in analogy with least-squares have distributions that are well-approximated by those for least-squares, as argued in Cleveland *et al.* (1988). This is good news because it means that familiar techniques can be used to make inferences based on the local-regression estimates.

Figure 2.4 provides an illustration of the smooth weight diagram $\boldsymbol{S}$. The weight function associated with the i-th point is used to compute the weights in the i-th row, $\boldsymbol{s}(x_i)$. $\boldsymbol{S}$ in Figure 2.4 has been computed with $\lambda = 19$, a polynomial of degree 3 and a triweight weight function with boundary correction *type 1*, see Section 2.3.4.



(a) $i, j = 0, \ldots, 25$ **(b)** $i, j = 25, \ldots, 75$ **(c)** $i, j = 75, \ldots, 98$

Figure 2.4: *Smoother S_{ij} computed with $\lambda = 19$, a polynomial of degree 3, a triweight weight function and boundary correction type 1.*

The weights are shown as the height along the i-th row of the surface. For values in the central region, the weights form a triweight kernel such as Figure 2.3, center panel. But as the point at which we are estimating the true curve moves towards the boundaries, the kernel overlaps the boundary and becomes asymmetric. Also some weights are negative. Moreover, the height of the kernel increases because fewer observations are available.

By fitting local polynomials models to series originating from life insurance, we observe a relatively high curvature in the boundaries. In consequence, the selection of the constellation of the smoothing parameters may be mainly driven by minimizing the criteria in the boundaries rather than for the whole set of data points. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table.

2.3.3 Effective dimension fo a linear smoother

The effective dimension of the fitted model is an important concept in modeling. For linear models, this concept is clear and intuitive. The number of parameters used in the model determines its dimension. In non-parametric settings, a different definition is needed.

In linear models, the hat matrix, $\boldsymbol{H}$, is idempotent, $\mathrm{tr}(\boldsymbol{H}\ \boldsymbol{H}^T) = \mathrm{tr}(\boldsymbol{H}) = \mathrm{rank}(\boldsymbol{H})$. Hence the trace of the hat matrix is equal to the number of parameters in the fitted model. Given this feature of classic linear models, the trace of the hat matrix can be used to assess the fitted degrees of freedom and hence the effective dimension of a smoother.

The *influence* or leverage values, denoted $\mathrm{infl}(x_i)$, are the diagonal elements $s_i(x_i)$ of the smooth weight diagram. They measure the sensitivity of the fitted curve to the individual data points. For local polynomial regression, we define the influence function at x_i by

$$\mathrm{infl}(x_i) = \boldsymbol{e}_1^T \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{e}_1.$$

The property of influence relates to the fact that as $\mathrm{infl}(x_i)$ approaches one, the corresponding residual approaches zero.

Hence, in analogy with parametric least-squares, we define

$$\begin{aligned} \upsilon_1 &= \sum_{i=1}^{n} \mathrm{infl}(x_i) = \mathrm{tr}(\boldsymbol{S}) \\ \upsilon_2 &= \sum_{i=1}^{n} \|\boldsymbol{s}(x_i)\|^2 = \mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^T). \end{aligned} \tag{2.10}$$

υ_1 and υ_2 are the fitted degrees of freedom (DF) of $\widehat{\psi}(x_i)$. For locally-weighted regression, as the bandwidth increases or as the degree of polynomial reduces, υ_2 tends to decrease, so we are using more equivalent degrees of freedom to explain the data. The fitted DF provide a mechanism by which different smoothers, with different smoothing parameters, can be compared. More extensive discussion of the degrees of freedom of a smoother can be found in Cleveland and Devlin (1988).

2.3.4 Specific treatments at the boundaries

To understand the boundary problem in the context of graduation, we study three specific treatments including symmetric and asymmetric weight systems.

i. *Type 1* uses an asymmetric weighting system. It always uses λ observations whatever the target point is. It means, for instance, that for a target point at the left boundary it uses all the observations κ

available at the left side, and $\lambda - 1 - \kappa$ at the right side. Reciprocally for the right boundary. This type of correction is found in most smoothing software such as the `loess()` or `locfit()` functions in R, R Development Core Team (2012).

ii. *Type 2* uses a different asymmetric weighting system. For instance at the left boundary, it uses all observations available at the left side, and $(\lambda - 1)/2$ observations at the right side. Reciprocally at the right boundary.

iii. *Type 3* is a combination of observed rates and an adaptive symmetric weighting system. This correction is only applied to the left boundary. From age 0 to $\nu_{p,W}$ the mortality rates equal the observed ones. $\nu_{p,W}$ depends on the polynomial degree p and on the weight function $W(.)$ to ensure sufficiently observations to fit a polynomial of degree p. Then from $\nu_{p,W} + 1$ to $(\lambda - 1)/2$, we use an adaptive symmetric window width with $2 \times (x_i - 1) + 1$ observations, where x_i is the target point. This correction is based on an idea presented by the Dutch Actuarieel Genootschap (the Dutch Actuarial Society), see Donselaar *et al.* (2007).

We apply these corrections to the smoothers of degree 0 to 4 with four weighting system shapes. Figure 2.5 shows the symmetric and asymmetric weighting system $\boldsymbol{s}(x_i)$ for $i = 5$ (left boundary) of the corrections mentioned above with $\lambda = 19$. It is apparent that the symmetric weights of correction *type 3* have the smallest dispersion around the central value while correction *type 1* implies more smoothing.

As we face a fixed design model, in which we have a single observed mortality rate at equally spaced ages, the amount of smoothing applied by the local polynomial regression is identical at the left and right boundary. Hence the amount of smoothing is lower at the left boundary than to the right as the number of exposures is larger. Table 2.2 presents the fitted DF for smoother $\boldsymbol{S}$ in the left boundary, that is for observations x_i for $i = 1, \ldots, 10$. The window width, λ, is fixed to 19 observations.

The fitted DF aid interpretation in providing a measure of the amount of smoothing applied. For instance, 1 DF represents a smooth model with very little flexibility while 7 DF represents a noisy model showing many features. It is obvious that the amount of smoothing decreases when increasing the degree of polynomial. In addition we observe that the amount of smoothing applied is higher when the weight function has a high dispersion around the central value. A rectangular weighting system shape implies very little flexibility, but a triweight weighting shape shows more features.

Note again that a least-squares fit to a polynomial of even degree gives the same result as that of the next odd degree for a symmetric weight function.

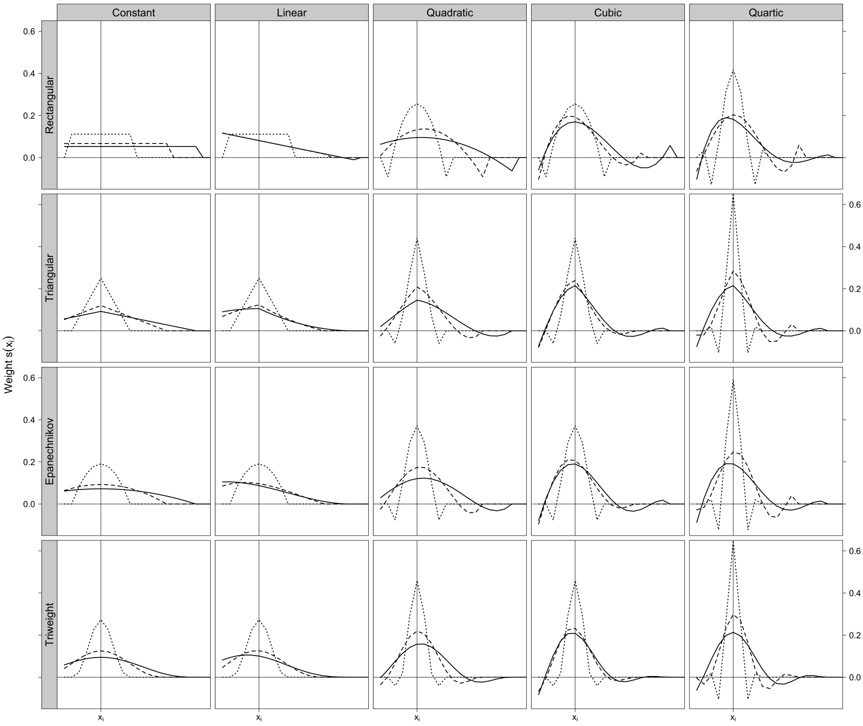


Figure 2.5: *Smooth weights $\boldsymbol{s}(x_i)$ for $i = 5$ (left boundary) with $\lambda = 19$ for correction type 1 (solid line), type 2 (dashed line) and type 3 (dotted line).*

It is apparent that boundary correction *type 1* induces more smoothing in the boundaries than *type 2* and *type 3*. Correction *type 3*, having smooth weights showing the smallest dispersion, has the property of showing more features.

2.3.5 Comparison with the Whittaker-Henderson model

It is interesting to compare the local polynomials approach with classical graduation methods. Among the classical methods we can mention the splines approach or the Whittaker-Henderson smoothing. As shown by Taylor (1992) and Planchet and Winter (2007) both approaches lead to very similar results. Taylor (1992, p.15) shows that natural spline graduation can be regarded as approximately Whittaker-Henderson graduation with statistically insignificant terms removed, concluding that the general spline function is preferable to Whittaker-Henderson graduation due to its greater flexibility. In this section we choose to use the Whittaker-Henderson model which is simpler to implement.

		Local Polynomial Regression				
	Weight fct.	$p=0$	$p=1$	$p=2$	$p=3$	$p=4$
Corr. *type 1*	Rectangular	0.40	1.03	1.47	2.04	2.51
	Triangular	0.65	1.17	1.73	2.24	2.80
	Epanechnikov	0.56	1.11	1.65	2.17	2.72
	Triweight	0.78	1.27	1.91	2.39	2.99
Corr. *type 2*	Rectangular	0.66	1.33	1.94	2.61	3.21
	Triangular	1.03	1.95	2.87	3.67	4.37
	Epanechnikov	0.92	1.81	2.72	3.55	4.25
	Triweight	1.25	2.23	3.15	3.93	4.59
Corr. *type 3*	Rectangular	3.11		4.77		6.08
	Triangular	4.22		5.75		6.94
	Epanechnikov	4.11		5.66		6.88
	Triweight	4.40		5.90		7.02

Table 2.2: *Fitted DF for local polynomial regression in the left boundary*

In the following, we show that the Whittaker-Henderson model falls into the class of linear smoothers. It will allow us to use the methodology developed in Section 2.3.3 for model comparisons and smoothing power.

The Whittaker-Henderson model is non-parametric and is a relatively simple and natural version of Bayesian smoothing, see Taylor (1992). The method relies on the combination of a fit and smoothness measure. The chosen parameters minimize a linear combination of these two criteria,

$$M = F + h \times S,$$

where F and S denote the fit and smoothness measures respectively and h is a parameter allowing more emphasis on the smoothness criterion. The fit and smoothness measures are

$$F = \sum_{i=1}^{n} v_i (y_i - \widehat{y_i})^2 \quad \text{and} \qquad S = \sum_{i=1}^{n-z} (\Delta^z y_i)^2,$$

where v_i represents the weight for observation i, taken generally as the ratio $l_i / \max(l_i)$ where l_i denotes the exposure, and z is another parameter representing the polynomial degree.

For this optimization problem, we solve the n equations given by the partial derivatives of M with respect to each of the y_i,

$$\frac{\partial M}{\partial y_i} = 0, \quad i = 1, \ldots, n.$$

With $\boldsymbol{y} = (y_i)_{1\leq i\leq n}$, $\widehat{\boldsymbol{y}} = (\widehat{y}_i)_{1\leq i\leq n}$ and $\boldsymbol{V} = \text{diag}(v_i)_{1\leq i\leq n}$, F can be written in matrix notation as

$$F = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T \boldsymbol{V} (\boldsymbol{y} - \widehat{\boldsymbol{y}}).$$

For the smoothness criterion, writing $\Delta^z \boldsymbol{y} = (\Delta^z y_i)_{1\leq i\leq n-z}$, leads to $S = (\Delta^z \boldsymbol{y})^T \Delta^z \boldsymbol{y}$.
To find $\Delta^z \boldsymbol{y}$, we introduce a matrix denoted K_z, of dimension $(n-z)\times z$, where the terms are binomial coefficients of order z and where the signs of the coefficients alternate and start being positive for z even, $\Delta^z \boldsymbol{y} = K_z * \boldsymbol{y}$. The M criterion can finally be written as

$$\begin{aligned} M &= (\boldsymbol{y} - \widehat{\boldsymbol{y}})^T \boldsymbol{V} (\boldsymbol{y} - \widehat{\boldsymbol{y}}) + h\boldsymbol{y}^T K_z^T K_z \boldsymbol{y} \\ &= \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{V} \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}}^T \boldsymbol{V} \widehat{\boldsymbol{y}} + h\boldsymbol{y}^T K_z^T K_z \boldsymbol{y}. \end{aligned}$$

It leads to $\frac{\partial M}{\partial \boldsymbol{y}} = 2\boldsymbol{V}\boldsymbol{y} - 2\boldsymbol{V}\widehat{\boldsymbol{y}} + 2hK_z^T K_z \boldsymbol{y}$. Solving $\partial M / \partial \boldsymbol{y} = 0$ leads to the expression:

$$\widehat{\boldsymbol{y}} = (\boldsymbol{V} + hK_z^T K_z)^{-1} \boldsymbol{V}\boldsymbol{y}. \tag{2.11}$$

We see that the form of the estimate is linear in the y_i.

Moreover, the smooth weights depend on the sample size $\sum_{i=1}^n l_i$. Hence, the amount of smoothing applied is no longer identical at the left and right boundaries. It is lower in the left boundary than to the right as the exposure is larger.

For ease of comparison with the Whittaker-Henderson model smoother, h is fixed to 5 in expression (2.11) leading to approximately 19 observations participating non-negligibly in the estimation, having weights higher than 0.01. The fitted DF in the left boundary can be found in Table 2.3.

Whittaker-Henderson smoother				
$z=0$	$z=1$	$z=2$	$z=3$	$z=4$
0.17	1.23	2.17	2.79	3.26

Table 2.3: *Fitted DF in the left boundary for the Whittaker-Henderson model, with* $h = 5$

The amount of smoothing in the left boundary implied by the Whittaker-Henderson model lies between corrections *type 1* and *type 2*, see Table 2.2. Hence the model can be slightly more flexible at the left boundary than when correction *type 1* is applied.

Figure 2.6 displays the influence values of the smoothers implied by the local polynomial regression and Whittaker-Henderson models.

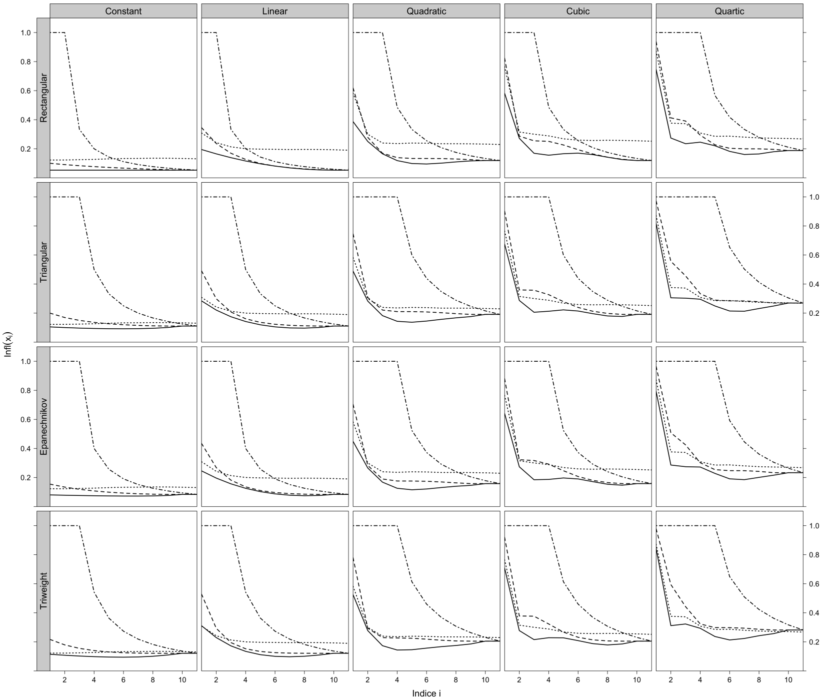


Figure 2.6: *Influence values in the left boundary for correction type 1 (solid black line), type 2 (dashed line), type 3 (dotted-dashed line) and Whittaker-Henderson model (dotted line).*

For correction *type 3*, from x_1 to $\nu_{p,W}$, the smoothed mortality rates equal the observed ones. In consequence, the corresponding influence values equal 1. The parameter $\nu_{p,W}$ depends on the degree of polynomial and on the weighting system to ensure that a sufficient number of observations is used to fit the corresponding polynomial.

Under a rectangular weighting system, corrections *type 1* and *type 2* give similar results. Then, by using a weighting system shape inducing less dispersion around the central value, the differences become more apparent. The shape of the influence functions drawn by a triangular, Epanechnikov or triweight weight function is relatively similar. Note that the influence values of the Whittaker-Henderson model lie mostly between corrections *type 1* and *type 2*.

By a constant fit, the influence values for corrections *type 1*, *type 2* and the Whittaker-Henderson model are approximatively equal to 0.1, indicating that y_i constitutes about 10 % of the fitted value. But the main feature is the boundary effect where the influence function shows a huge increase. This reflects the difficulty of fitting a polynomial at boundary regions. Note also that the effect is more pronounced as we increase the order of polynomial. This shows that boundaries are a main concern when choosing the order of approximation and, more generally, the constellation of smoothing parameters.

In the next section, we turn to the statistical properties of this smoother. As we will see, smoothing always means a compromise between bias and variance and the choice of the smoothing parameters will be driven by this trade-off.

2.4 Statistical properties

2.4.1 Assessment of bias and variance

Contrary to linear model fitting, there is no exact expression for the variance in a general case, because local polynomial regression models involve a non-linear (vector) function of the estimate $k(\widehat{\boldsymbol{\beta}})$. On the other hand, we can approximate the non-linear function using a first-order Taylor series expansion about $\boldsymbol{\beta}$. Assuming first order differentiability of $k(.)$, we have

$$k(\widehat{\boldsymbol{\beta}}) = k(\boldsymbol{\beta}) + \frac{\partial k(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + o\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|.$$

Then for $\widehat{\psi}(x_i) = \widehat{\beta}_0(x_i)$, see equation (2.6), we obtain

$$\widehat{\psi}(x_i) = \psi(x_i) + \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + o\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|,$$

and,

$$\mathbb{E}\left[\widehat{\psi}(x_i)\right] = \psi(x_i) + \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T}\mathbb{E}\left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right].$$

We obtain an approximation of the variance of the local polynomials estimate by

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] &\approx \mathbb{E}\left[\left(\widehat{\psi}(x_i) - \psi(x_i)\right)^2\right] \\ &\approx \mathbb{E}\left[\left(\frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^2\right] \\ &= \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T}\mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)^2\right]\frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0}. \end{aligned} \tag{2.12}$$

We still need to estimate $\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}(x_i)\right] = \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)^2\right]$. However, standard weighted least squares theory provides explicit mean and variance expressions of the solution (2.5),

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}(x_i)\right] = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{f} = \boldsymbol{\beta} + \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}, \tag{2.13}$$

where $\boldsymbol{f} = (\psi(x_1), \psi(x_2), \ldots, \psi(x_n))^T$ and $\boldsymbol{\epsilon} = \left\{\epsilon_j\right\}_{j=1}^n = \boldsymbol{f} - \boldsymbol{X\beta}$; and,

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}(x_i)\right] &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)^2\right] = \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)^T\right] \\ &= \mathbb{E}\left[\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{W}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\right] \\ &= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\boldsymbol{W}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}. \end{aligned} \tag{2.14}$$

From (2.2), $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \sigma^2(x_j)\boldsymbol{I}_n$. Using local homoscedasticity, namely that $\sigma(x_j) \approx \sigma(x_i)$ for x_j in a neighborhood of x_i, equation (2.14) can be approximated by

$$\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}(x_i)\right] = \sigma^2(x_i)\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}. \tag{2.15}$$

Therefore,

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] &= \sigma^2(x_i)\boldsymbol{e}_1^T\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{e}_1 \qquad (2.16)\\ &= \sigma^2(x_i)\boldsymbol{S}\,\boldsymbol{S}^T,\end{aligned}$$

since $\partial k(\boldsymbol{\beta})/\partial\beta_0^T = \boldsymbol{e}_1^T$. Then by (2.9) we obtain compact forms for the mean and variance of the local regression estimate, similar to Loader (1999b, p.288)

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}(x_i)\right] &= \sum_{j=1}^n s_j(x_i)\,\psi(x_j) \\ \mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] &= \sigma^2(x_i)\sum_{j=1}^n s_j^2(x_i) = \sigma^2(x_i)\|\boldsymbol{s}(x_i)\|^2. \end{aligned} \tag{2.17}$$

The variance reducing factor $\|\boldsymbol{s}(x_i)\|^2$ measures the reduction in variance due to the local regression. It usually decreases with the bandwidth.

The bias and variance in equations (2.13) and (2.14) are not directly accessible, as they depend on unknown quantities, the residual $\boldsymbol{\epsilon}$ and $\sigma^2(x_i)$. Finite sample estimates are needed to gain access to a smoothing parameter selection procedure and construction of pointwise confidence intervals. We now provide an estimate for the bias and variance of the local polynomial fit based on an idea introduced by Fan and Gijbels (1995a, p.218-219) and Fan and Gijbels (1995b, p.376-378).

The bias of the estimator $\widehat{\boldsymbol{\beta}}$ comes from the approximation error in the Taylor expansion. Recall the bias vector given in (2.13) and let

$$\epsilon(x_j) = \psi(x_j) - \sum_{p=0}^{P} \psi^{(p)}(x_i)(x_j - x_i)^p / p\,!$$

denote this approximation error at the point x_j. Assume that the $(p{+}a{+}1)$th derivative of the function ψ exists at the point x_i for some $a > 0$. Then, a further expansion of $\psi(x_j)$ gives an approximation to the approximation error

$$\epsilon(x_j) \approx \beta_{p+1}(x_j - x_i)^{p+1} + \ldots + \beta_{p+a}(x_j - x_i)^{p+a} \equiv \tau_j, \tag{2.18}$$

where $\beta_k = \psi^{(k)}(x_i)/k!$ and a denotes the order of the approximation. The choice of a has an effect on the performance of the estimated bias. A discussion of the choice of a can be found in Fan and Gijbels (1995b, p.376), who recommend using $a = 2$ for practical implementation.
The unknown parameters in $\tau = (\tau_1, \tau_2, \ldots, \tau_n)^T$ can be estimated from a local polynomial fit of order $p + a$ with a bandwidth h^*. Let $\widehat{\beta}^*_{p+1}, \ldots, \widehat{\beta}^*_{p+a}$ be the resulting estimated regression coefficients and denote the weighted residual sum of squares by

$$\widehat{\sigma}^{*2}(x_i) = \frac{1}{\mathrm{tr}(\boldsymbol{W}^*) - \mathrm{tr}\big((\boldsymbol{X}^{*T}\boldsymbol{W}^*\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{W}^{*2}\boldsymbol{X}^*\big)} \sum_{j=1}^{n} (y_j - \widehat{y}_j)^2 W\left(\frac{x_j - x_i}{h^*}\right), \tag{2.19}$$

where the $\widehat{y}_j$ are the fitted values from the $(p + a)$th order local polynomial fit. Moreover, $\boldsymbol{X}^*$ and $\boldsymbol{W}^*$, similar to $\boldsymbol{X}$ and $\boldsymbol{W}$, denote respectively the design matrix and weight matrix for the local $(p + a)$th order polynomial fit with bandwidth h^*. Substitution of the estimates for $\beta_{p+1}, \ldots, \beta_{p+a}$ into the vector τ gives $\widehat{\tau}$, leads to an estimated bias vector

$$\begin{aligned} \widehat{\mathrm{bias}}_p(x_i) &= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\widehat{\tau} && (2.20) \\ &= (\boldsymbol{T})^{-1} \begin{pmatrix} \widehat{\beta}^*_{p+1} t_{p+1} & \cdots & \widehat{\beta}^*_{p+a} t_{p+a} \\ & \vdots & \\ \widehat{\beta}^*_{p+1} t_{2p+1} & \cdots & \widehat{\beta}^*_{p+a} t_{2p+a} \end{pmatrix}, && (2.21) \end{aligned}$$

where $\boldsymbol{T} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix of which the (j,k) element is t_{j+k-2} with

$$t_k = \sum_{j=1}^{n} (x_j - x_i)^k W\left(\frac{x_j - x_i}{h}\right). \tag{2.22}$$

The variance matrix of the estimator (2.14) can be estimated by substituting $\widehat{\sigma}^{*2}(x_i)$, defined in (2.19), into (2.15). This provides an estimated variance matrix

$$\widehat{\text{var}}_p(x_i) = \widehat{\sigma}^{*2}(x_i)\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}. \tag{2.23}$$

Expressions (2.21) and (2.23) give the estimated bias and variance not only for $\widehat{\psi}(x_i)$ but also for $\widehat{\psi}^{(v)}(x_i) = v!\,\beta_v(x_i), v = 0, 1, \dots, P$.
The estimated bias for $\widehat{\psi}^{(v)}(x_i)$ is the $(v+1)$th element of (2.20), denoted by $\widehat{\text{bias}}_{p,v}(x_i)$, multiplied by $v!$. Its estimated variance is given by $(v+1)$th diagonal element of (2.23), denoted by $\widehat{\text{var}}_{p,v}(x_i)$, times $(v!)^2$. For instance,

$$\mathbb{E}\left[\widehat{\psi}(x_i)\right] - \psi(x_i) = \boldsymbol{e}_1^T\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{W}\widehat{\tau}, \tag{2.24}$$

and,

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] &= \widehat{\sigma}^{*2}(x_i)\,\boldsymbol{e}_1^T\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\boldsymbol{e}_1\\ &= \widehat{\sigma}^{*2}(x_i)\,\|\boldsymbol{s}(x_i)\|^2. \end{aligned} \tag{2.25}$$

Recall that the approximated bias (2.13) and variance (2.14) depend on the quantities $\epsilon_1, \dots, \epsilon_n$ and $\sigma^2(x_i)$ respectively, which are unknown. These quantities will be estimated by fitting a local polynomial of degree $p + a$ locally via equation (2.4), using a pilot bandwidth h^*. This gives estimates $\widehat{\beta}_0^*, \widehat{\beta}_1^*, \dots, \widehat{\beta}_{p+a}^*$ and $\widehat{\sigma}^{*2}(x_i)$, which are then substituted respectively into expressions (2.18), yielding estimates $\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_n$ of $\tau_1, \tau_2, \dots, \tau_n$, and (2.23) leading to the estimated variance. Finally, the estimated bias is computed by substituting the estimates $\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_n$ into (2.20).

As recommended by Fan and Gijbels (1995b, p.377) we modify the bias estimate in expression (2.21) to improve its finite sample performance, especially in case of higher order fits ($p \geq 2$). This slight modification consists of replacing the higher order terms $t_{p+a+1}, t_{p+a+2}, \dots, t_{2p+a}$ in (2.21) by 0. Fan and Gijbels (1995b, p.377) argue that it reduces collinearity effects among monomials $\{(x_j - x_i)^k\}$ such as $\{(x_j - x_i)^2\}$ and $\{(x_j - x_i)^4\}$. This operation has no effect on the asymptotic properties, since it only concerns the higher order terms and no leading terms.

Other authors have expressed the bias and the variance in other fashions, see Section 2.5.2 or Cleveland *et al.* (1988, p.100), however we do not provide here any comparisons between the approaches.

2.4.2 Construction of pointwise confidence intervals

Having estimates of the bias and variance, we are now able to compute pointwise confidence intervals for $\widehat{\psi}(x_i)$.

By (2.25) a local polynomial estimate $\widehat{\psi}(x_i)$ has the distribution

$$\frac{\widehat{\psi}(x_i) - \mathbb{E}\left[\psi(x_i)\right]}{\sigma(x_i)\|\boldsymbol{s}(x_i)\|} \sim N(0,1).$$

If $\widehat{b}(x_i) = \mathbb{E}\left[\widehat{\psi}(x_i)\right] - \psi(x_i)$, an estimated bias corrected confidence interval is

$$\widehat{I}(x_i) = \\ \left(\widehat{\psi}(x_i) - \widehat{b}(x_i) - c\,\widehat{\sigma}^*(x_i)\|\boldsymbol{s}(x_i)\|,\ \widehat{\psi}(x_i) - \widehat{b}(x_i) + c\,\widehat{\sigma}^*(x_i)\|\boldsymbol{s}(x_i)\|\right),$$

where c is the appropriate quantile of the standard normal distribution ($c = 1.96$ for 95 % confidence) and $\widehat{b}(x_i)$ is a bias estimate as defined in (2.24).

This approach based on a plug-in principle has been criticized in the literature. Loader (1999b, p.168) argue that plug-in bias estimates simply amount to increasing the order of the fit. For example, a double smoothing bias correction converts a local constant estimate into a local quadratic. In this case an estimated $\widehat{I}(x_i)$ is just a construction of an under-smoothed interval centered around the local quadratic estimate $\widehat{\psi}(x_i) - \widehat{b}(x_i)$.

2.4.3 A bias and variance trade-off

The bias measures the distance that the curve is away from the data points. We do not want this to be too large obviously, and too small would be an interpolation, so somewhere in between is desirable.

The variance measures how much the model depends on that one sample. Again, it is fairly obvious that we do not want this to be too big or too small.

The compact form obtained for the bias (2.24) and variance (2.25) expressions are suitable for our applications. However, they only give a limited view of the behavior of the bias and variance functions when the design, sample size or neighborhood change. Here we provide some simple asymptotic approximations to the bias and variance functions based on the derivations of Loader (1999b, p.38-42) and Fan and Gijbels (1996, p.101-107). These results stated below for one independent variable are not new. Tsybakov (1986) and Müller (1987) were among the first to derive these for local regression, although similar expressions for kernel regression and density estimation have been known for much longer.

To state asymptotic results, we need to make assumptions about how the sequence of design points $x_1, \ldots, x_n$ behaves as n increases. In case of equally spaced points, we refer to a regular design. More generally, a regular design generated by a density $\phi(u)$ defines $x_{i,n}$ to be the solution of

$$\frac{i - 0.5}{n} = \int_{-\inf}^{x_{i,n}} \phi(u)\, du.$$

Let $\widehat{\psi}(x_i)$ be a local polynomial fit of degree p. Assuming that $\psi(x_i)$ is $p+2$ times differentiable, we can expand $\psi(.)$ in a Taylor series around x_i:

$$\begin{aligned}\psi(x_j) =&\psi(x_i) + (x_j - x_i)\psi'(x_i) + \ldots + (x_j - x_i)^p \frac{\psi^{(p)}(x_i)}{p!}\\ &+ (x_j - x_i)^{p+1}\frac{\psi^{(p+1)}(x_i)}{(p+1)!} + (x_j - x_i)^{p+2}\frac{\psi^{(p+2)}(x_i)}{(p+2)!} + \ldots\end{aligned}$$

As an application of Henderson's theorem, we know that each row sums to 1, $\sum_{j=1}^n s_j = 1$, and $\sum_{j=1}^n s_j(x_i)\ (x_j - x_i)^k = 0$ for $1 \le k \le P$. This leads to

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}(x_i)\right] - \psi(x_i) =&\frac{\psi^{(p+1)}(x_i)}{(p+1)!}\sum_{j=1}^n s_j(x_i)\ (x_j - x_i)^{p+1} \\ &+ \frac{\psi^{(p+2)}(x_i)}{(p+2)!}\sum_{j=1}^n s_j(x_i)\ (x_j - x_i)^{p+2} + \ldots\end{aligned} \tag{2.26}$$

The bias has a leading term involving the $(p+1)$st derivative $\widehat{\psi}^{(p+1)}(x_i)$. We keep the $\widehat{\psi}^{(p+2)}(x_i)$ term in (2.26) because in case the design points are equally spaced, the rows of the smooth weight diagram are symmetric around the fitting point x_i. Then, if p is even, $p+1$ is odd and $\sum_{j=1}^n s_j(x_i)\ (x_j - x_i)^{p+1} = 0$ by symmetry, similarly to Müller (1987, p.234 Corollary 3) for kernel regression. Thus the first term in the bias expansion disappears. In that case the second term is dominant.

From expression (2.22), the matrix $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ has components t_k of the form $\sum_{i=1}^n w_j\big(x_j - x_i\big)^k$. Under mild conditions, in particular $nh^d \to \infty$,

$$\frac{1}{nh^d}\sum_{j=1}^n w_j^l \frac{(x_j - x_i)^k}{h^k} = \int W(v)^l v^k \phi(x_i + hv)\, dv + o(1). \tag{2.27}$$

This result is valid for fixed h. Under the additional assumption $h \to 0$, (2.27) simplifies to

$$\frac{1}{nh^d}\sum_{j=1}^n w_j^l \frac{(x_j - x_i)^k}{h^k} = \phi(x_i)\int W(v)^l v^k dv + o(1). \tag{2.28}$$

For regular design, the limit (2.28) follows from the theory of Riemann sums, see Loader (1999b, p.38-39). Applying (2.27) and (2.28) to the matrix $\boldsymbol{X}^T\boldsymbol{W}^l\boldsymbol{X}$ gives

$$\begin{aligned}&\frac{1}{nh^d}\boldsymbol{H}^{-1}\,\boldsymbol{X}^T\,\boldsymbol{W}^l\,\boldsymbol{X}\,\boldsymbol{H}^{-1}\\ &\quad= \begin{cases}\int W(v)^l\boldsymbol{c}(v)\boldsymbol{c}(v)^T\phi(x_i + hv)dv + o(1) & h \text{ fixed}\\ \phi(x_i)\int W(v)^l\boldsymbol{c}(v)\boldsymbol{c}(v)^T dv + o(1) & h \to 0,\end{cases}\end{aligned} \tag{2.29}$$

where $\boldsymbol{H}$ is a diagonal matrix with elements $1, h, \dots, h^P$ and $\boldsymbol{c}(v)$ is the vector of the fitting functions, $\boldsymbol{c}(v) = (1, v, \dots, v^p/p!)^T$.

Asymptotic approximations to quantities such as the bias and variance are now easily derived. Under the small bandwidth limits, the variance (2.25) has the following asymptotic approximation

$$\mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] = \frac{\sigma^{*2}(x_i)}{nh^d\ \phi(x_i)} \boldsymbol{e}_1^T \Lambda_1^{-1} \Lambda_2 \Lambda_1^{-1} \boldsymbol{e}_1 + o((nh)^{-1}), \tag{2.30}$$

where $\Lambda_l^{-1} = \int W(v)^l \boldsymbol{c}(v)\boldsymbol{c}(v)^T dv$. Substituting (2.29) into expression (2.6) for the local regression estimate leads to

$$\begin{aligned} \widehat{\psi}(x_i) &\approx \frac{1}{nh^d \phi(x_i)} \boldsymbol{e}_1^T \Lambda_1^{-1} \boldsymbol{H}^{-1}\ \boldsymbol{X}^T\ \boldsymbol{W}\boldsymbol{y} \\ &= \frac{1}{nh^d \phi(x_i)} \sum_{j=1}^n W^\circ\left(\frac{x_j - x_i}{h}\right) y_j, \end{aligned}$$

where

$$W^\circ(v) = \boldsymbol{e}_1^T \Lambda_1^{-1} \boldsymbol{c}(v) W(v). \tag{2.31}$$

The weight function $W^\circ(v)$ is the asymptotically equivalent kernel. Its depends on the degree of fit and the original weight function $W(v)$. Often equivalent kernels provide poor approximations but their merit is to simplify theoretical computations considerably, see Loader (1999b, p.40) and Fan and Gijbels (1996, p.101-107). The asymptotic variance (2.30) becomes

$$\mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] \approx \frac{\sigma^{*2}(x_i)}{nh^d\ \phi(x_i)} \int W^\circ(v)^2 dv.$$

The first term of the bias expansion (2.26) is approximated by

$$b(x_i) = \frac{h^{p+1}\psi^{(p+1)}(x_i)}{(p+1)!} \int v^{p+1} W^\circ(v) dv + o(h^{p+1}).$$

If p is even and $W(v)$ is symmetric, $\int v^{p+1} W^\circ(v) dv = 0$. The dominant bias arises from the second term of (2.26), which has size $o(h^{p+2})$. For p even, we obtain

$$b(x_i) = h^{p+2}\left(\frac{\psi^{(p+1)}(x_i)\phi'(x_i)}{(p+1)!\phi(x_i)} + \frac{\psi^{(p+2)}(x_i)}{(p+2)!}\right) \int v^{p+2} W^\circ(v) dv + o_p(h^{p+2}).$$

For more details and additional assumptions see Ruppert and Wand (1994), Loader (1999b, p.38-42) and Fan and Gijbels (1996, p.101-107) among others.

When we look at the asymptotic bias and variance, we find interesting features. In the leading term of the bias the smoothing parameter is found in the numerator while for the variance it is found in the denominator. Thus, for $\lambda \to 0$ the variance becomes large whereas the bias becomes low. As an illustration, Figure 2.7 shows the squared bias, variance and MSE in one graph. We see that the bias-variance trade-off is evident as well as the fact that the minimization of the mean squared error is a compromise between the two.

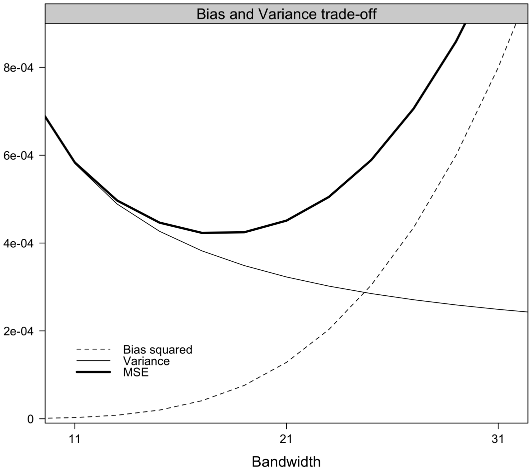


Figure 2.7: *Squared bias (thin dashed), variance (thin solid) and mse (thick solid) of a local polynomial fit for the Dutch male population, 2008. Source: HMD.*

The intuition behind this is as follows. When the local polynomial does not fit well, i.e. the bandwidth is too large, the bias is large and hence also the residual sum of squares. When the bandwidth is too small, the variance term tends to be larger. So the MSE quantity protects against both extreme choices.

In addition, there is a difference between p odd and p even, leading to the same order of the bias for $p = 0$ (constant) and $p = 1$ (local linear), as well as $p = 2$ (local quadratic) and $p = 3$ (local cubic), and so on. For instance, for $p = 0$ as well as for $p = 1$, the leading term of the bias contains h^2, whereas for $p = 2$ and $p = 3$ one obtains h^4.
One last feature as is seen in the formulas, is that for p odd the bias does not depend on the density $\phi(x_i)$; in this sense the estimate is *design adaptive* in the terminology of Fahrmeir and Tutz (2001). For p even, the term contains the density $\phi(x_i)$ in the denominator, meaning that bias is lower if the density $\phi(x_i)$ is high.

To give an illustration on how the trade-off between bias and variance works in practice, consider Figures 2.8 and 2.9.

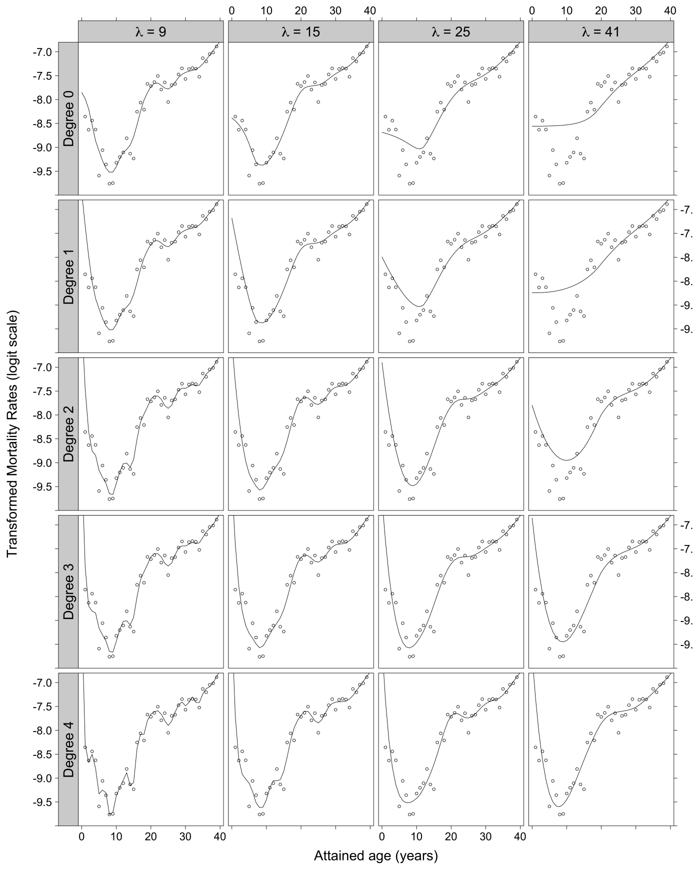


Figure 2.8: *Fits for four bandwidths and five local fitting methods for the Dutch male population, 2008. Source: HMD.*

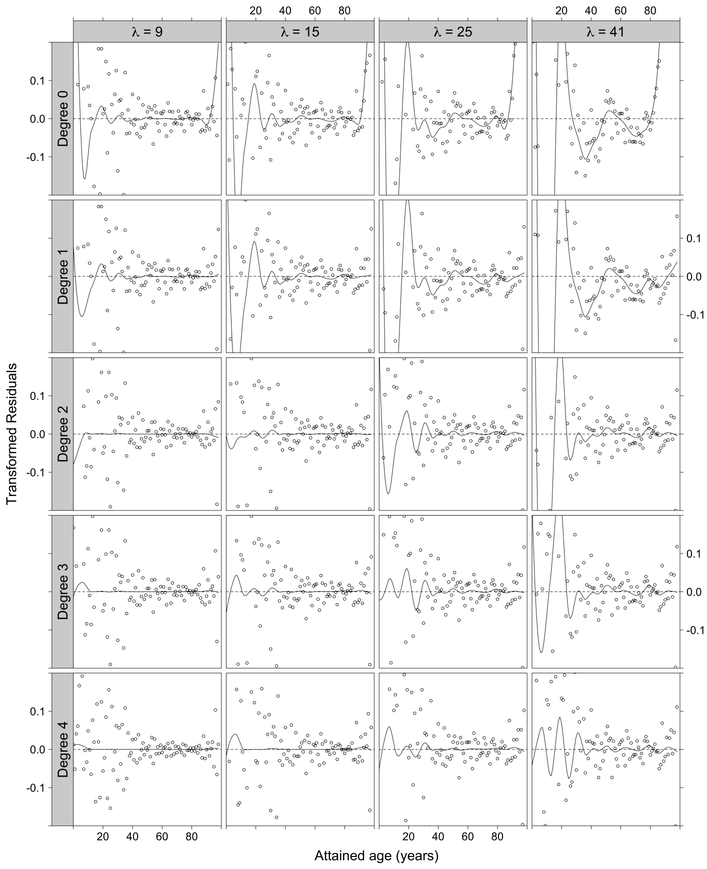


Figure 2.9: *Transformed Residuals plots for the fits in the left panel for the Dutch male population, 2008. Source: HMD.*

Figure 2.8 shows fits for Dutch Male data (year 2008) and age range from 0 to 36 where the curvature is the most pronounced. Each column contains fits for one value of λ ($\lambda = 9$ to 41). The rows show the fits for degrees 4 to 0. The fits have been computed using a triweight weight function.
Figure 2.9 shows the residuals for each of the 20 fits in Figure 2.8, but for the full age range, from 0 to 98. Superimposed on each plot is a *loess* smooth.

For local constant fitting, $p = 0$, a small λ is needed to capture the dependence of the probability of death on age without introducing an undue distortion. Even for $\lambda = 9$, the plot of residuals suggests a lack of fit at the youngest age, that is, at the left boundary, where there is a large curvature. Local constant fitting can neither capture a quadratic effect at the left boundary, nor the hump around 18 years old. A similar remark can be made for a local linear fitting, when $p = 1$, even for small values of λ.

As we increase λ to get a smoother fit, the local constant and linear fits introduce a major distortion, and miss the mortality patterns. As λ increases the neighborhood size increases, the bias tends to increase, and the variance tends to decrease. However, one can observe that a high polynomial degree will usually provide a better approximation than a low polynomial degree.

Thus as we increase the polynomial degree, we reduces the bias and the curvature at youngest ages is capture as it is illustrated in Figure 2.9.
To some extent, the effects of the polynomial degree and bandwidth are confounded. For example, if a local quadratic and a local linear estimate is computed using the same bandwidth, the local quadratic estimate is more variable. But the variance increase can be compensated by increasing the bandwidth.

For mortality data there is a pronounced dependence of the response on the independent variable, illustrated by valleys and peak at youngest ages. Therefore we might expect that locally, taking a small λ and a quadratic or cubic family provides a reasonable approximation. This, however, must be done judiciously, since there must be a sufficient number of observations to support the extra degrees of freedom.

The issue is how to choose the value of the smoothing parameters to get the right balance of bias and variance. The answer is to try and satisfy some optimality criteria and it is discussed in the following section.

2.5 Fitting criteria and choice of the smoothing parameters

Where do we look to make the choices of the smoothing parameters? The answer is, as we have emphasized, to treat choices of bandwidth, polynomial degree and weight function as modeling the data and to use formal model selection criteria and graphical diagnostics to provide guidance.

The development of methods of parametric regression has had a long history of using model selection criteria and diagnostic methods for parametric models fitted to regression data, see Cleveland and Loader (1996).

From parametric regression, there are two families of criteria based on prediction error and on estimation error, respectively.

2.5.1 Criteria based on prediction error

To evaluate the performance of the estimator we may focus on the prediction problem:

- If the fitted regression curve is used to predict new observations, how good will the prediction be?
- If a new observation is made at $x_i = x_0$, and the response y_0 is predicted by $\widehat{y}_0 = \widehat{\psi}(x_0)$, what is the prediction error?

One measure is

$$\mathbb{E}\left[(y_0 - \hat{y}_0)^2\right].$$

The method of cross-validation (CV) can be used to estimate this quantity. In turn, each observation (x_i, y_i) is omitted from the dataset, and is *predicted* by smoothing the remaining $n-1$ observations. This leads to the CV score

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{\psi}^{-i}(x_i)\right)^2. \tag{2.32}$$

where $\widehat{\psi}^{-i}(x_i)$ denotes the smoothed estimate when the single data point (x_i, y_i) is omitted from the dataset; only the remaining $n-1$ data points are used to compute the estimate.

The leave-one-out cross validation criterion was introduced for parametric models as the PRESS procedure (prediction error sum of squares). Formally computing each of the leave-one-out regression estimates $\widehat{\psi}^{-i}(.)$ would take a lot of computer time, and so at first sight computation of the CV as in (2.32) looks prohibitively expensive. But there is a remarkable simplification, valid for all common linear smoothers, involving correcting the weights computed for the full set of n data points. We can calculate all the leave-one-out smooths from the original smooth weight diagram $\boldsymbol{S}$.

Actually, it is not clear what leave-one-out means in the context of smoothing. In general there is not necessarily a relationship between a smoother for n data pairs and a smoother for $n-1$ data pairs. One method of finding such a relationship is to note that any reasonable smooth weight diagram is constant preserving. Thus if we want to use the same smooth weight diagram with the i-th row and column deleted resulting in an $(n-1)\times(n-1)$ smooth weight diagram, we must renormalize the rows to sum to one.

Let us recall that $s_i(x_i)$ denote the diagonal elements of the original $n \times n$ smooth weight diagram $\boldsymbol{S}$. When we delete the i-th column, then the i-th row sums to $1 - s_i(x_i)$. So that's what we divide by to renormalize. For linear smoothers $\widehat{\psi}(x_i) = \sum_j s_j(x_i)\, y_j$, one may choose

$$\widehat{\psi}^{-i}(x_i) = \frac{1}{1 - s_i(x_i)} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_j(x_i) y_j, \tag{2.33}$$

where the modified weights $s_j(x_i)/(1 - s_i(x_i))$ now sum to 1. Thus, one gets the simple form

$$\widehat{\psi}^{-i}(x_i) = \frac{1}{1 - s_i(x_i)} \widehat{\psi}(x_i) - \frac{s_i(x_i)}{1 - s_i(x_i)} y_i.$$

Then the essential term $y_i - \widehat{\psi}^{-i}(x_i)$ in (2.32) is given by

$$y_i - \widehat{\psi}^{-i}(s_i(x_i)) = \frac{y_i - \widehat{\psi}(x_i)}{1 - s_i(x_i)},$$

and may be computed from the regular fit $\widehat{\psi}(x_i)$ based on n observations and weights $s_i(x_i)$. By using (2.33) one gets the criterion

$$CV = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{\psi}(x_i)}{1 - s_i(x_i)} \right)^2 .$$

Generalized cross-validation (GCV), as introduced by Craven and Wahba (1979), replaces $s_i(x_i)$ by the average $\sum_i s_i(x_i)/n$. The resulting criterion is easier to compute as it is the single average squared error corrected by a factor.

$$\begin{aligned} GCV &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{\psi}(x_i)}{1 - \frac{1}{n} \sum_j s_j(x_j)} \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{\psi}(x_i)}{1 - \operatorname{tr}(\boldsymbol{S})/n} \right)^2 \\ &= \frac{1}{n(1 - \operatorname{tr}(\boldsymbol{S})/n)^2} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{\psi}(x_i) \right)^2 \\ &= \frac{n}{(n - \upsilon_1)^2} \sum_{i=1}^{n} \left(y_i - \widehat{\psi}(x_i) \right)^2 , \end{aligned}$$

In this form, the criterion is very sensitive to the design space. Table 2.4 presents the proportion of the residuals sum of squares (RSS) in the boundaries given by the local polynomial regression targeting the mortality rates q_i

		Male pop.		Female pop.	
Target	p	Left	Right	Left	Right
q_i	0	0.03	98.79	0.02	99.03
	1	0.19	88.91	0.25	86.29
	2	0.17	94.83	0.49	92.02
	3	0.11	94.73	0.31	91.49
	4	0.07	94.32	0.20	90.61
logit(q_i)	0	82.91	7.82	83.52	9.14
	1	82.98	0.83	86.24	0.28
	2	76.86	1.44	72.41	0.60
	3	72.32	1.67	60.63	0.80
	4	69.74	1.90	58.71	0.85

Table 2.4: *Proportion of the RSS in the boundaries (in %) by the local polynomial regression, computed with boundary correction type 1 and a triweight weight function and* $\lambda = 19$*, for the Dutch male and female population, 2008. Source: HMD.*

on the original scale and on the logit scale by fixing arbitrarily the bandwidth λ to 19 observations. The proportion of the RSS varies with the underlying structure of the data as well as the degree of polynomial p chosen. A constant fit leads to the highest disturbing nuisance. The performance in the boundaries increases with the degree of polynomial. By targeting the mortality rates q_i on the logit scale, most of the curvature appears in the right boundary. The proportion of the RSS in the right boundary represents at least 88.91 % and 86.29 %, for the male and female population respectively. It is apparent that the selection of the parameters is driven by minimizing the RSS in the left boundary rather than the whole data.

However, the generalized cross validation can be seen as a special case of minimizing

$$\log(\widehat{\sigma}^2) + \varphi(\boldsymbol{S}),$$

where $\varphi(.)$ is a penalty function that decreases with increasing smoothness of $\widehat{\psi}$ and $\widehat{\sigma}^2 = (1/n)\sum_i (y_i - \widehat{\psi}(x_i))^2$ is the average squared residuals, see Hurvich *et al.* (1998, p.273). Table 2.5 presents the proportion of the natural logarithm of RSS in the boundaries.

By taking the natural logarithm of the average square errors, the variability is reduced and the criterion less affected by the boundaries.

The choice $\varphi(\boldsymbol{S}) = -2\log(1 - \mathrm{tr}(\boldsymbol{S}/n))$ yields the GCV criterion, while $\varphi(\boldsymbol{S}) = 2\,\mathrm{tr}(\boldsymbol{S}/n)$ yields the AIC criterion

$$\log(\widehat{\sigma}^2) + 2\,\mathrm{tr}(\boldsymbol{S})/n. \tag{2.34}$$

		Male pop.		Female pop.	
Target	p	Left	Right	Left	Right
q_i	0	10.58	4.70	10.4	4.46
	1	10.07	5.65	9.92	5.87
	2	10.08	5.67	9.73	6.38
	3	10.01	5.72	9.71	6.26
	4	10.15	5.66	9.92	5.97
$\text{logit}(q_i)$	0	4.07	6.72	3.59	6.88
	1	2.92	10.37	2.39	13.67
	2	3.76	10.28	5.56	12.08
	3	3.13	10.43	4.56	12.50
	4	3.48	10.44	4.86	11.08

Table 2.5: *Proportion of the* $\log(RSS)$ *in the boundaries (in %) by the local polynomial regression, computed with boundary correction type 1 and a triweight weight function and* $\lambda = 19$*, for the Dutch male and female population, 2008. Source: HMD.*

The usual form of the AIC criterion is given by $AIC = -2\{\log(L) - p\}$, where $\log(L)$ is the maximal log-likelihood and p stands for the number of parameters. Under the assumption of normally distributed responses $y_i \sim N(\mu_i, \sigma^2)$, one obtains, apart from additive constants,

$$AIC = n\left(\log(\widehat{\sigma}^2) + \frac{2}{n}\,p\right).$$

In (2.34) the trace $\text{tr}(\boldsymbol{S})$ plays the role of the effective number of parameters used in the smoothing fit, see Loader (1999b). Thus, replacing p by $\text{tr}(\boldsymbol{S})$ leads to (2.34). If $\varphi(\boldsymbol{S}) = -\log\{1 - 2\,\text{tr}(\boldsymbol{S})/n\}$ is chosen, one obtains the criterion suggested by Rice (1984).

A last alternative can be mentioned. Hurvich *et al.* (1998, p.277) proposed to use the criterion $AICC$, a corrected version of the AIC,

$$AICC = \log(\hat{\sigma}^2) + \frac{1 + tr(\boldsymbol{S})/n}{1 - (tr(\boldsymbol{S}) + 2)/n} = \log(\hat{\sigma}^2) + 1 + \frac{2(\text{tr}(\boldsymbol{S}) + 1)}{n - \text{tr}(\boldsymbol{S}) - 2}. \quad (2.35)$$

The first term in (2.35) measures the quality of the adjustment while the second term evaluates the model complexity.

It follows from Härdle *et al.* (1988, p.88) that all the so-called *classical* selectors considered here are asymptotically equivalent. Given this, we might wonder why they might exhibit noticeably different performances in practice. The reason, exposed in Hurvich *et al.* (1998, p.277) is that the asymptotic theory assumes $\text{tr}(\boldsymbol{S}) \to 0$, a situation that is not consistent with a small smoothing parameter λ.

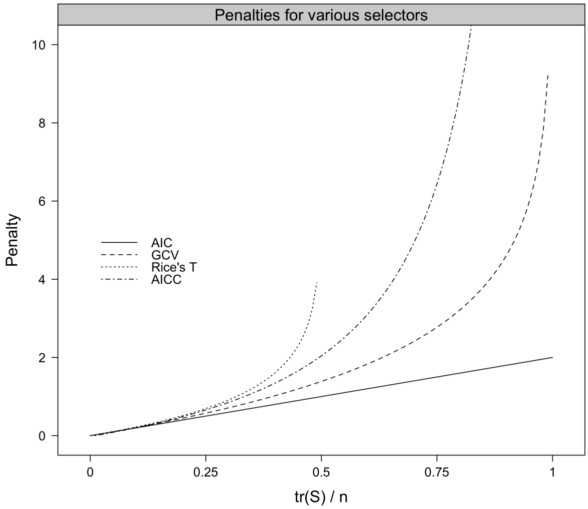


Figure 2.10: *$\varphi(.)$ penalties for various selectors as a function of $tr(\boldsymbol{S})/n$.*

Figure 2.10 makes this distinction clear. It gives the penalty functions $\varphi(\boldsymbol{S})$ as a function of $\text{tr}(\boldsymbol{S})$ for GCV, Rice's T statistic, the AIC and $AICC - 1$ (subtracting 1 from $AICC$ makes it comparable with the other selectors, and does not affect its smoothing parameter choices; since $AICC$ depends on n, its curve is given for $n = 100$).

All four functions become indistinguishable at the left-hand end of the plot, which corresponds to $\text{tr}(\boldsymbol{S})/n \to 0$ and the usual asymptotics. The criteria differ markedly for a small smoothing parameter (large $\text{tr}(\boldsymbol{S})/n$), however, with a sharper rise corresponding to a heavier penalty against under-smoothing. The AIC and GCV have relatively weak penalties; this accounts for their tendencies to lead to under-smoothing. Rice's T statistic, in contrast, has a very strong penalty, as it is effectively infinite for $\text{tr}(\boldsymbol{S})/n \geq 0.5$. This means that Rice's T must lead to over-smoothing when a very small smoothing parameter is appropriate. $AICC$ occupies a position between these two extremes, being less susceptible to both the under-smoothing of the AIC and GCV and the over-smoothing of Rice's T statistic.

In consequence, we would use the AIC or the GCV selector when the data present a smooth pattern, as we are more likely to look for an under-smoothed fit. While Rice's T statistic and $AICC$ would be used alternatively, as they lead to an over-smoothed fit which is satisfactory when the data are volatile.

2.5.2 Criteria based on estimation error

Alternatively, one can consider methods motivated by estimation error: how well does $\widehat{\psi}(x)$ estimate the true mean $\psi(x)$? A risk function measures the distance between the true regression function and the estimate; for example,

$$R(\psi, \widehat{\psi}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbb{E}\left[\left(\widehat{\psi}(x_i) - \psi(x_i)\right)^2\right]. \tag{2.36}$$

Ideally, a good estimate would be one with low risk. But since ψ is unknown, $R(\psi, \widehat{\psi})$ cannot be evaluated directly. Instead, the risk must be estimated. Focusing on the squared-error risk, we have the bias-variance decomposition

$$\sigma^2 \, R(\psi, \widehat{\psi}) = \sum_{i=1}^{n} \mathbb{V}\text{ar}\left[\widehat{\psi}(x_i)\right] + \sum_{i=1}^{n} \left(\mathbb{E}\left[\widehat{\psi}(x_i)\right] - \psi(x_i)\right)^2.$$

Cleveland *et al.* (1988, p.100) compute the expected value of the residual sum of squares of $\widehat{\psi}(x_i)$ as

$$\mathbb{E}\left[\sum_{i=1}^{n} \left(y_i - \widehat{\psi}(x_i)\right)^2\right] = \sum_{i=1}^{n} \mathbb{V}\text{ar}\left[y_i - \widehat{\psi}(x_i)\right] + \sum_{i=1}^{n} \left(\mathbb{E}\left[\widehat{\psi}(x_i)\right] - \psi(x_i)\right)^2.$$

Likewise, in matrix notation, knowing that

$$\begin{aligned}\mathbb{V}\text{ar}\left[\boldsymbol{y} - \widehat{\boldsymbol{\psi}}\right] &= \mathbb{V}\text{ar}\left[(\boldsymbol{I} - \boldsymbol{S})\,\boldsymbol{y}\right] \\ &= \sigma^2\,(\boldsymbol{I} - \boldsymbol{S})\,(\boldsymbol{I} - \boldsymbol{S})^T \\ &= \sigma^2\left(\boldsymbol{I} - \boldsymbol{S} - \boldsymbol{S}' + \boldsymbol{S}\boldsymbol{S}^T\right),\end{aligned}$$

where $\boldsymbol{y}$ is the vector of the response values and $\boldsymbol{I}$ is the identity matrix, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\,\boldsymbol{y} - \widehat{\boldsymbol{\psi}}\,\right\|^2\right] &= \sigma^2 \,\text{tr}\left(\boldsymbol{I} - \boldsymbol{S} - \boldsymbol{S}^T + \boldsymbol{S}\boldsymbol{S}^T\right) + \boldsymbol{b}^T\boldsymbol{b} \\ &= \sigma^2\left(n - 2\,\text{tr}(\boldsymbol{S}) + \text{tr}(\boldsymbol{S}\boldsymbol{S}^T)\right) + \boldsymbol{b}^T\boldsymbol{b} \\ &= \sigma^2(n - 2\,\upsilon_1 + \upsilon_2) + \boldsymbol{b}^T\boldsymbol{b},\end{aligned}$$

with $\boldsymbol{b}$ the bias vector. Hence Cleveland *et al.* (1988, p.100) estimate of the bias term $\boldsymbol{b}^T\boldsymbol{b}$ as

$$\mathbb{E}\left[\left\|\,\boldsymbol{y} - \widehat{\boldsymbol{\psi}}\,\right\|^2\right] - \sigma^2(n - 2\,\upsilon_1 + \upsilon_2). \tag{2.37}$$

With (2.25) and (2.37), and making the proper rearrangements, an unbiased estimate of (2.36) is

$$\widehat{R}(\psi, \widehat{\psi}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \widehat{\psi}(x_i)\right)^2 - n + 2\,\upsilon_1.$$

This statistic is known as the *Cp* criterion, and has been introduced by Mallows (1973) for parametric regressions. It provides an unbiased estimate of $R(\psi, \widehat{\psi})$. This statistic was extended to local regression by Cleveland and Devlin (1988). To implement the Cp criterion (or unbiased risk estimate) one needs to know an estimate of σ^2. The recommendation of Cleveland *et al.* (1988) is to replace it by an estimate from a local fit for which it seems reasonable to suppose the bias is small. This means estimating $\widehat{\sigma}^2$, where λ is small, by

$$\widehat{\sigma}^2 = \frac{1}{n - 2\,\upsilon_1 + \upsilon_2} \sum_{i=1}^{n} \Big(y_i - \widehat{\psi}(x_i) \Big)^2 .$$

2.5.3 Plug-in method and theoretical bandwidth

Since the choice of the smoothing parameters is of crucial importance to the performance of the estimator, this has been a topic of extensive research. The work has been most predominantly in the setting of kernel density estimation, see Loader (1999a). The bandwidth selection methods can be divided into two broad classes, the *classical* and *plug-in* methods.

Classical methods are Cp, CV, GCV and AIC and variations, introduced in Section 2.5.1 and 2.5.2. We have seen these are more or less natural extensions of methods used in parametric modeling.

On the other hand, plug-in methods rely on an approximation of the bias via Taylor series expansions. The bias of an estimate $\widehat{\psi}$ is written as a function of the unknown ψ, and is approximated through Taylor series expansions. A pilot estimate of ψ is then plugged in to derive an estimate of the bias and hence an estimate of the mean squared error. The optimal bandwidth minimizes this estimated measure of fit:

$$\widehat{MSE}_{p,v}(x_i, h) = (v!)^2 \left(\widehat{\text{bias}}^{\,2}_{p,v}(x_i) + \widehat{\text{var}}_{p,v}(x_i) \right). \tag{2.38}$$

With the estimated MSE, Fan and Gijbels (1995b, p.378) formulate a bandwidth selection rule as follows: Fit a polynomial of order $p + a$ (choosing $a = 2$) and find the *pilot* bandwidth h^* that minimizes the integrated residual squares criterion,

$$IRSC(h) = \int_{[x_{min}, x_{max}]} RSC(t, h)\, dt,$$

with the RSC defined as

$$RSC(x_i, h) = \widehat{\sigma}^{*2}(x_i)\,(1 + (p + 1)/N)\,, \tag{2.39}$$

where N^{-1} is the first diagonal element of the matrix

$$\left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^2 \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1},$$

and $\widehat{\sigma}^{*2}(x_i)$ is the normalized weighted residual sum of squares after fitting locally a $(p + a)$th order polynomial defined as expression (2.19). Note that N reflects the effective number of local data points since $\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}(x_i)\right] =$

$\sigma^2(x_i)/N$ by equation (2.15). The criterion is not suffering from the boundary effects as the RSS component is weighted by the variance term N. Because the variance is larger at the boundaries, the resulting contributions of the observations are lower.

The intuition behind statistic (2.39) is that when the local polynomial does not fit well (the bandwidth is too large), the bias is large and hence also the residual sum of squares $\widehat{\sigma}^{*2}(x_i)$. When the bandwidth is too small, the variance term N tends to be larger. So the RSC quantity protects against both extreme choices.

Thus, having the optimal bandwidth h^* for estimating β_{p+1}, obtain estimates $\widehat{\beta}^*_{p+1}(x_i)$, $\widehat{\beta}^*_{p+2}(x_i)$ and $\widehat{\sigma}^{*2}(x_i)$. With these estimated parameters, compute the estimated bias $\widehat{\text{bias}}_{p,v}(x_i)$ and variance $\widehat{\text{var}}_{p,v}(x_i)$ of $\widehat{\beta}_v$, which are the $(v+1)$st element of vector (2.20) and the $(v+1)$st diagonal element of the estimated expression (2.23), respectively. Combining these estimates yields (2.38) as the estimated MSE. This leads to the bandwidth selector

$$\widehat{h}_{p,v} = \arg\min_h \left\{ \int_{[x_{min}, x_{max}]} \widehat{MSE}_{p,v}(t,h)\, dt \right\}.$$

The key problem here is the bias estimation. The current approach makes it possible to assess the bias without going into deep asymptotics. It differs from the usual plug-in procedure (see for instance Park and Marron (1990), Sheather and Jones (1991), and Gasser *et al.* (1991)) in the sense that the t_k, defined by expression (2.22), are not further replaced by their asymptotic counterparts. The quantities t_k are already known, and Fan and Gijbels (1995b, p.377) argue that replacing them by their corresponding asymptotic quantities introduces not only some extra approximation but also extra unknown parameters such as the marginal density $\psi_X(x_i)$.

However, for higher order fits ($p \geq 2$) such as local quadratic or cubic fits, bias estimation essentially amounts to estimating fourth order derivatives about which the data contains little or no information, see Cleveland and Loader (1996, p.33). Hence plug-in bandwidth selection alone does not solve the bandwidth problem, but replaces the problem with the problem of choosing pilot bandwidths.

2.5.4 Graphical Diagnostics and heuristics

In practice one needs to choose λ and the fitting variables to balance the trade-off between bias and variance. To find such constellation, we can compute the criteria presented in Section 2.5.1 and 2.5.2 for different fits and select the fit with the lowest score.

However, as argued strongly by Cleveland and Devlin (1988), this discards much of the information about the trade-off between the contributions of variance and bias to the mean-square-error. Cleveland and Devlin then introduced graphical tools for displaying these statistics.

As an illustration, Figure 2.11 displays the AIC scores against the fitted degrees of freedom $\operatorname{tr}(\boldsymbol{S}\boldsymbol{S}^T)$. We use the fitted degrees of freedom, rather than the smoothing parameter, as the horizontal axis. This aids interpretation: 10 degrees of freedom represents a smooth model with very little flexibility while 30 degrees of freedom represents a noisy model showing many features. It also aids comparability as we can compute criteria scores for other polynomial degrees or for other smoothing methods and added to the plot.

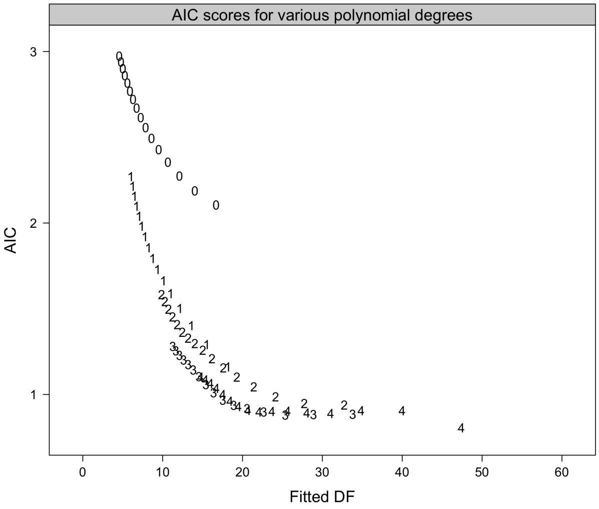


Figure 2.11: *AIC scores for various polynomial degrees and triweight weight function for Dutch Male population, 2008. Source: HMD.*

From Figure 2.11, the lowest score corresponds to a quartic fit with $\upsilon_2 = 47.41$, leading to a smoothing window of 11 points. Following Loader (1999b, p.33), any model with a score near the minimum is likely to have a similar predictive power. The flatness of the plot reflects the uncertainty in the data, and the resulting difficulty in choosing the smoothing parameters. Hence Cleveland and Devlin (1988) elect to use a larger λ and recommend to choose the smoothing parameters at the point when the criterion reaches a plateau after a steep descent. In consequence, we would select a cubic fit with $\upsilon_2 = 18.46$, corresponding to a bandwidth of 19 observations.

In parallel, we shall use fitting and corresponding residuals plots. Figure 2.12 shows the fits and corresponding residuals plot for the constellation picked by the lowest AIC score and the one elected using our graphical diagnostic. Both of the fits have been computed with a triweight weight function.

One always has to look at residual plots in conjunction with looking at plots of the fits. Superimposed on the residual plot is a *loess* smooth with local quadratic fitting and $\lambda = 19$. The smoothed curves help the search for clusters of residuals that may indicate lack of fit. Such residual plots provide a powerful diagnostic that nicely complements the selection criteria.
The diagnostic plots can show lack of fit locally, and we have the opportunity to judge the lack of fit based on our knowledge of both the mechanism generating the data and our knowledge of the performance of the smoothers used in the fitting. Here, the process is not to judge a fit adequate if a smooth curve on its residual plot is flat. A flat curve means simply that no systematic, reproducible lack of fit has been detected. The fit may well be too noisy as we can see from the fit computed with the lowest AIC score. It stays too close to an interpolation since trends in small parts of the data are interpreted as more widespread trends. Then, for small datasets, the fit is very nearly interpolating the data which results in unacceptably high variance.

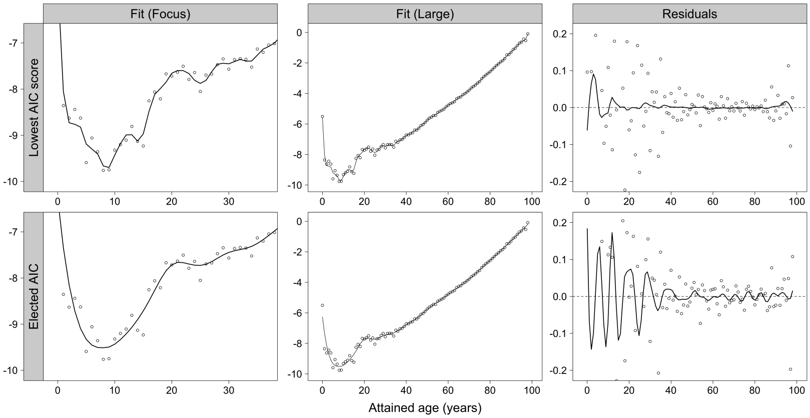


Figure 2.12: *Fits and residuals plots elected by the AIC score with a triweight weight function for Dutch Male population, 2008. Source HMD.*

Loader (1999a) has emphasized the importance of not relying blindly on any bandwidth selector to produce the right bandwidth automatically. If one applies a bandwidth selector and plots the fit, one gets a one-sided view of the bias-variance trade-off, seeing the variance but not the bias. It is extremely important to use appropriate residual diagnostics to look for lack of fit. Likewise, plotting the AIC or variations provides valuable diagnostic information as to how difficult the bandwidth selection is; a flat plot suggests that different features of the data may be competing for attention at different bandwidths. Plug-in approaches discard this information.

Plug-in approaches make substantial prior assumptions about the required bandwidth through the specification of tuning parameters for pilot estimates. They will fail if this information is wrong. The plug-in methods obtain much of their information from the data through the use of higher order pilot estimates. If classical approaches are also allowed to consider higher order methods, better estimates result. Loader (1999a) does not claim that classical approaches such as AIC and variations will produce the best estimates, but rather that, used properly, the results will often be more informative than other bandwidth selection.

To conclude, note that in practice relying exclusively on a global criterion is unwise because a global criterion does not provide information about where the contributions to bias and variance are coming from.

In the next section, we use two examples to graduate the mortality data through the choices of the weight function, the bandwidth, and the parametric family. We use the fitting criteria and graphical diagnostics to guide the modeling.

2.6 Applications

2.6.1 The data

In this section we present two applications of local polynomial fitting method for graduation. The computations are carried out with the help of the software R, R Development Core Team (2012). Figure 2.13 displays the observed statistics of the two datasets.

i. The data for the first application are reported by the Human Mortality Database (2012). The dependent variable is the observations in a logit scale of the one-year probability of death for the Dutch Male population for the year 2008 at age x_i with $i = 1, \ldots, 99$.

ii. The data for the second application are the Female counterpart.

2.6.2 Choice of the constellation of the smoothing parameters

We graduate the mortality data through the choices of the weight function, the bandwidth and the parametric family. In practice one needs to choose λ and the fitting variables to balance the trade-off between bias and variance. To find such a constellation, we use the criteria presented in Section 2.5 and graphical diagnostics to guide our modeling.

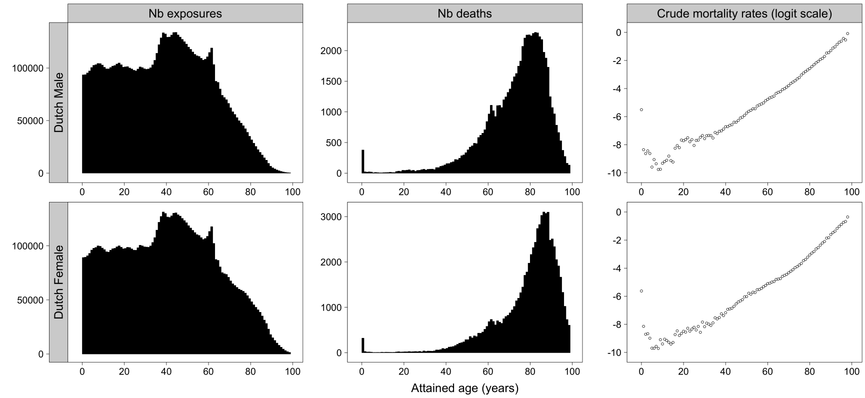


Figure 2.13: *Observed statistics for Dutch Male and Female population, 2008. Source: HMD.*

Both datasets present a relatively wiggly pattern. For these applications we picked the optimal constellation selected by Rice's T statistic and $AICC$ as the final fit. Due to strong penalties on $tr(S)/n$, these criteria tend to lead to over-smoothing, which, considering the underlying pattern of the data, is satisfactory. However, the selected bandwidth should not be too large to capture the structure at the left boundary and the accident hump which we believe as true.

Table 2.6 displays the elected optimal constellation of smoothing parameters for the local polynomials method together with the fitted degrees of freedom.

	λ	Degree	$W(.)$	Fitted DF
Dutch Male	19	3	Triweight	18.46
Dutch Female	21	3	Triweight	16.76

Table 2.6: *Elected optimal constellation of smoothing parameters and fitted degrees of freedom, $\lambda = 2\,h + 1$*

A local cubic fit is needed to capture the mortality patterns. The choices differ with the bandwidth elected. The weight function has much less effect on the bias-variance trade-off than the two other smoothing parameters. However, it influences the visual quality of the fitted regression curve.

The mortality patterns for the Dutch female population are less pronounced than for the male. A higher λ is then needed to smooth the structure at the left boundary and the accident hump which we believe less accentuated than the Male population. The corresponding fitted degrees of freedom for the female population are lower than the ones for the male, indicating that we have applied more smoothing.

Table 2.7 presents the theoretical optimal bandwidth provided by the plug-in method developed in Section 2.5.3. We fit a polynomial of degree 3 and use the corresponding optimal weight functions elected in Table 2.6. The values of λ are reported below.

	Pilot bandwidth	*Optimal* bandwidth
Dutch Male	$\lambda = 19$	$\lambda = 17$
Dutch Female	$\lambda = 32$	$\lambda = 21$

Table 2.7: *Pilot and optimal bandwidths selected by the plug-in method*

The optimal bandwidths confirmed our choices presented in Table 2.6, being relatively close and agreeing with our ranking.

2.6.3 Plots of the fits on the transformed scale

Figure 2.14 presents the mortality rates (logit scale) graduated by our local polynomials method with the optimal constellation of smoothing parameters displayed in Table 2.6.

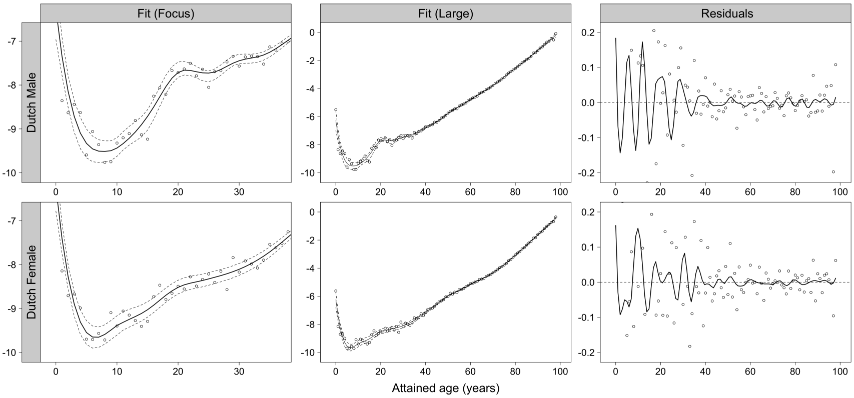


Figure 2.14: *Graduated mortality rates by local polynomial (logit scale) with* 95 % *pointwise confidence intervals and corresponding transformed residuals plots for Dutch Male and Female population, 2008. Source: HMD.*

In conjunction with the plots of the fits, we display the residuals plots.

Superimposed on the responses residuals is a *loess* smooth curve which helps for search of clusters of residuals that may indicate a lack of fit locally. This *loess* smooth curve has not detected any systematic and reproducible lack of fit. However, it shows an important lack of fit at the left boundary. Due to the underlying structure of the mortality data - high curvature at the youngest ages and a relatively linear trend after 30 years old - it is normal to get higher residuals at the left boundary than in the rest of the curve.

A last feature is shown by examining the confidence intervals in Figure 2.14. The width of the interval reveals the uncertainty associated with the graduated series. These widths are much larger for youngest ages, when the number of deaths is relatively low compared to the highest ages, as they depend on the variance of the estimates and hence on the volume of data available for graduation.

2.6.4 Plots of the smoothers

The weight function associated with the i-th point is used to compute the weights in the i-th row, $\boldsymbol{s}(x_i)$, of the 99×99 smoother $\boldsymbol{S}$ and is shown in Figures 2.15 and 2.16, below, with the influence values.

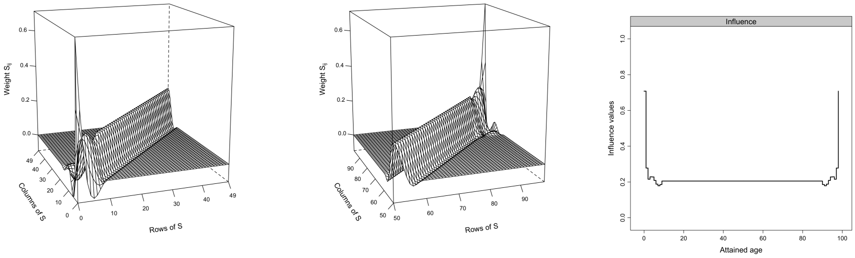


Figure 2.15: *Smoother S_{ij}: left panel: $i, j = 0, \dots, 49$, center panel: $i, j = 50, \dots, 98$ and influence values for the Dutch Male population, 2008. Source: HMD.*

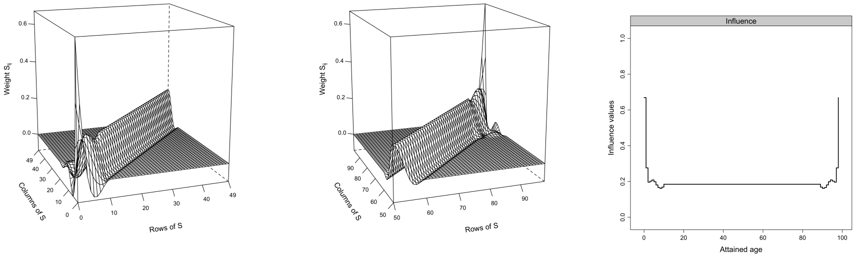


Figure 2.16: *Smoother S_{ij}: left panel: $i, j = 0, \dots, 49$, right panel: $i, j = 50, \dots, 98$ and influence values for the Dutch Female population, 2008. Source: HMD.*

The weights are shown as the height along the i-th row of the surface. For values in the central region, the weights form a triweight kernel. But as the point at which we are estimating the true curve moves towards the boundaries, the kernel overlaps the boundary and becomes asymmetric, and some weights are negative. Moreover, the height of the kernel increases because fewer observations are available.

For our applications, the boundary correcting kernel always uses λ observations wherever the target point is. For instance, for a target point at the left boundary, we use all the observations available k at the left side, and $2\ h - k$ at the right side of the point. Reciprocally for the right boundary. This type of correction is found in most smoothing software such as the `loess()` or `locfit()` functions in `R`, R Development Core Team (2012). Note that the criteria used for model selection have been computed over a restricted number of observations. Restricting the sum helps to reduce the boundaries effects, see Fan *et al.* (1998). At the boundaries, the residual sum of squares, RSC criterion and estimated derivatives can be too large because of numerical instabilities and scarcity of the data, see Section 2.5.

The influence values measure the sensitivity of the fitted curves $\widehat{\psi}(x_i)$ to the individual data points. It shows us the amount of smoothing applied locally. For instance, in Figure 2.15 right panel, $\text{infl}(x_7) = \text{infl}(x_{91}) \approx 0.18$, indicating that the observed values constitute about 18 % of the fitted values while the influence values for observations in the central region (≈ 0.21) shows that the observed values constitute about 21 % of the fitted values. It illustrates that locally we have applied more smoothing at age 7 and 91 than in the rest of the curve.

2.6.5 Plots of the graduated series and diagnostic checks

Having produced estimates on the transformed scale, we now back-transform the graduated rates. Figure 2.17 presents the mortality rates graduated on the original scale by our local polynomials method.

After graduating the crude rates and back transforming, one diagnostic mentioned by Gavin *et al.* (1995, p.183) uses the mean and variance of the binomial distribution to calculate the standardized deviation between the observed and expected deaths,

$$\frac{d_i - l_i\widehat{q}_i}{\sqrt{l_i\widehat{q}_i(1-\widehat{q}_i)}}, \quad \text{for } i = 1, \ldots, n.$$

Figure 2.18 displays the expected number of deaths with the statistic described above. We notice that most values are less than two and the statistic

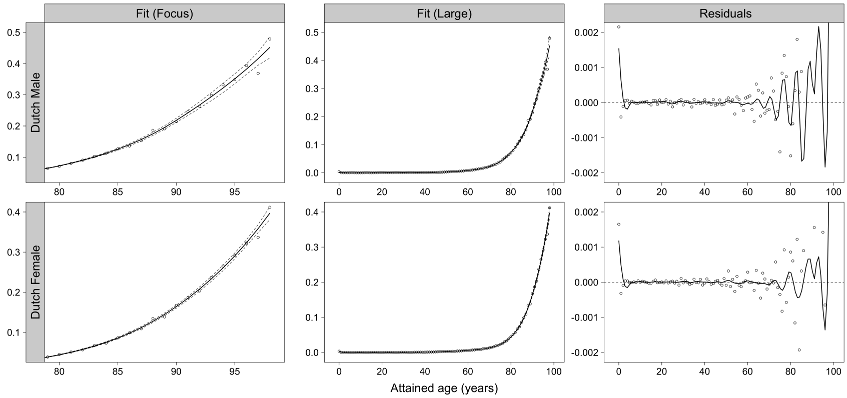


Figure 2.17: *Graduated mortality rates by local polynomials (original scale) with* 95 % *pointwise confidence intervals and corresponding residuals plots for Dutch Male and Female population, 2008. Source: HMD.*

has a mean close to zero for both populations, indicating that the assumptions made by the model are valid. Several other diagnostic plots and nonparametric tests could be considered, see Gavin *et al.* (1995) and Cleveland *et al.* (1988).

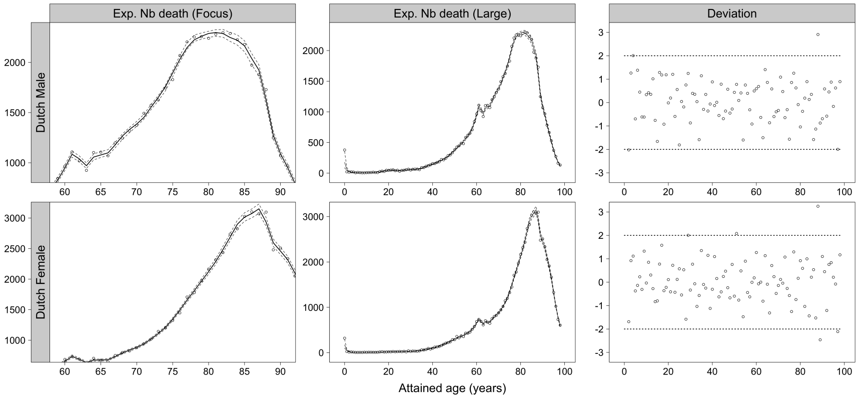


Figure 2.18: *Expected number of death with* 95 % *pointwise confidence intervals and deviation between actual and expected death for Dutch Male and Female population, 2008. Source: HMD.*

2.7 Comparisons with the Whittaker-Henderson model

Similarly to the local polynomials method, we apply the criteria presented in Section 2.5.1 to find the optimal value of parameters h and z. We picked the constellation $h = 5$ and $z = 3$ for the male, and $h = 20$ and $z = 3$ for the female population, given by Rice's T criterion, leading to 20.99 and 17.06 fitted degrees of freedom respectively. Figures 2.19 and 2.20 present graphical comparisons of the local polynomials approach and the Whittaker-Henderson model.

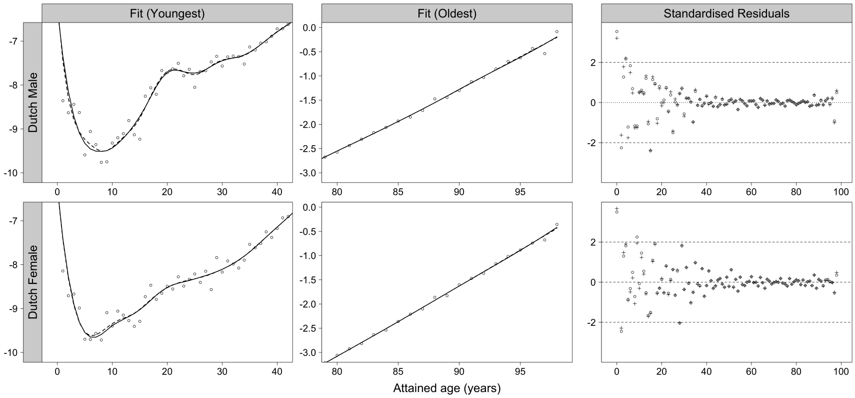


Figure 2.19: *Graphical comparisons between the local polynomials approach (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008: Graduated series and standardized residuals. Source: HMD.*

In Figure 2.19, the top left panel presents the graduated mortality rates (logit scale) for the Dutch Male population. The series graduated by local polynomials displays a smoother pattern. The corresponding degrees of freedom are lower than the ones obtained by the Whittaker-Henderson model, illustrating that the model is showing less features.

The bottom left panel shows the graduated mortality rates (logit scale) for the Dutch Female population. The graduated series are practically identical. The fitted degrees of freedom are very close, illustrating that the models show the same features.
The right panels display the standardized residuals. The circles represent the residuals from the local polynomials approach and the crosses the ones from the Whittaker-Henderson smoothing. The standardized residuals are mainly in the interval $[-2; 2]$ which indicates that the models adequately model the variability of these datasets.

In Figure 2.20, the influence values, obtained by the local polynomials for the male population, up to age 80 are below the ones computed with Whittaker-Henderson model, $\text{infl}_{WH}(x_i) = diag((\boldsymbol{V} + hK_z^T K_z)^{-1}\boldsymbol{V})$, top

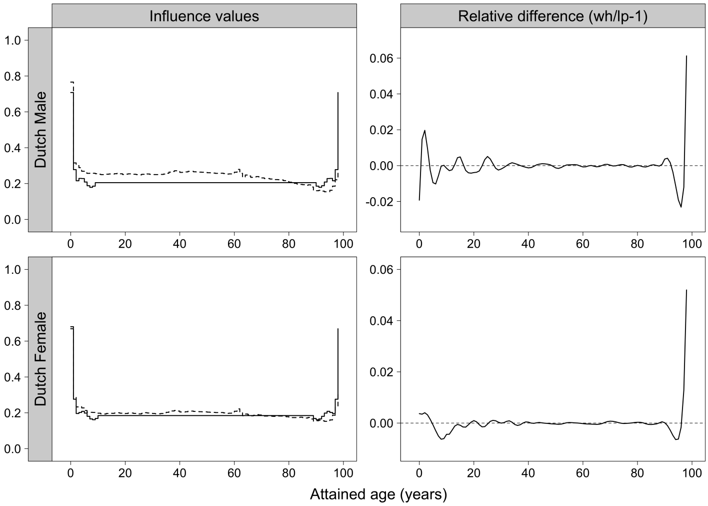


Figure 2.20: *Graphical comparisons between the local polynomials approach (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008: Influence values and relative difference between the graduated series. Source: HMD.*

left panel. It indicates that, up to age 80, more smoothing has been applied by the local polynomials approach. For instance, $\text{infl}_{LP}(x_{20}) \approx 0.21$, indicating that the observed value constitutes about 21 % of the fitted value, while the influence value obtained by the Whittaker-Henderson model for the same observation ($\text{infl}_{WH}(x_{20}) \approx 0.26$) shows that the observed value constitutes about 26 % of the fitted value.
The relative difference between the two approaches for the male population is more important at the boundaries, where the Whittaker-Henderson model does not need special treatment.
The influence values for the female population, bottom left panel, stay close. The relative difference is very low and, as for the male population, is larger in the boundaries.

We end the comparisons by applying the tests proposed by Forfar *et al.* (1988, p.56-58) and Debón *et al.* (2006, p.231). We have also obtained the values of the mean absolute percentage error $MAPE$ and R^2 used in Felipe *et al.* (2002). We compare the crude mortality rates to the graduated series to see whether the two approaches lead to similar graduation. Table 2.8 presents the results.

The two approaches display favorable results making it difficult to choose one of them. As an advantage for the Whittaker-Henderson method, we observe that is not necessary to give a special treatment to the observations

		Local Polynomial		Whittaker-Henderson	
		Male	Female	Male	Female
Degree of freedom		$18,46$	$16,76$	$20,99$	$17,06$
Computation time	(sec)	$0,857$	$0,860$	$0,008$	$0,008$
Standardized	> 2	5	5	4	4
Residuals	> 3	2	2	2	2
Signs	$+(-)$	54(45)	48(51)	51(48)	48(51)
Test	p-value	0.4215	0.8408	0.8408	0.8408
Runs	Nb of runs	59	67	59	63
Test	Value	1.8152	3.3460	1.7281	2.5371
	p-value	0.0695	0.0082	0.0840	0.0112
Kolmogorov	Value	0.0303	0.0404	0.0303	0.0404
Smirnov test	p-value	1	1	1	1
$\mathcal{X}^2$	Value	129.06	93.15	103.39	94.62
Test	p-value	0.0194	0.6196	0.3352	0.5779
R^2	Value	0.9983	0.9986	0.9985	0.9986
MAPE	(%)	10.41	9.61	9.05	8.99

Table 2.8: *Comparisons between the local polynomials approach and the Whittaker-Henderson smoothing for the Dutch Male and Female population, 2008. Source: HMD.*

in the boundary, and the computation time is 100 times smaller. However we have used a prototype implementation in R to perform the the local polynomials approach. This can be improved by at least a factor of 10, if a lower level language such as C is used.

2.8 Summary and outlook

This chapter gives an extensive overview of local regression techniques. Local regression is a popular form of non-parametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. It is very adaptable, and it is also convenient statistically since a lot is known about least squares theory, which is helpful when looking at bias and variance.

We have seen how local polynomial regression can be used to model the relation between the crude death rates and attained age with sufficient exposures. However, for the purpose of graduating series originating from life insurance, the transformation of the data is a real problem for two reasons. On one hand, due to the transformation, a high curvature appears in the left boundary. As a consequence, the selection of the smoothing parameters may be mainly driven by minimizing the residual sum of squares in the boundaries rather than for the whole set of data points. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table.

On the other hand when the volume of data is not sufficiently high, the datasets might present zero response for youngest and oldest ages and hence the logit transform can not be applied. We should point out that many authors achieve better fits by eliminating the early ages due to their irregular profile, which they justify by arguing that actuarial operations begin at more advanced age. We have chosen to include the young age groups to show the applicability and relevance of the approach to find structure in the presence of an irregular profile. Moreover, it is worth remembering that the double exponential, which appears in Heligman and Pollard (1980) and is related to parametric models, has been introduced to deal specifically with the difficulty of adjusting the younger ages.

Finally, it would be desirable to model situations where a non-Gaussian likelihood is appropriate. In local polynomial regressions, the response variable was assumed to be approximately Gaussian. If the response is binary or given by counts, the technique considered there is no longer applicable, because binary or count data have an expectation-variance structure that is different from the continuous, normally distributed responses. In the following chapter, the concepts of Sections 2.2 and 2.4 are incorporated and extended within the framework of local likelihood and localized Generalized Linear Models.

Chapter 3

Local likelihood approaches

This chapter is based on Tomas (2011). A local likelihood approach to univariate graduation of mortality, *Bulletin Français d'Actuariat*, **11**(22), pages 105-153.

3.1 Introduction

We discusse a simple extension of the local fitting technique presented in Chapter 2. We extend smoothing ideas to other kinds of data. In particular, data of which the relationship can be expressed through a likelihood function. If the response is binary or given by counts, the technique considered in the previous chapter is no longer applicable, because binary or count data have an expectation-variance structure that is different from the continuous, normally distributed responses.

Local kernel-weighted log-likelihood is introduced as a method of smoothing by local polynomials in non-Gaussian regression models. In the following, we incorporate and extend the concepts of the non-parametric regression technique of local polynomials within the framework of local likelihood and localized Generalized Linear Models.

In the last three decades, the use of Generalized Linear Models (GLMs) in actuarial statistics has received a lot of attention, starting with the applications of McCullagh and Nelder (1989). First, regression is no longer restricted to normal data, but extended to distribution from the exponential family. This allows appropriate modeling for frequency counts (number of deaths) and binary data (mortality rates). Second, a GLM models the effect of explanatory variables on a transformation of the mean instead of the mean itself. Third, the distribution of error-terms may be non-normal and heteroskedastic, having a variance that depends on its mean.

The chapter is organized as follows. Section 3.2 extends the theory of GLMs to local kernel-weighted likelihood and local GLMs. The statistical properties are covered in Section 3.3 and model diagnostics are discussed in

Section 3.4. We cover the graduation of both the probability of death and the force of mortality over the entire age range involving historic data from the Netherlands. We present a local binomial likelihood model when the number of initial policyholders exposed to the risk is available in Section 3.5, and Section 3.6 develops a local Poisson likelihood model when the number of central policyholders exposed to the risk is available. The initial exposed to risk is the number of individuals alive aged x at the start of the period of observation, while the central exposed to risk is the time exposed to risk of dying at age x. We provide comparisons with the Whittaker-Henderson model for the two approaches. Finally Section 3.7 summarizes the conclusions drawn in this chapter.

Related work is in Delwarde *et al.* (2004) and Debón *et al.* (2006), but our work examines the statistical properties of the estimators and the choice of the smoothing parameters by classical selectors as well as the plug-in methodology. In addition we provide a method for constructing pointwise confidence intervals that are not depending on the estimates using the variance stabilizing link. This method allows us to obtain confidence intervals in presence of zero-responses. The implementation of optimization algorithm is straightforward in standard statistical software such as R, R Development Core Team (2012). The basic idea is a simple extension of the local fitting technique presented in Chapter 2. We extend smoothing ideas to other kinds of data. In particular, data of which the relationship can be expressed through a likelihood function.

Suppose that we have n independent realizations $y_1, y_2, \ldots, y_n$ of the random variable Y with

$$Y_i \sim f(Y|\theta(x_i)), \quad \text{for } i = 1, 2, \ldots, n,$$

where $f(\cdot|\theta(x_i))$ is a probability mass/density function in the exponential dispersion family and $\theta(x_i)$ is called the natural parameter in the GLMs framework. The likelihood is given by

$$\mathcal{L}(\theta_1, \theta_2, \ldots, \theta_n) = \prod_{1}^{n} f(y_i, \theta_i).$$

A standard modeling would assume a simple parametric form for the $\theta(x_i)$'s, for instance $\theta(x_i) = \beta_0 + \beta_1 x_i$. Following the approach taken by Tibshirani and Hastie (1987) we enlarge this class by replacing the parsimonious covariate form with an unspecified smooth function $\psi(x_i)$: $\theta(x_i) = \psi(x_i)$. To estimate $\{\psi(x_1), \psi(x_2), \ldots, \psi(x_n)\}$, we could try to maximize $\mathcal{L}(\psi(x_1), \psi(x_2), \ldots, \psi(x_n))$. However, this would result in an unsatisfactory estimate due to over-fitting. It would simply reproduce the data. As an alternative, similarly to Section 2.2.1, we suppose that the function ψ has a $(p+1)$st continuous derivative at the point x_i. For data point x_j in a neighborhood of x_i we approximate $\psi(x_j)$ via a Taylor expansion by a polynomial of degree p:

$$\psi(x_j) \approx \psi(x_i) + \psi'(x_i)(x_j - x_i) + \ldots + \frac{\psi^{(p)}(x_i)}{p!}(x_j - x_i)^p \equiv \boldsymbol{x}^T \boldsymbol{\beta},$$

where $\boldsymbol{x} = (1, x_j - x_i, \dots, (x_j - x_i)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$, with $\beta_v = \psi^{(v)}(x_i)/v!$, $v = 0, 1, \dots, p$.

The contribution to the log-likelihood, for data points (x_j, y_j) in the neighborhood of x_i, is denoted by $l\left(y_j, \boldsymbol{x}^T\boldsymbol{\beta}\right)$. In addition it is weighted by w_j, where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$, and

$$w_j = \begin{cases} W(|x_j - x_i|/h) & \text{if } |x_j - x_i|/h \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $W(.)$ is one of the weight functions presented in Table 2.1 and $h = (\lambda - 1)/2$, λ being the window width.

It leads to the local log-likelihood, or *local kernel-weighted log-likelihood* as named by Fan *et al.* (1998):

$$L(\boldsymbol{\beta}|\lambda, x_i) = \sum_{j=1}^{n} l\left(y_j, \boldsymbol{x}^T\boldsymbol{\beta}\right) w_j. \tag{3.1}$$

Maximizing the local log-likelihood (3.1) with respect to $\boldsymbol{\beta}$ gives the vector of estimators $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \dots, \widehat{\beta}_p)^T$. Estimators $\psi^{(v)}(x_i)$, $v = 0, 1, \dots, p$, are given by

$$\widehat{\psi}^{(v)}(x_i) = v!\widehat{\beta}_v. \tag{3.2}$$

3.2 The local likelihood model

3.2.1 Localizing generalized linear models

A special case of model (3.1) occurs when the conditional density of Y given X belongs to the exponential dispersion family with a probability mass function which can be written in the form:

$$f_Y(y_j; \theta_j, \phi) = \exp\left\{\frac{y_j\theta_j - b(\theta_j, m_j)}{a_j(\phi)} + c(y_j, \phi)\right\},$$

for specific functions $a()$, $b()$ and $c()$ and where ϕ is called the dispersion parameter. It is a nuisance parameter not depending on x_j. If ϕ is known, then we call it an exponential family model with canonical parameter θ. If ϕ is unknown we have a two-parameter exponential family but the estimation procedure is unchanged because the local score function for θ does not involve $a_j(\phi)$. The functions a and c are such that $a_j(\phi) = \phi/m_j$ and $c = c(y_i, \phi/m_j)$, where m_j is a known weight for each observation x_j. The most important examples for our purposes are presented in Table 3.1.

One of our goals throughout this chapter is to clarify and demonstrate how the families, links and variations fit together in an understandable

Distribution of y_j	θ_j	m_j	$a_j(\phi)$	$b(\theta_j, m_j)$	$c(y_j, \phi)$
Normal(μ_j;σ^2)	μ_j	1	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\}$
Poisson(μ_j)	$\log(\mu_j)$	1	1	$\exp(\theta_j)$	$-\log y!$
Binomial(l_j; q_j)	$\log\left(\frac{q_j}{1-q_j}\right)$	l_j	$\frac{1}{l_j}$	$l_j \log(1 + \exp\theta_j)$	$\log\binom{l_j}{L_j d_j}$

Table 3.1: *Distributions of interest belonging to the Exponential Dispersion Family.*

framework. A local binomial likelihood model is used when the number of initial policyholders exposed to risk is available, and hence the graduated probabilities of death are given by $\widehat{\eta}(x_i)$, the linear predictor in the GLMs framework; while for those central exposed to risk, a local Poisson model is used and the graduated forces of mortality are derived as $\widehat{\mu}(x_i)/l_i$. The initial exposed to risk is the number of individuals alive aged x_i at the start of the period of observation, while the central exposed to risk is the time exposed to risk of dying at age x_i.

The unknown function $\mu(x_j) = \mathbb{E}[Y|X = x_j]$ is modeled in X by a link function $g(.)$ such as $g(\mu(x_j)/m_j) = \eta(x_i)$. Then $\mathbb{E}[Y_i]$ is tied to a linear combination,

$$\sum_{j=1}^{n} w_j m_j \sum_{p=0}^{n} \beta_p (x_j - x_i)^p,$$

of the parameters β, by a monotonous and differentiable function $g(.)$, not necessarily the identity. We proceed by forming the local likelihood as in (3.1) and estimate the $\boldsymbol{\beta}$.

This procedure can be viewed as an extension of the family of generalized linear models (GLMs), see Nelder and Wedderburn (1972) and McCullagh and Nelder (1989).

Extensive experience on graduation using GLMs has been built up in the actuarial literature with Renshaw (1991) and reviewed by Haberman and Renshaw (1996). We invite the reader to look at Kaas *et al.* (2008, Chap. 9-11) for a clear presentation about the use of GLMs in actuarial science. Local likelihood methods to graduate of mortality tables have been applied in Delwarde *et al.* (2004), Debón *et al.* (2006) and more recently in Gschlössl *et al.* (2011).

The GLMs provide a generalization of linear regression to likelihood models. Regression is no longer restricted to normal data, but extended to distributions from the exponential family. This allows appropriate modeling for frequency counts (number of deaths) and binary data (mortality rates). Also, a GLM models the effect of explanatory variables on a transformation of the mean instead of the mean itself.

The role of GLMs is that of a background model which is fitted locally. In a parametric generalized linear model, $\eta(x_i) = \beta_0 + \beta_1 x_i$ for some unknown parameter β_0 and β_1. In our non-parametric setting, there is no model assumption about $\eta(x_i)$. The primary goal is to estimate $\mu(x_i)/m_i$, or equivalently $\eta(x_i)$, non-parametrically, that is, $\beta_0 + \beta_1 x_i$ is generalized to $\psi(x_i)$. The obvious extension of this idea is to suppose that $\eta(x_i)$ is a pth degree polynomial in x_j, with x_j being an element of the neighborhood of x_i.

Therefore, fitting procedures that are familiar from GLMs are needed, but, of course, the modeling itself is smooth and no longer parametric.

3.2.2 The choice of the link function

The function $g(.)$ is called the link function, and it is assumed to be known. In parametric regression models, the choice of the link function is largely dictated by the data. If the true mean is log-linear, one has to use the log link. With local regression models, one does not assume the model is globally correct, so the choice of the link can be driven by convenience. This choice has a relatively small impact on the graduated series compared to the choice of the smoothing parameters. Hence, it can be driven by practical considerations, which could be the ease of computations or the construction of the confidence interval in the presence of zero responses. It is also conceivable to dispense with the link function and just estimate $\mu(x_i)$ directly. But there are several drawbacks to having the link equal to the identity. An identity link may lead to a non-convex likelihood, allowing for the possibility of multiple maxima, inconsistency and computational problems. The use of a canonical link guarantees convexity, see Fan *et al.* (1995). Furthermore, it ensures that the final estimate is in the correct range. A final reason is that the estimate $\widehat{\mu}(x_i)$ approaches the usual parametric estimate as the bandwidth becomes large.

For our purpose, we could use the canonical link. The canonical link is $\theta = g(\mu/m)$. When a local polynomial is used for $\theta(x_i)$, the local log-likelihood $L(\boldsymbol{\beta}; \lambda, x_i)$ (and hence $\widehat{\theta}(x_i)$) depends on the data only through

$$\sum_{j=1}^{n} w_j m_j \sum_{p=0}^{n} \beta_p (x_j - x_i)^p y_j .$$

This locally sufficient statistic simplifies theoretical computations. Each of the discussed distributions has a special link function for which there exists a sufficient statistic. Examples are presented in the Table 3.2 below.

Error	Canonical link
Normal	$\eta = \mu$
Poisson	$\eta = \log(\mu)$
Binomial	$\eta = \log((\mu/m)/(1 - \mu/m))$

Table 3.2: *Examples of canonical links*

However, an important result associated to the graduated series is the construction of the corresponding confidence intervals. For likelihood models, confidence intervals should ideally take into account the underlying family of distributions. But the theory for deriving such intervals is quite intractable. Hence, following the approach taken by Loader (1999b, p.171), we would rely on a method based on normal approximation presented in Section 3.3.2.

A problem that occurs with likelihood models is that $\mathbb{V}\mathrm{ar}\left[\widehat{\theta}(x_i)\right]$ usually depends on the unknown parameter $\theta(x_i)$, and simply substituting an estimate may not be satisfactory if we happen to observe $y = 0$. Using the logistic link function does not help since then $\widehat{\theta} = -\infty$ and the variance is also infinite. The simple solution, within the framework of normal approximations is to use the variance stabilizing link. Under this link, the variance of $\widehat{\theta}(x_i)$ is independent of the true parameter $\theta(x_i)$, at least asymptotically. It leads to confidence intervals whose widths depend only on the design points x_i, see Sections 3.5.3 and 3.6.3. Examples of variance stabilizing links are indicated in Table 3.3 below.

Error	Variance stabilizing link
Poisson	$g(\mu) = \sqrt{\mu}$
Binomial	$g(\mu/m) = \sin^{-1}(\sqrt{\mu/m})$

Table 3.3: *Examples of variance stabilizing link functions*

3.2.3 Local likelihood equations

In practice, the coefficients $\boldsymbol{\beta} = \beta_0, \ldots, \beta_p$ are unknown and have to be estimated based on data in the neighborhood of the target point x_i.
In the following, we focus on the estimation of the $\boldsymbol{\beta}$ by maximum likelihood. It consists of maximizing the local log-likelihood

$$\begin{aligned} L\left(\boldsymbol{\beta}|\boldsymbol{y}, w_j, \phi, m_j\right) &= \sum_{j=1}^{n} w_j \; l\left(y_j|\theta_j, \phi, m_j\right) \qquad (3.3) \\ &= \sum_{j=1}^{n} w_j \log f_Y\left(y_j|\theta_j, \phi, m_j\right) \\ &= \sum_{j=1}^{n} w_j \frac{y_j\; \theta_j - b(\theta_i, m_j)}{\phi/m_j} + \sum_{j=1}^{n} w_j \; c(y_j, \phi/m_j), \end{aligned}$$

where $\mathbb{E}[Y_j] = b'(\theta_j, m_j) = \mu_j$ and $g(\mu_j/m_j) = \sum_{p=0}^{n} \beta_p (x_j - x_i)^p = \eta_j$, with $g(.)$ denoting the link function.
Since we want to maximize the log likelihood for $\beta_0, \beta_1, \ldots, \beta_p$ we look for

a solution of the set of normal equations to be fulfilled by the maximum likelihood parameter estimates $\boldsymbol{\beta}$:

$$A_v = 0 \quad \text{for} \quad v = 0, 1, \ldots, p,$$

where

$$\begin{aligned} A_v &= \frac{\partial L\left(\beta_v | \boldsymbol{y}, w_j, \phi, m_j\right)}{\partial \beta_v} \\ &= \sum_{j=1}^{n} w_j \frac{\partial \log f_Y\left(y_j | \theta_j, \phi, m_j\right)}{\partial \beta_v} \\ &= \sum_{j=1}^{n} w_j \frac{\partial}{\partial \beta_v}\left(\frac{y_j\ \theta_j - b(\theta_j, m_j)}{\phi/m_j} + c(y_j, \phi/m_j)\right). \end{aligned}$$

To obtain A_v, we apply the chain rule to the log likelihood

$$\frac{\partial \log f_Y\left(y_j | \theta_j, \phi, m_j\right)}{\partial \beta_v} = \frac{\partial \log f_Y\left(y_j | \theta_j, \phi, m_j\right)}{\partial \theta_j} \frac{\partial \theta_j}{\partial \mu_j} \frac{\partial \mu_j}{\partial \eta_j} \frac{\partial \eta_j}{\partial \beta_v}.$$

As $\mu_j = b'(\theta_j, m_j)$, this leads to

$$\begin{aligned} \frac{\partial \log f_Y\left(y_j | \theta_j, m_j, \phi\right)}{\partial \theta_j} &= \frac{y_j - b'(\theta_j, m_j)}{\phi/m_j} = \frac{y_j - \mu_j}{\phi/m_j}, \\ \frac{\partial \mu_j}{\partial \theta_j} &= b''(\theta_j, m_j), \end{aligned}$$

and

$$\frac{\partial \eta_j}{\partial \beta_v} = (x_j - x_i)^v.$$

Hence, we obtain

$$\frac{\partial \log f_Y(y_j | \theta_j, \phi, m_j)}{\partial \beta_v} = \frac{(y_j - \mu_j)\ (x_j - x_i)^v}{(\phi/m_j)\ b''(\theta_j, m_j)} \frac{\partial \mu_j}{\partial \eta_j}.$$

The link function $\eta = g(\mu/m)$ determines

$$\partial \mu_j / \partial \eta_j = \partial g^{-1}(\eta_j) / \partial \eta_j = 1/g'(\mu_j/m_j).$$

So finally,

$$A_v = \sum_{j=1}^{n} w_j\ \frac{(y_j - \mu_j)\ (x_j - x_i)^v}{b''(\theta_j, m_j)(\phi/m_j)g'(\mu_j/m_j)}.$$

Hence,

$$A_v = 0 \Leftrightarrow \frac{1}{\phi} \sum_{j=1}^{n} w_j\ m_j\ (y_j - \mu_j) \frac{(x_j - x_i)^v}{b''(\theta_j, m_j) g'(\mu_j/m_j)} = 0. \qquad (3.4)$$

Likewise, in matrix notation, the local likelihood equations can be written as

$$\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\ \boldsymbol{V}(\boldsymbol{y}-\boldsymbol{\mu})=0, \tag{3.5}$$

where

$$\boldsymbol{X}=\begin{bmatrix}1 & x_1-x_i & (x_1-x_i)^2 & \dots & (x_1-x_i)^P\\ 1 & x_2-x_i & (x_2-x_i)^2 & \dots & (x_2-x_i)^P\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n-x_i & (x_n-x_i)^2 & \dots & (x_n-x_i)^P\end{bmatrix},$$

and $\boldsymbol{V}$ is a diagonal matrix with elements

$$v_{jj}=\frac{m_j}{b''(\theta_j,m_j)}\frac{\partial\mu_j}{\partial\eta_j}. \tag{3.6}$$

If the canonical link is chosen, the local likelihood equations become

$$\frac{1}{\phi}\sum_{j=1}^{n}w_j m_j(y_j-\mu_j)\ (x_j-x_i)^v=0.$$

3.2.4 Fisher's scoring method

These equations are usually non-linear, and so the solution must be obtained through iterative methods. One way to solve those is to use Newton-Raphson iterations. We note $\boldsymbol{A}(\boldsymbol{\beta})$ the gradient vector of the log-likelihood; from which the vth component is

$$A_v(\boldsymbol{\beta})=\frac{\partial}{\partial\beta_v}\ L(\boldsymbol{\beta}|\boldsymbol{y}),$$

and we denote by $\boldsymbol{H}(\boldsymbol{\beta})$ the Hessian matrix of $l(\boldsymbol{\beta}|\boldsymbol{y})$, i.e., the one of which element (v,k) is

$$\frac{\partial^2}{\beta_v\beta_k}\ L(\boldsymbol{\beta}|\boldsymbol{y}).$$

For $\boldsymbol{\beta}$ close to $\boldsymbol{\beta}^*$, using a linear approximation, we have

$$0=\boldsymbol{A}(\boldsymbol{\beta}^*)\approx\boldsymbol{A}(\boldsymbol{\beta})+\boldsymbol{H}(\boldsymbol{\beta})(\boldsymbol{\beta}^*-\boldsymbol{\beta}),$$

which leads to

$$\boldsymbol{A}(\boldsymbol{\beta})+\boldsymbol{H}(\boldsymbol{\beta})(\boldsymbol{\beta}^*-\boldsymbol{\beta})\approx 0,$$

or

$$\boldsymbol{\beta}^*\approx\boldsymbol{\beta}-\boldsymbol{H}^{-1}(\boldsymbol{\beta})\ \boldsymbol{A}(\boldsymbol{\beta}). \tag{3.7}$$

The algorithm of Nelder and Wedderburn replaces the Hessian by its expected value. It uses the information matrix. The technique that arises in this way is called the Fisher's scoring method:

$$\boldsymbol{H}(\boldsymbol{\beta})\approx\mathbb{E}\big[\boldsymbol{H}(\boldsymbol{\beta})\big]=-\mathcal{I}(\boldsymbol{\beta}).$$

Hence an alternative to (3.7) is

$$\boldsymbol{\beta}^* \approx \boldsymbol{\beta} + \mathcal{I}^{-1}(\boldsymbol{\beta})\ \boldsymbol{A}(\boldsymbol{\beta}). \tag{3.8}$$

Note that $\mathbb{E}\big[\boldsymbol{H}\big] = -\mathbb{E}\big[\boldsymbol{A}\boldsymbol{A}^T\big]$. In terms of quantity of information, if $\boldsymbol{H}$, and hence $-\mathcal{I}$, is small, the likelihood will have a slight curvature, and the determination of the maximum likelihood estimate will be less trivial.
The element (v,k) of the Fisher information matrix $\mathcal{I}$ is given by $\mathcal{I}_{vk} = \mathbb{E}\big[\boldsymbol{A}_v\boldsymbol{A}_k\big]$:

$$\begin{aligned}\mathcal{I}_{vk} &= \mathbb{E}\Bigg[\sum_{j=1}^{n} w_j \frac{\partial \log f_Y(y_j)}{\partial \beta_v} \sum_{l=1}^{n} w_l \frac{\partial \log f_Y(y_l)}{\partial \beta_k}\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{j=1}^{n} w_j \frac{(y_j-\mu_j)(x_j-x_i)^v(x_j-x_i)^k}{b''(\theta_j,m_j)\ \phi/m_j}\left(\frac{\partial \mu_j}{\partial \eta_j}\right) \\ &\qquad \times \sum_{l=1}^{n} w_l \frac{(y_l-\mu_l)(x_l-x_i)^v(x_l-x_i)^l}{b''(\theta_l,m_l)\ \phi/m_l}\left(\frac{\partial \mu_l}{\partial \eta_l}\right)\Bigg].\end{aligned} \tag{3.9}$$

Note that

$$\mathbb{E}\big[(y_j-\mu_j)(y_l-\mu_l)\big] = \mathbb{C}\mathrm{ov}\big[y_j,y_l\big] = 0 \quad \text{for} \quad j \neq l,$$

as we supposed the observations independent. For $j = l$, we obtain

$$\mathbb{E}\left[(y_j-\mu_j)^2\right] = \mathbb{V}\mathrm{ar}\big[y_j\big].$$

Since $\mathbb{V}\mathrm{ar}\big[y_j\big] \equiv b''(\theta_j,m_j)$, we obtain

$$\begin{aligned}\mathcal{I}_{vk} &= \frac{1}{\phi}\sum_{j=1}^{n} w_j\ \frac{\mathbb{V}\mathrm{ar}\big[y_j\big]}{(b''(\theta_j,m_j)\phi/m_j)^2}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2 (x_j-x_i)^v(x_j-x_i)^k \\ &= \frac{1}{\phi}\sum_{j=1}^{n} w_j\ \omega_{jj}(x_j-x_i)^v(x_j-x_i)^k \\ &= \frac{1}{\phi}\Big\{\boldsymbol{X}^T\boldsymbol{W}\ \boldsymbol{\Omega}\boldsymbol{X}\Big\}_{vk},\end{aligned} \tag{3.10}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with elements

$$\omega_{jj} = \frac{m_j^2}{b''(\theta_j,m_j)}\left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2, \tag{3.11}$$

depending on the variance and link function. Since $\eta_j = g(\mu_j)$, we have $\partial\eta_j/\partial\mu_j = g'(\mu_j)$, hence in using the canonical link, ω_{jj} reduces to μ_j.

Using those weights $v_{jj} = (\omega_{jj}/m_j)(\partial\eta_j/\partial\mu_j)$ the local likelihood equations

(3.4) become

$$\begin{aligned}\frac{\partial}{\partial \beta_v}\, L(\boldsymbol{\beta}|\boldsymbol{y}) &= \frac{1}{\phi}\sum_{j=1}^{n} w_j\, \omega_{jj}\, (x_j - x_i)^v \frac{y_j - \mu_j}{m_j}\, g'(\mu_j/m_j)\\ &= \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\, \boldsymbol{u},\end{aligned}$$

where $u_j = \frac{y_j - \mu_j}{m_j} g'(\mu_j/m_j)$. Then by (3.8),

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \mathcal{I}^{-1}\frac{\partial}{\partial \beta_v}\, L(\boldsymbol{\beta}|\boldsymbol{y}),$$

or equivalently, $\mathcal{I}\,(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = \frac{\partial}{\partial \beta_v}\, L(\boldsymbol{\beta}|\boldsymbol{y})$.
Let $\widehat{\eta}_j$ and $\widehat{\mu}_j$ be the vector of linear predictors and fitted values when the parameter vector equals $\boldsymbol{\beta}$, so

$$\widehat{\eta}_j = \sum_{p=0}^{n} \beta_p (x_j - x_i)^p \qquad \text{and} \qquad \widehat{\mu}_j = m_j g^{-1}(\widehat{\eta}_j).$$

Then by (3.10),

$$\mathcal{I}\boldsymbol{\beta} = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\, \boldsymbol{\Omega}\, \boldsymbol{X}\, \boldsymbol{\beta} = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\, \boldsymbol{\Omega}\, \widehat{\eta}_j.$$

So we can rewrite the Fisher scoring iteration equation as

$$\mathcal{I}\boldsymbol{\beta}^* = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\, \boldsymbol{\Omega}\, \boldsymbol{z}$$

where

$$z_j = \widehat{\eta}_j + \frac{y_j - \widehat{\mu}_j}{m_j} g'(\widehat{\mu}_j/m_j). \tag{3.12}$$

The elements of $\boldsymbol{z}$ are called the working dependent variables.

Hence, a maximum likelihood estimate of $\boldsymbol{\beta}$ is found by the following iterative process:

> Repeat $\boldsymbol{\beta}^* := \left(\boldsymbol{X}^T\boldsymbol{W}\, \boldsymbol{\Omega}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\, \boldsymbol{\Omega}\, \boldsymbol{z}$;
>
> using $\boldsymbol{\beta}^*$, update the working weights $\boldsymbol{\Omega}$, as well as the working dependent variables $\boldsymbol{z}$ until convergence.

Estimation of $\boldsymbol{\beta}$ is performed using a Fisher's scoring method search in each neighborhood, going in order as i runs from 1 to n.
Note that simplification occurs for the canonical links where the expected value and the actual value of the Hessian matrix coincide. The Fisher's scoring method and the Newton-Raphson method thus reduce to the same algorithm.

Since the local likelihood estimate does not have an explicit representation, statistical properties cannot be derived as easily as in the local regression case. But a Taylor series expansion of the local likelihood gives an approximate linearization of the estimate, leading to theory parallel to that developed in Section 2.4 for local polynomial regression.

The generalization to multiple predictors is similar to Section 2.2.2. The derivations are presented briefly in Section 4.4.2.

3.3 Statistical properties

3.3.1 Assessment of bias and variance

We focus on how to estimate the bias and variance of the local likelihood estimate. The estimated bias and variance will be used to construct confidence intervals in Section 3.3.2. Due to the nonlinear definition of $\widehat{\boldsymbol{\beta}}$, it is not possible to derive exact means and variances of $\widehat{\boldsymbol{\beta}}$. We now provide an estimate for the bias and variance based on the same idea introduced in Section 2.4.1 and extended for local likelihood in Fan *et al.* (1998, p.594-597). The bias assessment relies on the difference of two maximum likelihood fits with different accuracies. Recall the bias of the estimator $\widehat{\boldsymbol{\beta}}$ comes from an approximation error in the Taylor expansion. Let

$$r(x_j) = \psi(x_j) - \sum_{v=1}^{p} \psi^{(v)}(x_i)(x_j - x_i)^v / j!$$

denote the approximation error at the point x_j. Suppose that the $(p+a+1)$st derivative of ψ exists at the point x_i for some $a > 0$. A further expansion of $\psi(x_j)$ gives then an approximation of the approximation error

$$r(x_j) \approx \beta_{p+1}(x_j - x_i)^{p+1} + \ldots + \beta_{p+a}(x_j - x_i)^{p+a} \equiv r_j, \tag{3.13}$$

where a denotes the order of approximation. Again for practical implementation, we have chosen $a = 2$.
Suppose for a moment that the quantities r_j are known. Then a more precise local log-likelihood is

$$L^\circ(\boldsymbol{\beta}) = \sum_{j=1}^{n} w_j l(y_j, \boldsymbol{x}^T\boldsymbol{\beta} + r_j). \tag{3.14}$$

Let $\widehat{\boldsymbol{\beta}}^\circ$ denote the maximizer of the local log-likelihood $L^\circ(\boldsymbol{\beta})$. The bias of $\widehat{\boldsymbol{\beta}}$ can then be estimated as $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^\circ$. Let $\boldsymbol{A}^\circ(\boldsymbol{\beta})$ and $\boldsymbol{H}^\circ(\boldsymbol{\beta})$ denote the gradient vector and the hessian matrix of the local log-likelihood $L^\circ(\boldsymbol{\beta})$, respectively. Since $\widehat{\boldsymbol{\beta}}^\circ$ is the maximizer of $L^\circ(\boldsymbol{\beta})$, a linear approximation gives

$$0 = \boldsymbol{A}^\circ(\widehat{\boldsymbol{\beta}}^\circ) \approx \boldsymbol{A}^\circ(\widehat{\boldsymbol{\beta}}) + \boldsymbol{H}^\circ(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^\circ - \widehat{\boldsymbol{\beta}}),$$

and we obtain the estimated bias vector

$$\begin{aligned} \widehat{\boldsymbol{b}}(\boldsymbol{\beta}) &= \boldsymbol{H}^\circ(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{A}^\circ(\widehat{\boldsymbol{\beta}}) \\ &= \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{V}\boldsymbol{r}. \end{aligned} \tag{3.15}$$

Therefore, the estimated bias of the linear predictor or equivalently of $\psi(x_i)$ is given by

$$\widehat{\boldsymbol{b}}(\eta_i) = \boldsymbol{e}_1^T \big(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Omega} \boldsymbol{X}\big)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{V} \boldsymbol{r}. \tag{3.16}$$

The approximated bias (3.15) depends on quantities r_j that are unknown. These quantities will be estimated by fitting locally a polynomial of degree $p+a$ via equation (3.3), using a pilot bandwidth h°. This gives estimates $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_{p+a})^T$, which are substituted into expression (3.13), leading to the estimates $\widehat{r}_j$ of r_j. These estimates are then substituted into (3.14), yielding the estimated bias as in (3.15). The choice of the pilot bandwidth h° will be discussed in Section 3.4.2.

To obtain the variance, using a linear approximation, we have

$$0 = \boldsymbol{A}(\widehat{\boldsymbol{\beta}}) \approx \boldsymbol{A}(\boldsymbol{\beta}) + \boldsymbol{H}(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

This leads to

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\boldsymbol{H}(\boldsymbol{\beta})^{-1} \boldsymbol{A}(\boldsymbol{\beta}),$$

and an approximation of the variance is

$$\begin{aligned} \mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big] &\approx \mathbb{E}\big[-\boldsymbol{H}(\boldsymbol{\beta})^{-1}\big]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \mathbb{E}\big[\boldsymbol{A}(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})^T\big]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \mathbb{E}\big[-\boldsymbol{H}(\boldsymbol{\beta})^{-1}\big]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \\ &= \mathcal{I}_{\widehat{\boldsymbol{\beta}}}^{-1} \left(\sum_{j=1}^{n} w_j \mathbb{E}\left[\frac{\partial \log f_Y(y_j)}{\partial \boldsymbol{\beta}^T} \frac{\partial \log f_Y(y_j)}{\partial \boldsymbol{\beta}} \right] \right) \mathcal{I}_{\widehat{\boldsymbol{\beta}}}^{-1}. \end{aligned}$$

Filling in the expressions (3.5) of $\mathbb{E}\big[\frac{\partial \log f_Y(y_j)}{\partial \boldsymbol{\beta}^T}\big]$ and (3.10) of $\mathcal{I}_{\widehat{\boldsymbol{\beta}}}^{-1}$ yields

$$\mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big] = \big(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega} \boldsymbol{X}\big)^{-1} \big(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{V} \mathbb{E}\big[\boldsymbol{y} - \boldsymbol{\mu}\big]^2 \boldsymbol{V} \boldsymbol{W} \boldsymbol{X}\big) \big(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega} \boldsymbol{X}\big)^{-1},$$

where $\boldsymbol{V}$ and $\boldsymbol{\Omega}$ are diagonal matrices with elements v_{jj} and ω_{jj} defined by (3.6) and (3.11) respectively.
Since $\mathbb{E}\big[y_j - \mu_j\big]^2 \equiv b''(\theta_j, m_j)$, we get

$$\boldsymbol{V} \mathbb{E}\big[\boldsymbol{y} - \boldsymbol{\mu}\big]^2 \boldsymbol{V} = \Omega,$$

therefore

$$\mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big] \approx \big(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega} \boldsymbol{X}\big)^{-1} \big(\boldsymbol{X}^T \boldsymbol{W}^2 \boldsymbol{\Omega} \boldsymbol{X}\big) \big(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega} \boldsymbol{X}\big)^{-1}. \tag{3.17}$$

An estimate of $\mathbb{V}\mathrm{ar}\big[\widehat{\eta}\big]$ is obtained similarly as in the local regression case, see Section 2.4.1. For $\widehat{\eta}(x_i) = \widehat{\beta}_0(x_i)$,

$$\widehat{\eta}(x_i) = \eta(x_i) + \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) + o \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|,$$

and,

$$\mathbb{E}\left[\widehat{\eta}(x_i)\right] = \eta(x_i) + \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T} \mathbb{E}\left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right].$$

From this we obtain the popular *sandwich estimate* of the variance, see Liang and Zeger (1986, p.15) as in equation 2.12,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\eta}(x_i)\right] = \frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0^T}\mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}(x_i) - \boldsymbol{\beta}(x_i)\right)^2\right]\frac{\partial k(\boldsymbol{\beta})}{\partial \beta_0}.$$

Since $\partial k(\boldsymbol{\beta})/\partial\beta_0^T = \boldsymbol{e}_1^T$, substituting (3.17) for $\mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big]$, we obtain

$$\mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big] \approx \boldsymbol{e}_1^T\big(\boldsymbol{X}^T\boldsymbol{W}\,\boldsymbol{\Omega}\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{\Omega}\boldsymbol{X}\big)\big(\boldsymbol{X}^T\boldsymbol{W}\,\boldsymbol{\Omega}\boldsymbol{X}\big)^{-1}\boldsymbol{e}_1. \quad (3.18)$$

We can also express the variance (3.18) in terms of $\boldsymbol{s}(x_i)$, the ith rows of the smooth weight diagram defined by equation 2.9. Rewriting the variance of the estimate leads to

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big] &\approx \\ &\boldsymbol{e}_1^T\boldsymbol{\Omega}^{-1}\boldsymbol{V}\mathbb{E}\big[\boldsymbol{y}-\boldsymbol{\mu}\big]^2\boldsymbol{V}\boldsymbol{\Omega}^{-1}\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\big)\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\big)^{-1}\boldsymbol{e}_1 \\ &= \boldsymbol{\Omega}^{-1}\boldsymbol{S}\boldsymbol{S}^T, \qquad (3.19)\end{aligned}$$

where $\boldsymbol{S}$ is the smooth weight diagram. Hence the variance approximation reduces to the following compact form,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big] &= \big[\omega_{ii}\big]^{-1}_{\mu=\widehat{\mu}}\sum_{j=1}^{n} s_j^2(x_i) \\ &= \frac{b''(\theta_i, m_i)}{m_i^2}\big(g'(\widehat{\mu}_i/m_i)\big)^2\|\boldsymbol{s}(x_i)\|^2. \qquad (3.20)\end{aligned}$$

By the delta method and as $\widehat{\mu}_i = m_i g^{-1}(\widehat{\eta}_i)$, we obtain an estimate of the variance of μ_i,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big[\widehat{\mu}(x_i)\big] &\approx \big[\partial g^{-1}(\eta_i)/\partial\eta_i\big]_{\eta=\widehat{\eta}} m_i^2\mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big] \\ &= \big[\partial g^{-1}(\eta_i)/\partial\eta_i\big]_{\eta=\widehat{\eta}} b''(\theta_i, m_i)\big(g'(\widehat{\mu}_i/m_i)\big)^2\|\boldsymbol{s}(x_i)\|^2. \qquad (3.21)\end{aligned}$$

Careful theoretical analysis of local likelihood is important. Many statistical software packages include functions for fitting generalized linear models, for instance the `glm()` function in `R`, R Development Core Team (2012). Since these functions usually allow weights for each observation, local likelihood models can be fitted by calling GLMs repeatedly, with a new set of weights for each fitting point. This approach produces correct estimates but incorrect inferences. The problem is that `glm()` interprets weights as a sample size. This appears as a multiplier for the $\boldsymbol{\Omega}$ matrix in $\big(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\big)^{-1}$, rather than the required $\boldsymbol{W}$. In particular, this implies the matrix $\big(\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{\Omega}\boldsymbol{X}\big)$ is computed incorrectly, and the standard errors are not correct, even asymptotically.

The assessed bias and variance have important applications in constructing confidence intervals. We can use the estimated bias and variance and rely on the asymptotic normality of the estimator to construct pointwise confidence intervals.

3.3.2 Pointwise confidence intervals

The confidence intervals should ideally take into account the underlying family of distributions. However, the theory for deriving such intervals seems quite intractable, see Loader (1999b, p.171).
Hence we must rely on methods based on normality assumptions, using the approximate variance. The local maximum likelihood estimator is usually asymptotically normal. This has been shown by Fan *et al.* (1995, p.143-145) in the context of generalized linear models

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \to N\big(0, \mathbb{V}\mathrm{ar}\big[\widehat{\boldsymbol{\beta}}\big]^{1/2}\big).$$

Within the framework of normal approximations, the simple solution is to use the variance stabilizing link. Under this link, the variance of $\widehat{\eta}(x_i)$ is, at least asymptotically, independent of the true parameter. It leads to confidence intervals whose widths depend only on the design points x_i.

By invoking asymptotic normality, we construct the pointwise confidence intervals adjusted to allow for bias as follows. With approximately $1-\alpha$ coverage probability, the unknown function $\eta(x_i)$ falls in the random interval

$$\begin{aligned} &\widehat{\eta}(x_i) - \widehat{\boldsymbol{b}}(\eta_i) \pm c\; \mathbb{V}\mathrm{ar}\big[\widehat{\eta}(x_i)\big]^{1/2}, \\ \text{or equivalently } &\widehat{\eta}(x_i) - \widehat{\boldsymbol{b}}(\eta_i) \pm c\; \big[\omega_{ii}\big]^{-1/2}_{\mu=\widehat{\mu}} \|\boldsymbol{s}(x_i)\|, \end{aligned} \tag{3.22}$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution.

Since $\boldsymbol{b}$ is unknown, a bias estimate is needed to form the estimated confidence intervals. The most common approaches as (3.15) are based on the *plug-in principle*, and plug-in bias estimates simply amount to increasing the order of the fit. In such cases, Loader (1999b, p.168) argue that an estimated interval is simply a construction of an under-smoothed interval centered around the estimate $\widehat{\eta}(x_i) - \widehat{\boldsymbol{b}}(\eta_i)$.

We can also compute confidence intervals for $\widehat{\mu}(x_i)$ and for transforms of the force of mortality or of the mortality rates by a function $k(.)$.
With (3.21), a $(1-\alpha)100\,\%$ confidence interval for $\widehat{\mu}(x_i)$ is

$$\begin{aligned} &\widehat{\mu}(x_i) - \widehat{\boldsymbol{b}}(\mu_i) \pm c\; \mathbb{V}\mathrm{ar}\big[\widehat{\mu}(x_i)\big]^{1/2}, \\ \text{equivalently } &\widehat{\mu}(x_i) - \widehat{\boldsymbol{b}}(\mu_i) \pm c\; m_i \big[\partial g^{-1}(\eta_i)/\partial \eta_i\big]_{\eta=\widehat{\eta}} \big[\omega_{ii}\big]^{-1/2}_{\mu=\widehat{\mu}} \|\boldsymbol{s}(x_i)\|. \end{aligned} \tag{3.23}$$

Finally, a confidence interval for $k(\widehat{q}(x_i))$ is given by

$$\begin{aligned} &k\big\{\widehat{q}(x_i) - \widehat{\boldsymbol{b}}(q_i)\big\} \pm c\; \big[\partial k(q_i)/\partial q_i\big]_{q=\widehat{q}} \mathbb{V}\mathrm{ar}\big[\widehat{q}_i\big]^{1/2}, \\ \text{or equivalently } &k\big\{\widehat{q}(x_i) - \widehat{\boldsymbol{b}}(q_i)\big\} \pm c\; \big[\partial k(q_i)/\partial q_i\big]_{q=\widehat{q}} \big[\partial g^{-1}(\eta_i)/\partial \eta_i\big]_{\eta=\widehat{\eta}} \mathbb{V}\mathrm{ar}\big[\widehat{\eta}_i\big]^{1/2}. \end{aligned}$$

When $k(.)$ is the logit function, we obtain

$$\operatorname{logit}\{\widehat{q}(x_i) - \widehat{\boldsymbol{b}}(q_i)\} \pm c\, 1/(\widehat{q}_i(1-\widehat{q}_i))\big[\partial g^{-1}(\eta_i)/\partial\eta_i\big]_{\eta=\widehat{\eta}}\big[\omega_{ii}\big]^{-1/2}_{\mu=\widehat{\mu}}\|\boldsymbol{s}(x_i)\|. \tag{3.24}$$

3.3.3 Effective dimension of a non-linear smoother

As in the local polynomials method, the fitted degrees of freedom are useful for assessing a single fit and comparing two fits. Similarly to the local polynomial method, we define the influence function at x_i:

$$\operatorname{infl}(x_i) = \boldsymbol{e}_1^T\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\right)^{-1}\boldsymbol{e}_1.$$

An interpretation of the influence function is the leave-i-out cross validation approximation which will be used when we assess the goodness of fit,

$$\widehat{\eta}_{-i}(x_i) = \widehat{\eta}(x_i) - \operatorname{infl}(x_i)\partial\log f_Y/\partial\widehat{\eta}(x_i). \tag{3.25}$$

Since $\mathbb{E}\ \left(\partial\log f_Y/\partial\,\widehat{\eta}(x_i)\right)^2 = -\mathbb{E}\ \left(\partial^2\log f_Y/\partial\,\widehat{\eta}^2(x_i)\right)$, the fitted degrees of freedom are defined as

$$\begin{aligned}\upsilon_1 &= \sum_{i=1}^{n}\operatorname{infl}(x_i)\mathbb{E}\left(-\partial^2\log f_Y/\partial\,\widehat{\eta}^2(x_i)\right)\\ &= \sum_{i=1}^{n}\operatorname{infl}(x_i)\ \omega_{ii},\end{aligned} \tag{3.26}$$

where ω_{ii} is defined as in (3.11). Another definition of the fitted degrees of freedom for a local likelihood model is the sum of the variance of the fitted values:

$$\upsilon_2 = \sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left[\widehat{\eta}(x_i)\right]\ \omega_{ii}. \tag{3.27}$$

3.4 Diagnostics for local likelihood

This section covers diagnostic and model selection issues for local likelihood. First, we discuss techniques similar to the ones used in parametric generalized linear models by McCullagh and Nelder (1989). These are the *classical* selectors. Second, we present a *plug-in* methodology to choose the theoretical bandwidth.

3.4.1 Classical selectors

These techniques are natural extensions of the local regression methodology introduced in Section 2.5. In local polynomial regression we developed diagnostics methods based on the residuals $y_i - \widehat{\mu}(x_i)$, and the residual sum of squares. For local likelihood models, these tools are less natural. In this case, it is more natural to consider diagnostics based on the ratio $y_i/\widehat{\mu}(x_i)$ rather than the difference.

The predictor of a future observation at a point x_i is $g^{-1}(\widehat{\eta}(x_i))$ where $g(.)$ is the link function. One possible loss function is the deviance (or *scaled deviance*) for a single observation (x_i, y_i), defined by

$$\begin{aligned} D(y_i, \widehat{\theta}(x_i)) &= 2\left(\sup_{\theta} l\big(y_i, \theta(y_i)\big) - l\big(y_i, \theta(\widehat{\mu}_i)\big)\right) \\ &= 2/\phi \; m_i \left(y_i\big(\theta(y_i) - \theta(\widehat{\mu}_i)\big) - b\{\theta(y_i)\} + b\{\theta(\widehat{\mu}_i)\}\right). \end{aligned}$$

It is easily seen that $D\big(y_i, \theta(\widehat{\mu}_i)\big) \geq 0$, and $D\big(y_i, \theta(\widehat{\mu}_i)\big) = 0$ if $y_i = g^{-1}(\widehat{\eta}_i)$. Since it is based on the likelihood, the deviance provides a measure of the evidence an observation y_i provides against $\widehat{\eta}(x_i)$ being the true value of $\eta(x_i)$. The total deviance is defined as

$$\sum_{i=1}^{n} D(y_i, \hat{\theta}(x_i)). \tag{3.28}$$

This generalizes the residual sum of squares for a regression model. Examples of the form of deviances are given in Table 3.1.

GLM	Scaled Deviance
Normal	$1/\phi \; \sum_i m_i (y_i - \widehat{\mu}_i)^2$
Poisson	$2/\phi \; \sum_i m_i \left(y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\right)$
Binomial	$2/\phi \; \sum_i m_i \left(y_i \log(y_i/\widehat{\mu}_i) + (n_i - yi) \log\left((n_i - y_i)/(n_i - \widehat{\mu}_i)\right)\right)$

Table 3.4: *Examples of forms of scaled deviance*

We can extend the cross-validation and Cp methods introduced for local polynomial regression. It is natural to base these methods directly on the likelihood or the deviance functions.

The likelihood (or deviance) cross validation criterion is defined by substituting the leave-x_i-out estimate $\theta_{-i}(\widehat{\mu}_i)$ in the total deviance (3.28);

$$\begin{aligned} LCV(\theta(\widehat{\mu}_i)) &= \sum_{i=1}^{n} D(y_i, \theta_{-i}(\widehat{\mu}_i)) \\ &= C - 2\sum_{i=1}^{n} \partial \log f_Y / \partial\, \widehat{\eta}_{-i}(x_i), \end{aligned}$$

where C depends on the observations y_i, but not on the estimate $\theta(\widehat{\mu}_i)$ and hence not on the bandwidth nor on the local polynomial degree.

The computation of the n leave-i-out estimates can be expensive. An alternative to deletion methods is the method of infinitesimal perturbations developed in Cook (1977) for linear models, and Pregibon (1981) for logistic regression models. These techniques relate the deletion estimate $\mu_{-i}(x_i)$ with the estimate $\widehat{\mu}(x_i)$ and the influence function $\mathrm{infl}(x_i)$.

In the likelihood setting, the simplification of CV no longer holds. Instead, we have to develop some approximations. First, we identify an influence function such as (3.25). Substituting (3.25) into the deviance and using a one-term Taylor series gives

$$D(y_i, \theta_{-i}(\widehat{\mu}_i)) \approx D(y_i, \theta(\widehat{\mu}_i)) + 2\,\mathrm{infl}(x_i)\left(\partial \log f_Y / \partial\, \widehat{\eta}(x_i)\right)^2 .$$

Summing this over all observations gives an approximation to the likelihood cross validation statistic. It leads to a generalization of the Akaike information criterion to local likelihood models

$$AIC(\theta(\widehat{\mu}_i)) = \sum_{i=1}^{n} D(y_i, (\theta(\widehat{\mu}_i))) + 2\,\upsilon_1,$$

where υ_1 is the degrees of freedom for the local likelihood fit.

One has to keep in mind that graduation, and hence model selection, is a very effective compromise between two objectives, the elimination of irregularities and the achievement of a desired mathematical shape to the progression of the mortality rates. This underlines the importance of experience, and above all, of thorough investigation of data as the prerequisites of reliable judgment, as we must first inspect the data and take the decision as the type of irregularity we wish to retain.

In practice, one needs to choose λ and the fitting variable to balance the trade-off between bias and variance. To find the right constellation, we use graphical tools for displaying the whole profile of the selectors curves as introduced in Section 2.5.4. For that, it is important to note that relying exclusively in practice in a global criterion is unwise because a global criterion does not provide information about where in the design space the contributions to bias and variance are coming from.

In conjunction with looking at the plots, one always has to look at residual plots. In the case of generalized linear models, we denote:

i. The response residual: $r_i = y_i - \widehat{\mu}_i$;

ii. The Pearson residual: $r_i = (y_i - \widehat{\mu}_i)/(\sqrt{\mathbb{V}\mathrm{ar}\,[\widehat{\mu}_i]})$;

iii. The deviance residual: $r_i = \mathrm{sign}(y_i - \widehat{\mu}_i) D(y_i, \theta(\widehat{\mu}_i))^{1/2}$.

Such residual plots provide a powerful diagnostic that nicely complements the selection criteria. The diagnostic plots can show lack of fit locally and we have the opportunity to judge the lack of fit based on our knowledge of both the mechanism generating the data and of the performance of the smoothers used in the fitting. Superimposed on the response and Pearson residuals plots is a *loess* smooth. If a local likelihood model correctly models a dataset, no strong patterns should appear in the response and Pearson residuals.

There is no deterministic method to obtain the constellation of smoothing parameters with the classical selectors. The purpose for which the mortality table is required must be kept clearly in mind, and the final choice of graduation is always a matter of judgment. The statistical criteria described should be regarded as aids in the assessment of the graduation and not interpreted too rigidly.

3.4.2 Plug-in method and theoretical bandwidth

With the estimated MSE (2.38), by analogy of the local least squares problem, Fan *et al.* (1998, p.599-600) formulate a bandwidth selection rule as follows: Fit a polynomial of order $p+a$ (choosing $a=2$) and find the *pilot* bandwidth h° that minimizes the integrated extended residual squares criterion,

$$IERSC(h) = \int_{[x_{min}, x_{max}]} ERSC(t,h)dt,$$

with the $ERSC$ defined as

$$ERSC(x_i,h) = \widehat{\sigma}_\circ^2(x_i)\left(1+(p+1)/N\right), \tag{3.29}$$

where N^{-1} is the first diagonal element of the matrix

$$\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^2\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}$$

and $\widehat{\sigma}_\circ^2(x_i)$ is the normalized weighted residual sum of squares using the working dependent variable $\boldsymbol{z}$ defined as expression (3.12) after fitting locally a $(p+a)$th order polynomial. The justification of this is simple. Firstly, the bias of $\widehat{\boldsymbol{\beta}}$ comes from the local polynomial approximation of ψ. Hence it is the same for the local likelihood method as for the local least squares problem. Secondly, comparing equation (3.20) with equation 2.17, the asymptotic variance of the local likelihood problem corresponds to that of the least squares problem with $\sigma^* = \omega = \sigma_\circ$. Treating $\boldsymbol{\beta}^\circ$ in equation (3.12) as fixed, the working dependent variable z_i has the same variance structure, namely

$$\mathbb{V}\mathrm{ar}\left[z_i\right] \approx \sigma_\circ(x_i) \approx \omega.$$

Thus, having the bandwidth h° for estimating β_{p+1}, obtain estimates $\widehat{\beta}_{p+1}^\circ(x_i)$, $\widehat{\beta}_{p+2}^\circ(x_i)$ and $\widehat{\sigma}_\circ^2(x_i)$. With these estimated parameters, compute the estimated bias $\widehat{\mathrm{bias}}_{p,v}(x_i)$ and variance $\widehat{\mathrm{var}}_{p,v}(x_i)$ of $\widehat{\beta}_v$, which are respectively the $(v+1)$st element of vector (3.15) and the $(v+1)$st diagonal element of the estimated expression (3.17). Combining these estimates yields the estimated MSE (2.38). This leads to the bandwidth selector

$$\widehat{h}_{p,v} = \arg\min_h \left\{ \int_{[x_{min}, x_{max}]} \widehat{MSE}_{p,v}(t,h)dt \right\}.$$

3.5 Model for the probabilities of death

3.5.1 The local likelihood binomial model

Let us suppose that l_j persons come under observation at age x_j and continue to be under observation until they survive to x_j+1 or die before that age. In this case we denote the number of policy holders initially exposed to risk as l_j. Moreover, let us suppose that the probability of death during the year for each one of them is q_j, and that the death or survival of one is independent of the death or survival of the others. If we call D_j the random variable that represents the number of deaths that occur in the year, we will use the usual model for the number of deaths,

$$D_j \sim \text{Binomial}(l_j, q_j),$$

and the observed death rate, which is the maximum likelihood estimate of q_j is

$$\dot{q}_j = \frac{d_j}{l_j}.$$

The binomial probability function is expressed as

$$f_D(d_j, l_j, q_j) = \binom{l_j}{d_j} q_j^{d_j} (1-q_j)^{l_j - d_j}.$$

In exponential family form, the binomial distribution may be written as

$$\begin{aligned} f_D(d_j, l_j, q_j) &= \exp\left\{ d_j \log(q_j) + l_j \log(1-q_j) - d_j \log(1-q_j) + \log\binom{l_j}{d_j} \right\} \\ &= \exp\left\{ d_j \log\left(\frac{q_j}{1-q_j}\right) + l_j \log(1-q_j) + \log\binom{l_j}{d_j} \right\}. \end{aligned} \tag{3.30}$$

The local log-likelihood at x_i is then

$$\begin{aligned} L(q_i) &= \sum_{j=1}^{n} w_j \log f_D(d_j, l_j, q_j) \\ &= \sum_{j=1}^{n} w_j \left\{ d_j \log\left(\frac{q_j}{1-q_j}\right) + l_j \log(1-q_j) \right\}, \end{aligned} \tag{3.31}$$

where the constant $c(d_j, \phi, w_j)$ function of d_j not involving q_j, namely

$$\sum_{j=1}^{n} w_j \log\binom{l_j}{d_j},$$

has been omitted. From (3.30) the canonical parameter θ and the cumulant functions $b(\theta)$ are given by

$$\begin{aligned} \theta_j &= \log\left(\frac{q_j}{1-q_j}\right) \\ b(\theta_j, m_j) &= -l_j \ \log(1-q_j). \end{aligned}$$

The mean and variance functions are calculated as the first and second derivatives of the cumulant function:

$$
\begin{aligned}
b'(\theta_j, m_j) &= \frac{\partial b}{\partial q_j}\frac{\partial q_j}{\partial \theta_j} = l_j q_j \\
b''(\theta_j, m_j) &= \frac{\partial^2 b}{\partial q_j^2}\left(\frac{\partial q_j}{\partial \theta_j}\right)^2 + \frac{\partial b}{\partial q_j}\frac{\partial^2 q_j}{\partial \theta^2} = l_j q_j (1 - q_j).
\end{aligned}
$$

Hence, the relationship of q_j and μ_j is given by $\mu_j = l_j q_j$ and the variance is

$$
\mathbb{V}\mathrm{ar}\left[\mu_j\right] = b''(\theta_j, m_j) = \mu_j \left(1 - \frac{\mu_j}{l_j}\right).
$$

The systematic part of the model specifies the relation between the vector $\boldsymbol{q}$ and the experimental conditions as summarized by the model matrix $\boldsymbol{X}$ of order $n \times p$. Using the variance stabilizing link indicated in Table 3.3, this relationship takes the form

$$
g(q_j) = \eta_j = \sin^{-1}\left(\sqrt{\frac{\mu_j}{l_j}}\right) = \sum_{p=0}^{n} \beta_p (x_j - x_i)^p
$$

The inverse link can easily be derived from the above as

$$
\begin{aligned}
\mu_j &= g^{-1}(\eta_j) = l_j \sin^2(\eta_j) \\
&= l_j \sin^2\left(\sum_{p=0}^{n} \beta_p (x_j - x_i)^p\right). \qquad (3.32)
\end{aligned}
$$

Since $q_j = \sin^2(\eta_j)$, we have $1 - q_j = \cos^2(\eta_j)$. Substituting (3.32) into (3.31) gives

$$
\begin{aligned}
&L(\boldsymbol{\beta}) = 2\sum_{j=1}^{n} w_j \\
&\times\left(d_j \log\left\{\sin\left[\sum_{p=0}^{n}\beta_p (x_j - x_i)^p\right]\right\} + (l_j - d_j)\log\left\{\cos\left[\sum_{p=0}^{n}\beta_p (x_j - x_i)^p\right]\right\}\right).
\end{aligned}
$$

3.5.2 Estimation method

Following the general method given in Section 3.2.4, we now derive the local likelihood equations for the parameters $\boldsymbol{\beta}$. The derivative of the local log-likelihood function with respect to $\boldsymbol{\beta}$ is

$$
\frac{\partial}{\partial \beta_v} L = \sum_{j=1}^{n} w_j m_j \frac{d_j - \mu_j}{b''(\theta_j, m_j)} \frac{\partial \mu_j}{\partial \eta_j} (x_j - x_i)^v. \qquad (3.33)
$$

The derivative of the arcsine square root link function is

$$
g'(\mu_j / m_j) = \frac{\partial \eta_j}{\partial \mu_j} = \frac{\partial}{\partial \mu_j}\sin^{-1}\left(\sqrt{\frac{\mu_j}{l_j}}\right) = \frac{l_j}{2\sqrt{\mu_j (l_j - \mu_j)}}
$$

From expressions (3.9) and (3.11), with $\mathbb{V}\mathrm{ar}\big[d_j\big] = \mu_j \left(1 - \frac{\mu_j}{l_j}\right)$, the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathcal{I}_{vk} &= \sum_{j=1}^{n} w_j \mathbb{V}\mathrm{ar}[d_j] \left(\frac{m_j}{b''(\theta_j, m_j)}\right)^2 \left(\frac{\partial \mu_j}{\partial \eta_j}\right)^2 (x_j - x_i)^v (x_j - x_i)^k \\ &= \left\{\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\right\}_{vk},\end{aligned}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with elements $\omega_{jj} = 4l_j$.

Then parameter estimates are obtained in the following way. Given initial estimates $\widehat{\boldsymbol{\beta}}^*$, we compute the vector $\widehat{\boldsymbol{\mu}}^*$ and $\widehat{\boldsymbol{\eta}}^*$. Using these values, define the adjusted dependent variable $\boldsymbol{z}$ with components

$$z_j = \widehat{\eta}_j + \frac{d_j - \widehat{\mu_j}}{l_j} g'(\widehat{\mu}_{j/m_j}),$$

all quantities being computed at the initial estimate $\widehat{\boldsymbol{\eta}}^*$. Maximum likelihood estimates satisfy the equations

$$\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{z},$$

which are solved iteratively. The revised estimate is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{z}.$$

To compute the criteria used for model selection, we need to determine the deviance.
Recall that the deviance is calculated as $D = 2\left\{L(d_j, d_j) - L(\mu_j, d_j)\right\}$. We list the appropriate calculations of the deviance for specific observations. Given the local log-likelihood (3.31),

$$\begin{aligned}&D(l_j > 1; 0 < d_j < l_j) = \\ &\qquad 2d_j \log\left(\frac{d_j}{1-d_j}\right) + l_j \log(1-d_j) - d_j \log\left(\frac{\mu_j}{1-\mu_j}\right) + l_j(1-\mu_j) \\ &\qquad = 2d_j \log\left(\frac{d_j}{\mu_j}\right) + (l_j - d_j)\log\left(\frac{l_j - d_j}{l_j - \mu_j}\right) \\ &D(l_j > 1; d_j = l_j) = 2d_j \log\left(\frac{l_j}{\mu_j}\right) \\ &D(l_j > 1; d_j = 0) = 2d_j \log\left(\frac{d_j}{d_j - \mu_j}\right).\end{aligned}$$

Finally, the total deviance at x_i is computed as the sum of the single deviances weighted by w_j.

3.5.3 Statistical Inference

Within the framework of normal approximation, the variance approximation (3.19) reduces to

$$\begin{aligned}\mathbb{V}\text{ar}\big[\widehat{q}_i\big] &= \mathbb{V}\text{ar}\big[\widehat{\eta}(x_i)\big] \\ &\approx \omega_{ii}^{-1}\|\boldsymbol{s}(x_i)\|^2 \\ &= (4\,l_i)^{-1}\|\boldsymbol{s}(x_i)\|^2 \qquad (3.34)\end{aligned}$$

Then, having the plug-in bias $\widehat{\boldsymbol{b}}(\eta_i)$ from expression (3.16), we construct the pointwise confidence intervals adjusted to allow for bias as follows. With approximately $(1-\alpha)$ coverage probability, the unknown function $q(x_i)$ falls in the random interval

$$\widehat{q}(x_i) - \widehat{\boldsymbol{b}}(q_i) \pm c\,\frac{1}{2}l_i^{-1/2}\|\boldsymbol{s}(x_i)\|,$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution. We can also construct pointwise confidence intervals for $\mu(x_i)$, the number of deaths, and $\text{logit}(q(x_i))$.

A confidence interval for $\widehat{\mu}(x_i)$ is

$$\widehat{\mu}(x_i) - \widehat{\boldsymbol{b}}(\mu_i) \pm c\, m_i\big[\partial g^{-1}(\eta_i)/\partial\eta_i\big]_{\eta=\widehat{\eta}}\big[\omega_{ii}\big]_{\mu=\widehat{\mu}}^{-1/2}\|\boldsymbol{s}(x_i)\|$$

or equivalently $\widehat{\mu}(x_i) - \widehat{\boldsymbol{b}}(\mu_i) \pm c\,\widehat{q}_i(1-\widehat{q}_i)l_i^{1/2}\|\boldsymbol{s}(x_i)\|$.

And a confidence interval for $\text{logit}(\widehat{q}(x_i))$ is given by

$$\text{logit}\big(\widehat{q}(x_i) - \widehat{\boldsymbol{b}}(q_i)\big) \pm c\,\frac{1}{2}(l_i\widehat{q}_i(1-\widehat{q}_i))^{-1/2}\|\boldsymbol{s}(x_i)\|.$$

3.5.4 Applications

We now consider non-parametric logistic regression to illustrate the method. To present the local likelihood approach, we discuss the two applications presented in Section 2.6.

Choice of the constellation of the smoothing parameters

The pattern displayed by the crude mortality rates is relatively smooth for both datasets. We use the *AIC* and *LCV* criteria and graphical diagnostics presented in Section 3.4 to guide the modeling. The *AIC* and the *LCV* criteria are relatively close. We notice however that *LCV* tends to select a smoother constellation of parameters than the *AIC*, which, considering the underlying pattern of the data, is satisfactory. The selected bandwidth should not be too large to capture the structure at the right boundary.

Table 3.5 displays the chosen constellation of smoothing parameters for the local likelihood binomial approach and for each dataset with the corresponding fitted degrees of freedom. Recall $\lambda = 2\,h + 1$.

	λ	Degree	$W(.)$	Fitted DF
Dutch Male	15	2	Gaussian	18.45
Dutch Female	17	2	Gaussian	16.39

Table 3.5: *Elected constellation of smoothing parameters and fitted degrees of freedom*

A local quadratic fit is needed to capture the mortality patterns. The choice differs by the chosen bandwidth. The mortality patterns for the Dutch female population are less pronounced than for the male. A higher λ is then needed to smooth the structures which we believe less accentuated than for the Male population. The corresponding fitted degrees of freedom for the female population are lower than the ones for the male, indicating that we have applied more smoothing.

Table 3.6 presents the theoretical bandwidth provided by the plug-in method developed in Section 3.4.2. We fit a polynomial of degree 2 and use the corresponding weight functions elected in Table 3.5. The values of λ are reported below.

	Pilot bandwidth	Bandwidth
Dutch Male	$\lambda = 7$	$\lambda = 13$
Dutch Female	$\lambda = 7$	$\lambda = 15$

Table 3.6: *Pilot and bandwidths selected by the plug-in method*

As the amount of curvature of the observed probability of death is relatively similar whichever dataset is considered, the selected pilot bandwidths are the same. The bandwidths confirmed our choices presented in Table 3.5, being relatively close and agreeing with our ranking.

Plots of the fits and residuals plots

Figure 3.1 presents the mortality rates graduated by the local Binomial likelihood approach with the smoothing parameters displayed in Table 3.5 above.

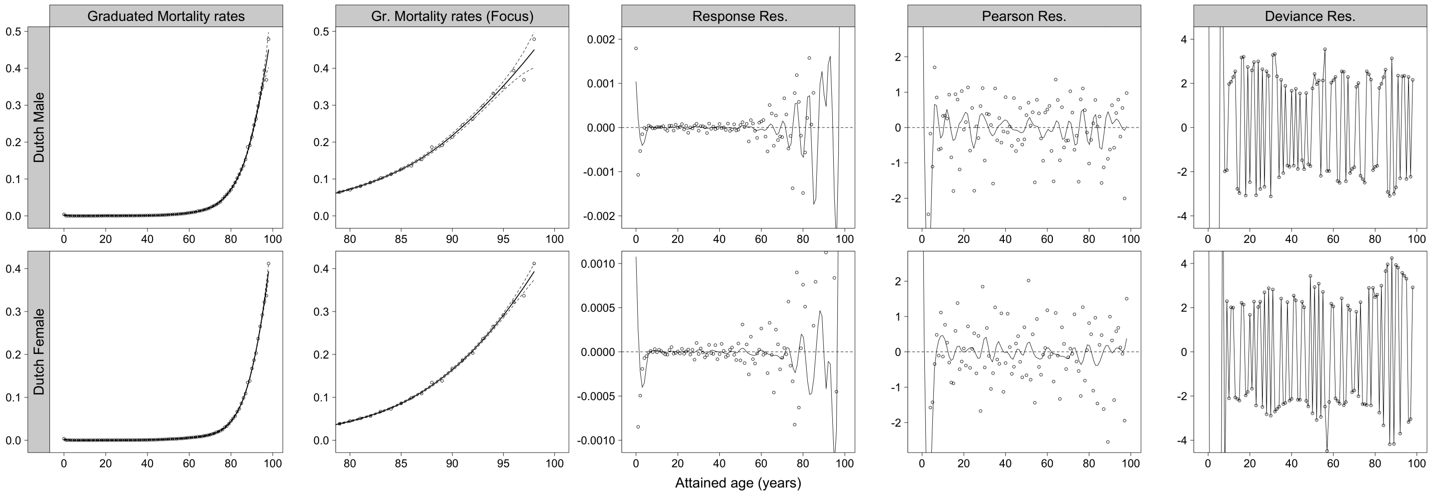


Figure 3.1: *Graduated mortality rates by local Binomial model with 95 % pointwise confidence intervals and corresponding residuals plots for Dutch Male and Female population, 2008. Source: HMD.*

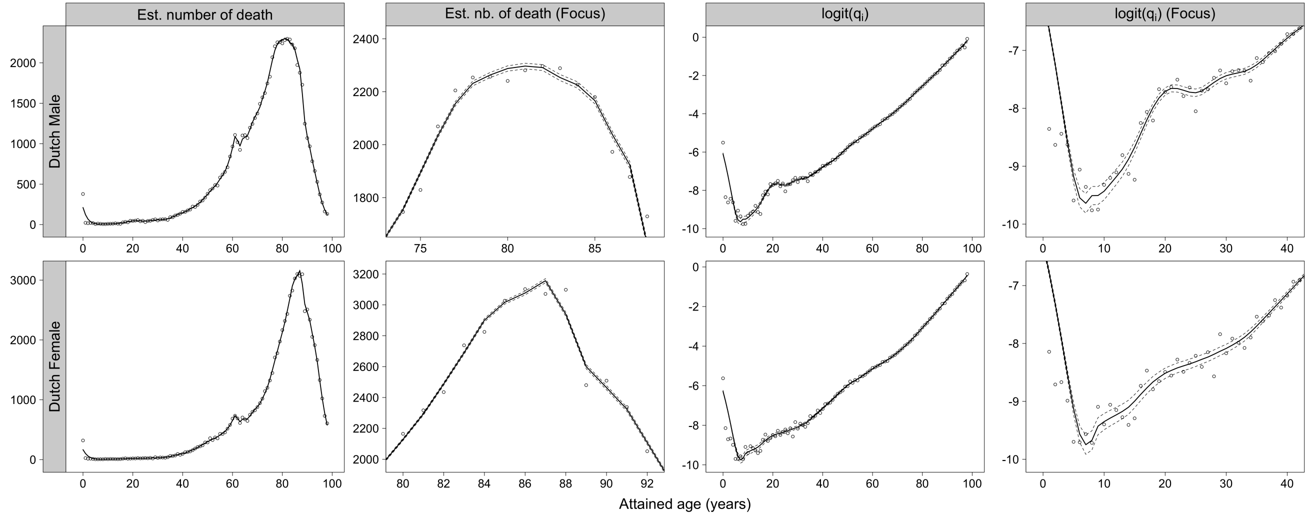


Figure 3.2: *Estimated number of death and Logit transform of the graduated mortality rates by local Binomial model with* 95 % *pointwise confidence intervals for Dutch Male and Female population, 2008. Source: HMD.*

Next to the plots of the fits, we display the residuals plots. Superimposed on the responses and Pearson residuals is a *loess* smooth curve.

This *loess* smooth curve shows an important lack of fit at the right boundary, when the data have a large curvature. The clusters of residuals are even more important when the dataset has a small volume of observations and the observations are sparse. However, due to the underlying structure of the mortality data it is normal to get higher residuals at the right boundary than in the rest of the curve.

For both datasets, the Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the model adequately captures the variability of these datasets. However, a clear lack of fit is shown by the Pearson residuals at the left boundary, which is confirmed by the shape of the deviance residuals. The deviance residuals present, for the youngest ages, several successive residuals having the same sign. It illustrates that the mortality rates are over-smoothed locally and hence we strongly overestimate the probability of death for the youngest ages as we can see in Figure 3.2.
A smaller bandwidth or a higher polynomial degree shall be used to capture the structure but it would be at the expense of a lack of fit in the middle of the table. When the dataset presents a high structure in the boundaries, a global constellation of smoothing parameters fails to provides an adequate fit to the data. To deal with such problems, we would rather use locally adaptive smoothing methods, which vary the amount of smoothing in a location dependent manner, so as to obtain a satisfactory fit over the whole range. These approaches are covered in Chapter 4.

Plots of the smoothers

Since the local likelihood estimate does not have an explicit representation, the smooth weight diagram can not be derived as in the local regression case. However, we can provide an illustration of the weight function associated with the i-th point at the last iteration. The weight function associated with the i-th point is used to compute the weights in the i-th row of the 99×99 smoother $\left(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{\Omega}$, and is shown in Figures 3.3 and 3.4, below, with the influence values.

For our applications we used the boundary correcting kernel *type 1*, see Section 2.3.4. Note again that the criteria used for model selection have been computed over a restricted number of observations.

The influence values measure the sensitivity of the fitted curves to the individual data points. It shows us the amount of smoothing applied locally. For instance, in Figure 3.3 right panel, the influence values at the boundaries are lower than the ones in the central region. It indicates that locally we have applied more smoothing in the boundaries than in the rest of the curve. On the other hand, $\operatorname{infl}(x_{61})$ (≈ 0.254) is larger than the influence values for observations in the central region (on average ≈ 0.227). It illustrates that

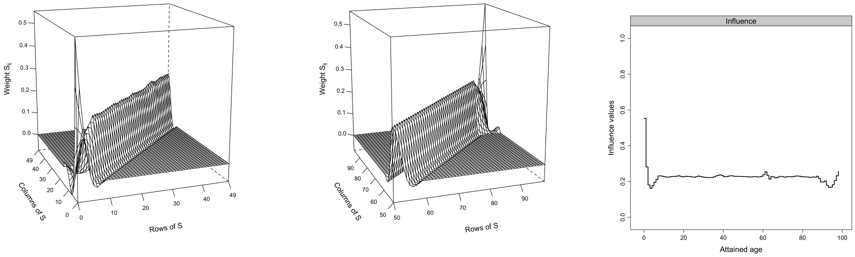


Figure 3.3: *Smoother S_{ij}: left panel: $i, j = 0, \ldots, 49$, center panel: $i, j = 50, \ldots, 98$ and influence values for the Dutch Male population, 2008. Source: HMD.*

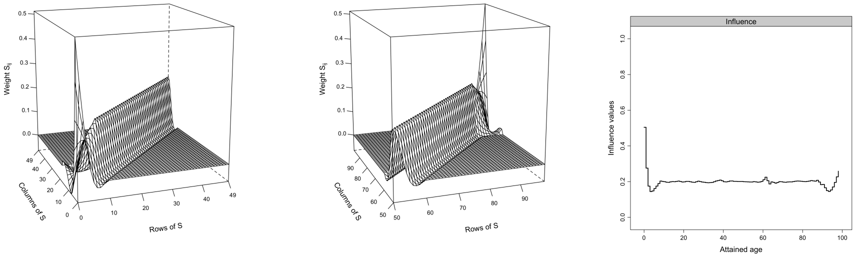


Figure 3.4: *Smoother S_{ij}: left panel: $i, j = 0, \ldots, 49$, right panel: $i, j = 50, \ldots, 98$ and influence values for the Dutch Female population, 2008. Source: HMD.*

observation x_{61} contributes more than average to the fitted value and thus less smoothing has been applied locally.

Comparison with Whittaker-Henderson smoothing

Similarly to the local polynomials method, we apply the criteria presented in Section 2.5.1 to find the value of parameters h and z. We picked the constellation $h = 1$ and $z = 2$ for the male, and $h = 2$ and $z = 2$ for the female population, given by Rice's T criterion, Rice (1984), leading to 24.18 and 20.73 fitted degrees of freedom respectively. Figure 3.5 presents graphical comparisons of the local binomial approach and the Whittaker-Henderson model.

The top left panel presents the graduated mortality rates for the Dutch male population. The graduated series by the local binomial model displays a smoother pattern. The corresponding degrees of freedom are lower than the ones obtained by the Whittaker-Henderson model, illustrating that the model is showing less features. The influence values obtained by the local binomial models are, up to the right boundary, below the ones computed with the Whittaker-Henderson model, $\text{infl}_{WH}(x_i) = diag((\boldsymbol{V} + hK_z^T K_z)^{-1}\boldsymbol{V})$,

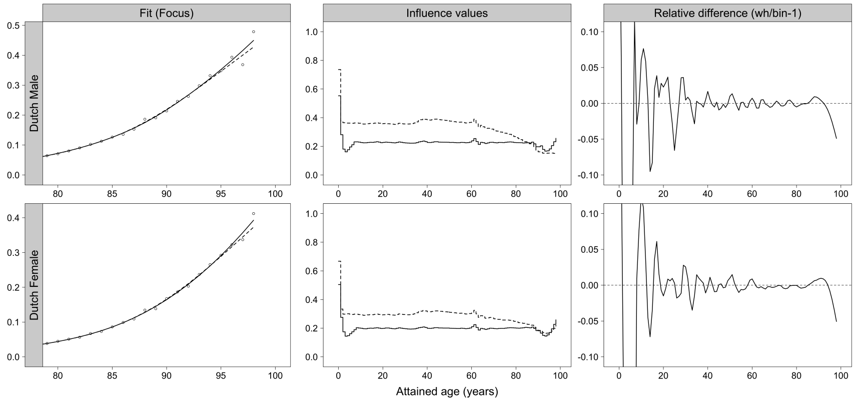


Figure 3.5: *Graphical comparisons between the local binomial approach (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008. Source: HMD.*

top center panel. It indicates that, up to the right boundary, more smoothing has been applied by the local binomial approach. The relative difference is more important at the boundaries, where the Whittaker-Henderson model does not need special treatment.

The bottom left panel shows the graduated mortality rates for the Dutch female population. Similar remarks can be made. The fitted degrees of freedom obtained by the Whittaker-Henderson smoothing are larger, illustrating that the model is showing more features. The influence values, bottom center panel, are higher. As for the male population, the relative difference is larger in the boundaries.

The graduated series are less smooth than the ones obtained by the local binomial approach. The smoothing parameters of the Whittaker-Henderson model have been chosen by minimizing Rice's T statistic, one of the so-called classical criteria. Other methods, described in Section 2.5, could have been tried.

3.6 Model for the forces of mortality

3.6.1 The local likelihood Poisson model

Let us now suppose that l_j persons enter observation under the hypothesis that the force of mortality (instantaneous mortality rate) is a constant during the period of observation and that the death or survival of each one is independent. In this case l_j represents those central exposed to risk, whereas in the previous section l_j denoted initial exposures.

Hence the force of mortality, φ_j, is the average risk to which the population is subjected during its passage through the year of age $x_j + 1$, and is a different concept from q_j, which represents the total effect of mortality in terms on proportion who fail to survive the whole year of age $x_j + 1$ without reference to the variation of mortality risk over the course of that year.

The number of deaths that occur in the period of observation, D_j, will have a Poisson distribution with mean and variance equal to μ_j. We consider the graduation of μ_j/l_j, with

$$D_j \sim \text{Poisson}(\mu_j).$$

The Poisson probability distribution function is

$$f_D(d_j;\mu_j) = e^{-\mu_j}\mu^{d_j}/d_j!,$$

or in exponential family form as

$$f_D(d_j;\mu_j) = \exp\{d_j \log(\mu_j) - \mu_j - \log\ d_j!\}.$$

The local log-likelihood function at x_i can be deduced from the exponential form of the distribution:

$$\begin{aligned} L(\mu_i) &= \sum_{j=1}^{n} w_j \log f_D(d_j;\mu_j) \\ &= \sum_{j=1}^{n} w_j \left\{d_j \log(\mu_j) - \mu_j - \log\ d_j!\right\}. \end{aligned} \tag{3.35}$$

When the response $d_j = 0$, the individual log-likelihood functions reduce to

$$L_j(\mu_j; d_j = 0) = -\mu_j.$$

The link and the cumulant function are then derived as

$$\begin{aligned} \theta_j &= \log(\mu_j) \\ b(\theta_j, m_j) &= \mu_j. \end{aligned}$$

The mean and variance functions are calculated as the first and second derivative with respect to θ_j, so

$$\begin{aligned} b'(\theta_j, m_j) &= \frac{\partial b}{\partial \mu_j}\frac{\partial \mu_j}{\partial \theta_j} = \mu_j, \\ b''(\theta_i, m_j) &= \frac{\partial^2 b}{\partial \mu_j^2}\left(\frac{\partial \mu_j}{\partial \theta_j}\right)^2 + \frac{\partial b}{\partial \mu_j}\frac{\partial^2 \mu_j}{\partial \theta^2} = \mu_j. \end{aligned}$$

The dependence of μ_j on the covariate vector is specified by the link function. Using the variance stabilizing link indicated in Table 3.3, we have

$$g(\mu_j) = \eta_j = \sqrt{\mu_j} = \sum_{p=0}^{n} \beta_p (x_j - x_i)^p.$$

The inverse link is easily derived

$$\mu_j = g^{-1}(\mu_j) = \eta_j^2 = \big(\sum_{p=0}^{n} \beta_p (x_j - x_i)^p\big)^2.$$

Using (3.35), the log-likelihood can also be parameterized in terms of $\sum_{p=0}^{n} \beta_p(x_j - x_i)^p$. Substituting the inverse link in place of each μ_j gives

$$L(\boldsymbol{\beta}) = \sum_{j=1}^{n} w_j \left(2\, d_j \log\bigl(\sum_{p=0}^{n} \beta_p(x_j - x_i)^p\bigr) - \bigl(\sum_{p=0}^{n} \beta_p(x_j - x_i)^p\bigr)^2 - \log\ d_j! \right)$$

3.6.2 Estimation method

The derivative of the local Poisson log-likelihood function with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial}{\partial \beta_v} L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j} \frac{\partial \mu_j}{\partial \eta_j} (x_j - x_i)^v. \tag{3.36}$$

The derivative of the link function is calculated as

$$g'(\mu_j/m_j) = g' = \frac{\partial \eta_j}{\partial \mu_j} = \frac{\partial}{\partial \mu_j} \mu_j^{1/2} = \frac{1}{2} \mu_j^{-1/2}.$$

From expressions (3.9) and (3.11), the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned} \mathcal{I}_{vk} &= \sum_{j=1}^{n} w_j \frac{1}{\mu_j} \left(\frac{\partial \mu_j}{\partial \eta_j} \right)^2 (x_j - x_i)^v (x_j - x_i)^k \\ &= \sum_{j=1}^{n} w_j\ 4 (x_j - x_i)^v (x_j - x_i)^k \\ &= \left\{ \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Omega} \boldsymbol{X} \right\}_{vk}, \end{aligned}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with elements $\omega_{jj} = 4$. Again, in case of log-linear models, i.e., using the canonical link, equation (3.36), when written in matrix notation, reduces to

$$\frac{\partial}{\partial \beta_v} L = \boldsymbol{X}^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}).$$

Following the general Fisher scoring procedure, Section 3.2.4, we obtain the estimates. Given initial estimates $\widehat{\boldsymbol{\beta}}^*$, we may compute the vector $\widehat{\boldsymbol{\mu}}^*$ and $\widehat{\boldsymbol{\eta}}^*$. Using these values, we define the adjusted dependent variable $\boldsymbol{z}$ with components

$$\begin{aligned} z_j &= \widehat{\eta}_j + (d_j - \widehat{\mu}_j) g'(\mu_j/m_j) \\ &= \widehat{\eta}_j + \frac{(d_j - \widehat{\mu}_j)}{2\, \sqrt{\mu_i}}, \end{aligned}$$

all quantities being computed at the initial estimate $\widehat{\boldsymbol{\eta}}^*$. Maximum likelihood estimates satisfy the equations

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Omega} \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Omega} \boldsymbol{z},$$

which are solved iteratively. The revised estimate is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{W\Omega X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W\Omega z}.$$

Finally, given the local log-likelihood (3.35), the deviance function is derived as

$$\begin{aligned} D &= 2\sum_{j=1}^{n} w_j \left\{d_j \log(d_j) - d_j - d_j \log(\mu_j) + \mu_j\right\} \\ &= 2\sum_{j=1}^{n} w_j \left\{d_j \log\left(\frac{d_j}{\mu_j}\right) - (d_j - \mu_j)\right\}. \end{aligned}$$

Again when the response is zero, the individual deviance function reduces to $D(d_j = 0) = 2\,\mu_j$.

3.6.3 Statistical Inference

Within the framework of normal approximation, the variance approximation (3.19) reduces to

$$\begin{aligned} \mathbb{V}\text{ar}\left[\widehat{\mu}_i\right] &= \mathbb{V}\text{ar}\left[g^{-1}(\widehat{\eta}(x_i))\right] \\ &\approx \left(\frac{\partial}{\partial \eta_i} g^{-1}(\widehat{\eta}(x_i))\right)^2 \omega_{ii}^{-1} \|\boldsymbol{s}(x_i)\|^2 \\ &= \widehat{\mu}_i \|\boldsymbol{s}(x_i)\|^2 \end{aligned} \tag{3.37}$$

With the plug-in bias $\widehat{\boldsymbol{b}}(\mu_i)$ from expression (3.16), we construct the pointwise confidence intervals adjusted to allow for bias as follows. From (3.23) the unknown function $\mu(x_i)$ falls in the random interval with approximately $(1-\alpha)$ coverage probability,

$$\widehat{\mu}(x_i) - \widehat{\boldsymbol{b}}(\mu_i) \pm c\, \widehat{\mu}_i^{1/2} \|\boldsymbol{s}(x_i)\|,$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution. We can also construct pointwise confidence intervals for $\varphi(x_i)$, the force of mortality, and for $\text{logit}(\varphi(x_i))$. A $(1-\alpha)100\,\%$ confidence interval for $\widehat{\varphi}(x_i)$ is

$$\widehat{\varphi}(x_i) - \widehat{\boldsymbol{b}}(\varphi_i) \pm c\, \mathbb{V}\text{ar}\left[\widehat{\mu}_i\right]^{1/2} L_i^{-1}$$
$$\text{or equivalently } \widehat{\varphi}(x_i) - \widehat{\boldsymbol{b}}(\varphi_i) \pm c\, \widehat{\mu}_i^{1/2}\, L_i^{-1} \|\boldsymbol{s}(x_i)\|.$$

A confidence interval for $\text{logit}(\widehat{\varphi}(x_i))$ is given by

$$\text{logit}\big(\widehat{\varphi}(x_i) - \widehat{\boldsymbol{b}}(\varphi_i)\big) \pm c\, \widehat{\mu}_i^{1/2} (L_i \widehat{\varphi}_i (1 - \widehat{\varphi}_i))^{-1} \|\boldsymbol{s}(x_i)\|.$$

3.6.4 Applications

We now consider non-parametric Poisson regression. We revisit the previous examples to examine the efficacy of the local Poisson approach. We graduate the mortality data through the choices of the smoothing parameters, using the *AIC* and *LCV* fitting criteria and graphical diagnostics to guide the modeling.

Choice of the constellation of the smoothing parameters

The pattern displayed by the observed number of deaths is relatively smooth for both datasets. The selected bandwidth should not be too large to miss the structure in the middle of the table. Table 3.7 displays the elected constellation of smoothing parameters for the local likelihood Poisson approach and for each dataset with the corresponding fitted degrees of freedom.

	λ	Degree	$W(.)$	Fitted DF
Dutch Male	19	3	Gaussian	15.93
Dutch Female	21	3	Gaussian	14.52

Table 3.7: *Elected constellation of smoothing parameters and fitted degrees of freedom*

Again whatever the volume of data, a local cubic fit is needed to capture the patterns displayed by the observed number of deaths. The choice differs by the elected bandwidth. A higher λ is then needed to smooth the structures for the female population which we believe less accentuated than the male. The corresponding fitted degrees of freedom for the female population are lower than the ones for the male, indicating that we have applied more smoothing.

Table 3.8 presents the theoretical bandwidth provided by the plug-in method developed in Section 3.4.2. We fit a polynomial of degree 3 and use the corresponding weight functions elected in Table 3.7. The values of λ are reported below.

	Pilot bandwidth	Bandwidth
Dutch Male	$\lambda = 9$	$\lambda = 17$
Dutch Female	$\lambda = 9$	$\lambda = 19$

Table 3.8: *Pilot and bandwidths selected by the plug-in method*

As the amount of curvature of the observed number of death is more or less similar for both datasets, the selected pilot bandwidths are the same. The bandwidths confirmed our choices presented in Table 3.7, being rather close and agreeing with our ranking.

Plots of the fits and residuals plots

Figure 3.6 presents the number of deaths estimated by the local Poisson likelihood approach, computed with the constellation of smoothing parameters displayed in Table 3.7.
Next to the plots of the fits, we present the corresponding residuals plots. Superimposed on the responses and Pearson residuals is a *loess* smooth curve.

The Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the model adequately captures the variability of these datasets. However, a clear lack of fit is shown by the *loess* smooth curve on the responses and Pearson residuals at the left boundary, which is confirmed by the deviance residuals. The deviance residuals present, for the youngest ages, several successive residuals having the same sign. It illustrates that the expected number of deaths is over-smoothed locally. As the sign is positive, we strongly underestimate the probability of death for the youngest ages as we can see on Figure 3.2.

We notice as well for both datasets a peak showing an important bias around attained age 60. This peak indicates departure of the graduated series from the observed number of deaths, which display, based on our knowledge, an abnormal hump certainly due to a cohort effect. This locally over-smoothed figure can be found in the deviance residuals as well, displaying several successive residuals having a positive sign. It illustrates that we underestimate the expected number of deaths around 60 years old.

For the male population the deviance residuals show, around attained age 80, several successive residuals having a negative sign. It indicates that we overestimate locally the expected number of death. For the female population, the deviance residuals exhibits, around attained age 85, several successive residuals having a positive sign. It illustrates that here we are underestimating the expected number of deaths. A smaller bandwidth or a higher polynomial degree could be used to capture the structure but it would be at the expense of a lack of smoothing in the middle of the table. The fit would be too noisy, and would stay to close to an interpolation since trends in small parts of the data are interpreted as more widespread trends. It would result in an unacceptably high variance.

Figure 3.7 shows the forces of mortality in the original and logit scale with the corresponding pointwise confidence intervals. As the constellations of smoothing parameters presented in Table 3.7 lead to relatively low fitted degrees of freedom compared to the binomial model, the patterns displayed by the graduated series are smoother.

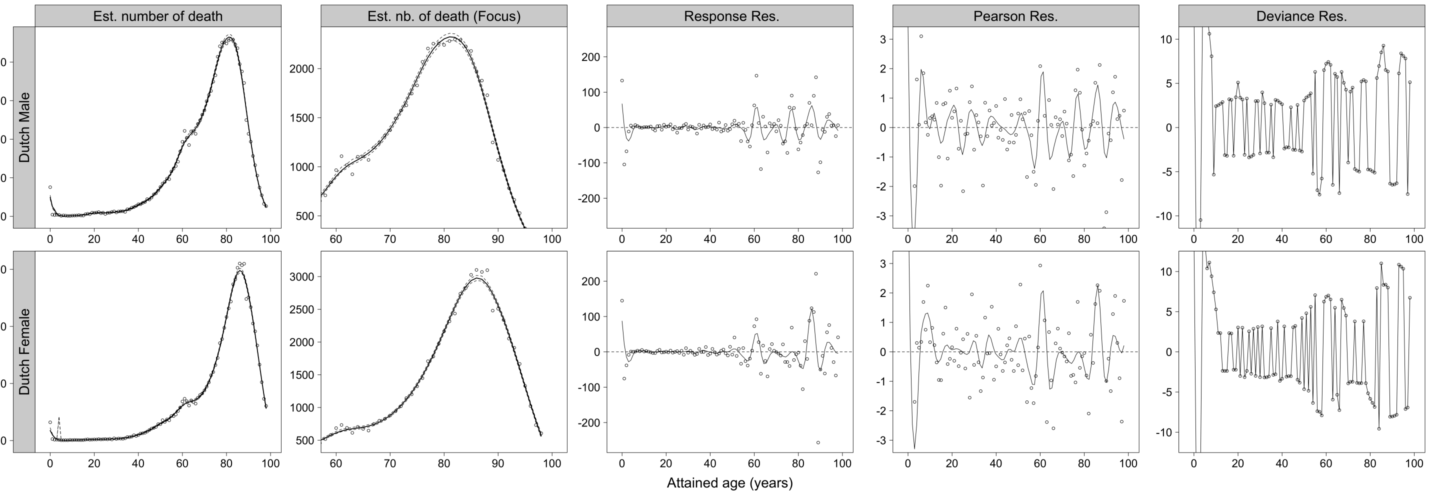


Figure 3.6: *Estimated number of death by local Poisson model with* 95 % *pointwise confidence intervals and corresponding residuals plots for Dutch Male and Female population, 2008. Source: HMD.*

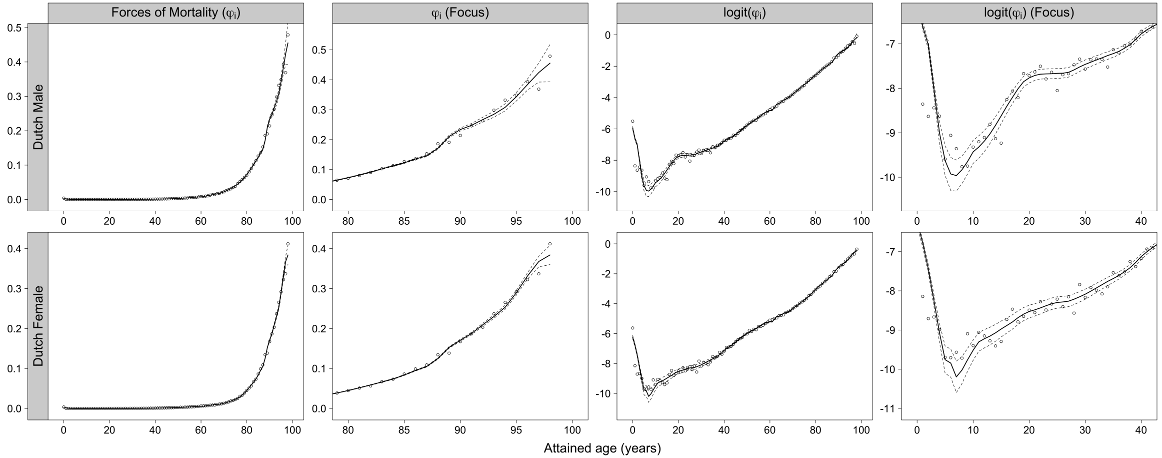


Figure 3.7: *Graduated forces of mortality and Logit transform of the graduated forces of mortality by local Poisson model with 95 % pointwise confidence intervals for Dutch Male and Female population, 2008. Source: HMD.*

Plots of the smoothers

We provide an illustration of the amount of smoothing applied similarly to the local binomial case. Figures 3.8 and 3.9 show the weight function associated with the i-th point at the last iteration with the influence values.

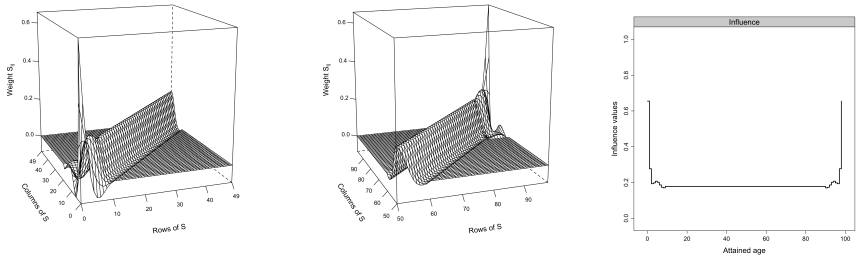


Figure 3.8: *Smoother S_{ij}: left panel: $i, j = 0, \ldots, 49$, center panel: $i, j = 50, \ldots, 98$ and influence values for the Dutch Male population, 2008. Source: HMD.*

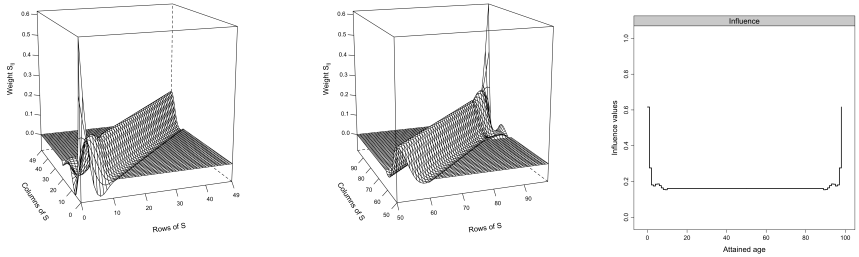


Figure 3.9: *Smoother S_{ij}: left panel: $i, j = 0, \ldots, 49$, right panel: $i, j = 50, \ldots, 98$ and influence values for the Dutch Female population, 2008. Source: HMD.*

Similar remarks as for the local binomial model can apply. For values in the central region, the weights form a Gaussian kernel. But as the point at which we are estimating the true curve moves towards the boundaries, the kernel overlaps the boundary, becomes asymmetric and leads to some weights being negative. Finally, the height of the kernel increases because fewer observations are available. In Figure 3.8 right panel, the influence values at the boundaries are lower than the ones in the central region. It indicates that locally we have applied more smoothing in the boundaries than to the rest of the curve.

However, contrary to the local binomial model, we notice that the influence values in the central region are constant. It is explained by the use of the normal approximation and the variance stabilizing link. It leads to approximating ω_{jj} by a constant, namely 4, while for the local binomial model we have $4 \times l_j$.

Comparison with Whittaker-Henderson smoothing

Similarly we apply the criteria presented in Section 2.5.1 to find the value of parameters h and z. We picked the constellation $h = 146$ and $z = 3$ for the male, and $h = 82$ and $z = 3$ for the female population, given by Rice's T criterion, Rice (1984), leading to 13.26 and 13.92 fitted degrees of freedom respectively. Figure 3.10 presents graphical comparisons of the local Poisson model and the Whittaker-Henderson method. The left panels present the estimated number of deaths for the two populations. The graduated series by the local Poisson model are showing more features. The corresponding degrees of freedom are larger than the ones obtained by the Whittaker-Henderson model.

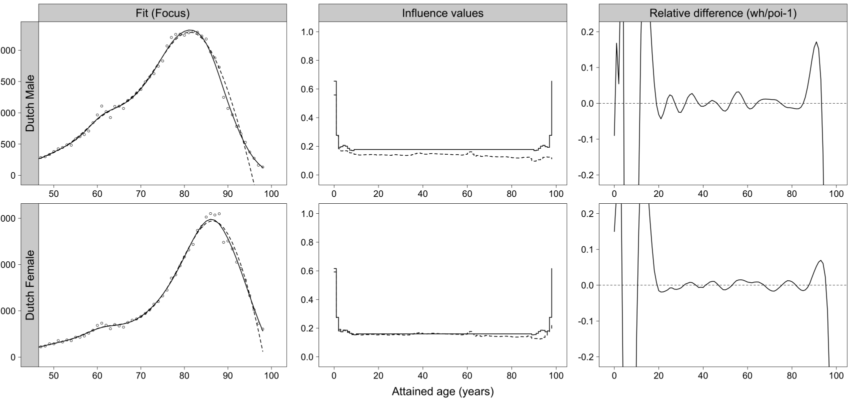


Figure 3.10: *Graphical comparisons between the local Poisson model (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008. Source: HMD.*

The influence values, obtained by the local Poisson model for the male population, are above the ones computed with the Whittaker-Henderson model, top center panel. It indicates that less smoothing has been applied by the local Poisson approach. The relative difference is more important at the boundaries, where no special treatment is needed when using the Whittaker-Henderson model.

For the female population, the fitted degrees of freedom are close, illustrating that the models sensibly show the same amount of features. The influence values, bottom center panel, stay close up to attained age 60. Then, the influence values obtained by the Whittaker-Henderson smoothing are slightly lower, indicating that more smoothing has been applied. Finally, just as for the male population, the relative difference is larger at the boundaries.

3.7 Summary and outlook

We have investigated the extension of the non-parametric regression technique of local polynomials to localized generalized linear models and local likelihood contexts. In the ordinary regression case, fitting by local polynomials has been seen to have several appealing features in terms of intuitive and mathematical simplicity. This is especially true for low odd-degree polynomial fits, such as linears and cubics. These properties have been shown to carry over to localized generalized linear models.

We have seen that the extension to local likelihood settings overcomes the problems encountered while applying the local polynomial regression to graduation of experience data. Local likelihood has been introduced as a method of smoothing by local polynomials in non-Gaussian regression models. A local Binomial likelihood model has been proposed when the number of initial policyholders exposed to risk is available, and a local Poisson likelihood model for those central exposed to risk. The variance stabilizing link has been used to produce confidence intervals not depending on the estimates, and provide an illustration of the uncertainty involved even in the presence of zero-responses.

An important issue that will receive further attention in the next chapter is a locally adaptive graduation method. In graduating mortality data, we face a situation where data have a low noise and a large amount of structure. We have seen that is possible that no global smoothing parameter or degree of local polynomial will provide an adequate fit to the data. In this case, it may be desirable to use locally adaptive smoothing methods, which vary the amount of smoothing in a location dependent manner, so as to obtain a satisfactory fit.

Chapter 4

Adaptive local kernel-weighted log-likelihood methods

This chapter is based on Tomas and Planchet (2012), Multidimensional smoothing by adaptive local kernel-weighted log-likelihood with application to long-term care insurance, *ISFA - Laboratoire SAF Working paper 2012.8, submitted to Insurance: Mathematics & Economics*, , 1-28; and on Tomas (2012b), Essays on boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models, *Insurance and Risk Management*, forthcoming.

4.1 Introduction

Local fitting techniques combine excellent theoretical properties with conceptual simplicity. They are very adaptable and also convenient statistically. In Chapter 3, we have seen the applicability of local kernel-weighted log-likelihoods to model the relation between the forces of mortality - or the crude death rates - and attained age.

Unfortunately, as we face situations where data have a low noise and a large amount of structure, the simplicity of a local modeling has flaws. For instance, at the boundary, the smoothing weights are asymmetric and the estimate may have substantial bias. Bias can be a problem if the regression function has a high curvature in the boundary. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to under-smoothing in the middle of the table, see Section 4.2.

As a consequence, in some cases no global smoothing parameter or degree of local polynomial provides an adequate fit to the data. Rather than restricting the smoothing parameters to a fixed value, a more flexible approach is to allow the constellation of smoothing parameters to vary across the observations.

We can restrict the observations contributing to the criteria to the central region and apply weights according to the reliability of the data to enhance the optimization criteria and refine the choice of the smoothing parameters. But weighting the criteria is an illustration for a need of an adaptive smoothing procedure. Rather than weighting the criteria and restricting the observations to the central region, we would use a more flexible approach. It would be to vary the amount of smoothing in a location dependent manner and to allow adjustment based on the reliability of the data. It may be advantageous for several reasons. The estimator can adapt the reliability of the data to take into account the nature of the risk, smoothing more when the volume of observations is low, and less when the corresponding amount of observations is large. We distinguish a locally adaptive smoothing pointwise method using the intersection of confidence intervals rule, as well as a global method using local bandwidth correction factors. Part of our work is an extension of the adaptive kernel methods proposed by Gavin *et al.* (1995) to adaptive local kernel-weighted log-likelihoods techniques. The techniques can be implemented without difficulty in standard statistical software such as R, R Development Core Team (2012).

Section 4.2 presents the motivation for adaptive smoothing, studying the influence of the boundaries on the choice of the smoothing parameters and the possibility of taking the nature of the risk into account. The adaptive methods are introduced in Section 4.3. Section 4.4 illustrates how the methods can be applied to multidimensional smoothing. We are interested in the variation of mortality of individuals subscribing long-term care insurance. We analyze the incidence of mortality as a function of both the age of occurrence of the pathology and the duration of the care. Tests and single indices summarizing the lifetime probability distribution are used to compare the graduated series with those obtained from global non-parametric approaches, p-splines and local likelihood, in Section 4.5. Finally, Section 4.6 summarizes the conclusions drawn in this chapter.

4.2 Motivations for an adaptive smoothing

4.2.1 Influence of the boundaries on a global criterion

Table 4.1 presents the proportion of the contribution to the AIC criterion in the boundaries given by the local likelihood models for the Dutch male and female data, computed with a triweight weight function and $\lambda = 19$, see Section 2.3. It is apparent that the contribution varies with the underlying structure of the data. The mortality patterns of females are less pronounced than those of the males, and thus the resulting contribution to the criterion is smaller. The local Poisson model is less influenced by the boundaries than the local Binomial model as most of the curvature appears in the central region.

		Local Poisson model						Local Binomial model					
		Male pop.			Female pop.			Male pop.			Female pop.		
Treatment	p	Left	Right	All	Left	Right	All	Left	Right	All	Left	Right	All
type 1	2	49.21	12.82	62.03	45.07	14.64	59.71	78.48	2.53	81.01	73.20	4.52	77.72
	3	45.22	12.91	58.13	32.99	20.77	53.76	71.87	2.95	74.82	60.59	7.20	67.79
	4	38.42	15.15	53.57	26.54	25.48	52.02	63.31	3.41	66.72	48.74	8.77	57.51
type 2	2	35.45	16.29	51.74	29.22	18.86	48.08	23.99	8.95	32.94	16.67	14.01	30.68
	3	27.86	17.00	44.86	19.37	24.99	44.36	14.33	8.99	23.32	8.80	16.65	25.45
	4	19.00	19.94	38.94	12.40	30.38	42.78	8.90	8.47	17.37	4.66	16.31	20.97
type 3	2	1.14	25.92	27.06	0.94	26.37	27.31	4.06	11.32	15.38	2.76	16.46	19.22
	3	1.42	23.27	24.69	0.98	30.72	31.7	4.10	10.09	14.19	2.62	17.81	20.43
	4	1.56	24.31	25.87	1.14	34.30	35.44	4.17	8.93	13.1	2.61	16.67	19.28

Table 4.1: *Contribution to the AIC in the boundaries (in %), computed with a triweight weight function and $\lambda = 19$, for the Dutch male and female population, 2008. Source: HMD.*

Correction *type 1* leads to the highest contributions to the AIC. This treatment induces the highest amount of smoothing in the boundaries and thus leads to the highest disturbance when choosing the constellation of smoothing parameters. When summing the contribution coming from the left and right boundaries, we observe that the boundaries represent at least 52.02 % and 57.51 % to 62.03 % and 81.01 % of the AIC, respectively, for the local Poisson and Binomial models. It is obvious that the selection of the smoothing parameter is driven by minimizing the criterion in the boundaries rather than for the whole set of data points.

The disturbance is reduced when treatment *type 2* is used. However the contribution to the AIC is still relatively high with at most 51.74 % and 32.94 %, for the local Poisson and Binomial models respectively.

Correction *type 3* implies smooth weights having the smallest dispersion around the central value. In consequence, it leads to the smallest disturbing nuisance. The contribution to the criterion for observations in the left boundary has strongly reduced while the contribution in the right boundary has inflated. This type of treatment leads to under-smoothed figures in the left boundary and its merit would depend on the underlying smoothness of the data.
It shows the resulting difficulty of applying a global smoothing approach when the true curve presents rapid changes in the curvature.

A solution for a homogeneous contribution of the design space to the criterion would be to modify the AIC by taking the logarithm of the deviance or weighting the criterion by the variance of the fitted values. It would lead, as for the criteria for linear smoothers, to a reduction in the variability, and the criterion would be less affected by the boundaries. Further, we consider restricting the computation of the criteria to observations in the central region and study where the contribution to these criteria are coming from in the design space.

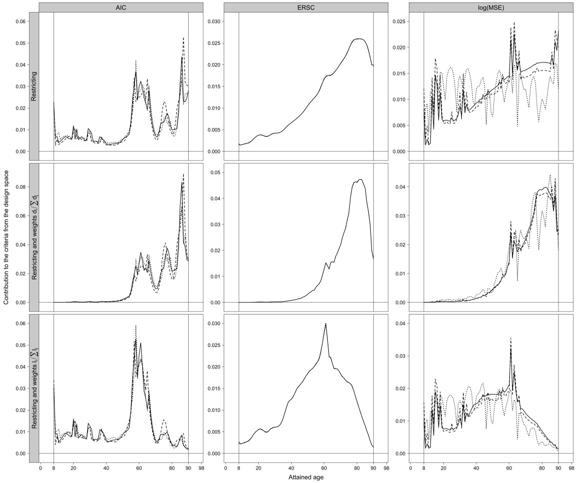


Figure 4.1: *Pointwise contribution to the criteria when restricting and weighting the observations for the local Poisson model targeting the number of deaths, d_i, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

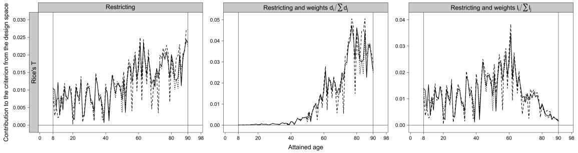


Figure 4.2: *Pointwise contribution to Rice's T criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the number of deaths, d_i, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

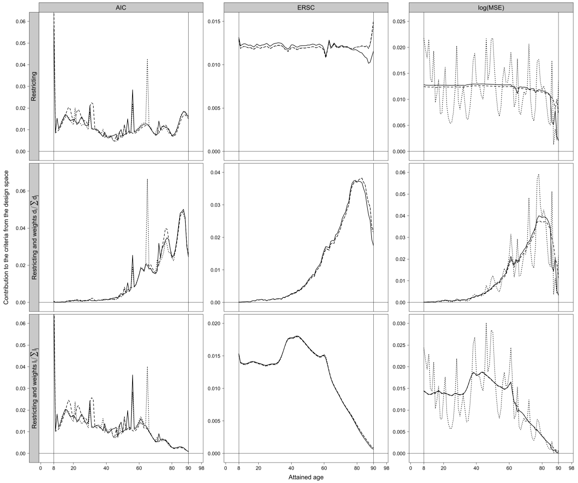


Figure 4.3: *Pointwise contribution to the criteria when restricting and weighting the observations for the local Binomial model targeting the mortality rate, q_i, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

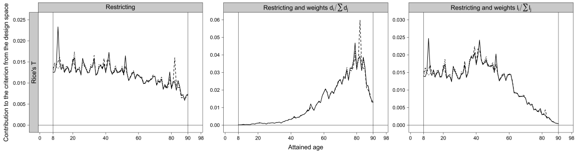


Figure 4.4: *Pointwise contribution to Rice's T criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the mortality rate, q_i, Dutch male population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

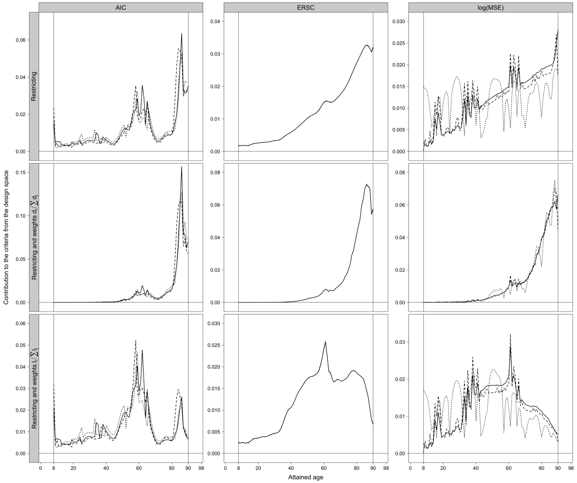


Figure 4.5: *Pointwise contribution to the criteria when restricting and weighting the observations for the local Poisson model targeting the number of deaths, d_i, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

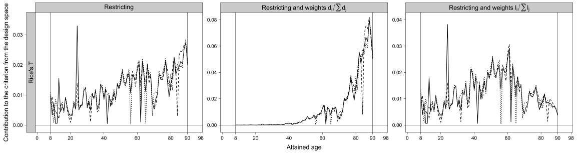


Figure 4.6: *Pointwise contribution to Rice's T criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the number of deaths, d_i, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

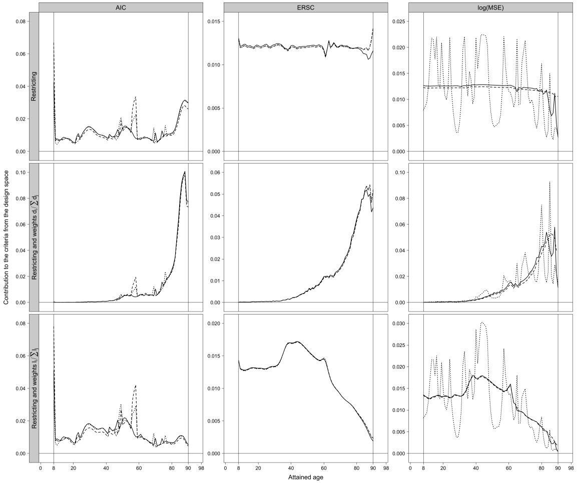


Figure 4.7: *Pointwise contribution to the criteria when restricting and weighting the observations for the local Binomial model targeting the mortality rate, q_i, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

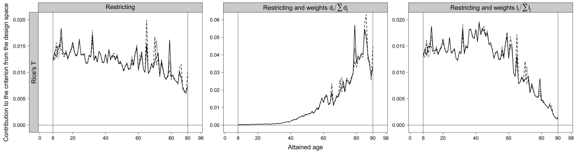


Figure 4.8: *Pointwise contribution to Rice's T criterion when restricting and weighting the observations for the Whittaker-Henderson model targeting the mortality rate, q_i, Dutch female population, 2008. Quadratic fit (dashed line), cubic fit (full line) and quartic fit (dotted line). Source: HMD.*

Restricting the observations participating in the computation of the criteria helps to reduce the boundary effects, see Fan *et al.* (1998). At the boundaries, the pointwise contributions are too large because of numerical instabilities, underlying structure and scarcity of the data. Figures 4.1, 4.3 and 4.5, 4.7, first row, show the pointwise contributions to the criteria when restricting the contribution to observations in the central region for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population.
The pointwise contributions to the criteria differ due to the underlying structure of the data as the mortality patterns are more pronounced for the male than the female population. We observe that observations around age 18, corresponding to the *accident hump*, as well as observations around 60, corresponding to a cohort effect, contribute more to the criteria for the male population when fitting both of the local likelihood models. By fitting the local Poisson model, we notice an increase of the pointwise contribution with the number of deaths. This is particularly visible for the $ERSC$ and $\log(MSE)$. On the other hand, in case of a local Binomial model, the pointwise contribution to the $ERSC$ and $\log(MSE)$ tends to decrease as the curvature of the observed mortality rates increases.

These features can be also seen in the pointwise contribution to Rice's T criterion used for linear smoothing, shown in Figures 4.2, 4.4 and 4.6, 4.8, for the Whittaker-Henderson model targeting the number of deaths and the mortality rates on the original scale, for the Dutch male and female population respectively.
In graduating the mortality rates, however, the decrease of the pointwise contribution with the increasing curvature can be a problem. It may force the criterion to select a larger bandwidth and this may lead to over-smoothing at the end of the table. It results in underestimating the mortality rates and in missing the mortality pattern of the oldest ages.

4.2.2 The nature of the risk

In practice, the search for an optimal criterion depends not only on statistical considerations but also on the nature of the risk considered. A smoothing method well suited for annuities may not be suited for death benefits. In the first case, we have to represent effectively the remaining life expectancy in the regions where the exposure is high. In the second case, we have to represent the observed deaths well where the number of deaths is large, and these regions may not necessarily be those where there is more exposure, such as the female population. Therefore, the criteria can be weighted according to the nature of the risk considered to refine the choice of the smoothing parameters:

i. by $l_i/\sum_j l_j$ in case of annuities, and

ii. by $d_i/\sum_j d_j$ in case of death benefits.

Table 4.2 presents the contribution to the criteria for observations in the age range representing 80 % of the exposure and number of deaths for the Dutch male and female population after weighting the criteria according to the nature of the risk considered.

				Local Poisson			Local Binomial		
Population	Age range	l_i	Rice's T	AIC	$ERSC$	$\log(MSE)$	AIC	$ERSC$	$\log(MSE)$
Male	8-67	80	91.27	85.85	72.87	82.92	90.06	89.20	89.84
Female	8-70	80	90.38	81.86	67.98	79.54	84.74	88.36	86.40
Population	Age range	d_i	Rice's T	AIC	$ERSC$	$\log(MSE)$	AIC	$ERSC$	$\log(MSE)$
Male	59-90	80	89.76	90.07	93.09	89.91	86.83	85.05	83.49
Female	46-90	80	98.53	99.26	99.34	98.52	98.21	96.34	96.28

Table 4.2: *Contribution to the criteria (in %) for observations in the age range representing* 80 % *of the exposure and number of deaths for the Dutch male and female population, 2008. Computed with a cubic fit and a triweight weight function. Source: HMD.*

For the male population, 80 % of the exposure appears in the age range $8-67$. For the female population the age range corresponds to $8-70$. In case of annuities, by weighting by $l_i/\sum_j l_j$ most of the criteria applied to the local Poisson model (force of mortality) and to the Binomial model (probability of death) provides a good representation. The contribution to these criteria, for observations in the age range considered, are mostly above 80 %. Only the $ERSC$ provides a poor representation when fitting the local Poisson model, due to the distribution of the criterion following broadly the observed number of deaths.

For the male and female population, 80 % of the deaths appears in the age ranges $59-90$ and $46-90$, respectively. In case of death benefits, by weighting the criteria by $d_i/\sum_j d_j$, the proportion of the contributions from observations in the age range are above 80 % showing a good representation of the risk considered.

For linear smoothers, the representation of the risk given by Rice's T and variations of the classical criteria is satisfactory.

In consequence, weighting the criteria by the reliability of the data leads to a better representation of the nature of the risk considered whatever the model used for graduating the forces of mortality or the probabilities of death.

Figures 4.1, 4.3 and 4.5, 4.7, second and third row, show the pointwise contribution to the criteria for the local Poisson model and the local Binomial model respectively, for the Dutch male and female population, when restricting the contribution to observations in the central region and weighting according the reliability of the data.

We have restricted the observations contributing to the criteria to the central region and applied weights according to the reliability of the data.

These practical considerations enhance clearly the optimization criteria, and the choice of the constellation of the smoothing parameters is refined, leading to a good representation of the risk considered.

It should be noted that graduating mortality data through the Whittaker-Henderson model by selecting the parameters through the log transform of classical criteria, see Section 2.5, performs relatively well, and taking into account the nature of the risk improves the smoothing.

Weighting the criteria according to the reliability of the data illustrates the need of an adaptive smoothing procedure. Rather than weighting the criteria and restricting the observations to the central region, we would use a more flexible approach. It would be to allow the constellation of smoothing parameters to vary across the observations to vary the amount of smoothing in a location dependent manner. It would allow adjustments based on the reliability of the data and on the nature of the risk considered. It may be advantageous for several reasons. The estimator could adapt to the reliability of the data to take into account the nature of the risk, smoothing more when the volume of observations is low, and less when the corresponding amount of observations is large.

4.3 Adaptive Methods

This section presents the adaptive methods and covers model selection issues. We treat the choices of bandwidth, polynomial degree and weight function as modeling the data and choose the constellation of smoothing parameters to balance the trade-off between bias and variance. We distinguish a locally adaptive pointwise smoothing method using the intersection of confidence intervals rule and a global method using local bandwidth correction factors. We vary the amount of smoothing in a location dependent manner and allow adjustments based on the reliability of the data.

It is well known that of the smoothing parameters, the weight function has much less influence on the bias and variance trade-off than the bandwidth or the order of approximation. The choice is not too crucial, at best it changes the visual quality of the regression curve. For convenience, we use the Epanechnikov weight function in expression (4.6) throughout this chapter, as it is computationally cheaper to use a truncated kernel. Moreover, it has been shown that the Epanechnikov kernel is optimal in minimizing the mean squared errors for local polynomial regression, see Fan *et al.* (1997). The biweight and triweight kernel, which behave very similarly, could have also been chosen. The choice remains subjective.

The data used for the following illustrations are presented in Section 4.4.4. In brief, the data come from observations of individuals subscribing to Long-Term Care (LTC) insurance policies originating from a portfolio of a French insurance company. We focus on measuring the forces of mortality as a function of the age v of occurrence of the pathologies and the duration u of the care.

4.3.1 Intersection of confidence intervals

The intersection of confidence intervals was introduced by Goldenshulger and Nemirovski (1997) and further developed by Katkovnik (1999). Application of the ICI rule in case of Poisson local likelihood for adaptive scale image restoration has been studied in Katkovnik *et al.* (2005). Chichignoud (2010) in his Ph.D. thesis presents a comprehensive illustration of the method. The intersection of confidence intervals (ICI) provides an alternative method of assessing local goodness of fit.
We start by defining a finite set of window sizes

$$\Lambda = \{\lambda_1 < \lambda_2 < \ldots < \lambda_K\},$$

and determines the optimal bandwidth by evaluating the fitting results. Let $\widehat{\psi}(x_i, \lambda_k)$ be the estimate at x_i for the window λ_k. To select the optimal bandwidth, the ICI rule examines a sequence of confidence intervals of the estimates $\widehat{\psi}(x_i, \lambda_k)$:

$$\begin{aligned}
\widehat{I}(x_i, \lambda_k) &= \left[\widehat{L}(x_i, \lambda_k), \widehat{U}(x_i, \lambda_k)\right], \\
\widehat{U}(x_i, \lambda_k) &= \widehat{\psi}(x_i, \lambda_k) + c\,\widehat{\sigma}(x_i)\|\boldsymbol{s}(x_i, \lambda_k)\|, \\
\widehat{L}(x_i, \lambda_k) &= \widehat{\psi}(x_i, \lambda_k) - c\,\widehat{\sigma}(x_i)\|\boldsymbol{s}(x_i, \lambda_k)\|,
\end{aligned}$$

where c is a threshold parameter of the confidence interval. Then, from the confidence intervals, we define

$$\begin{aligned}
&\overline{\widehat{L}}(x_i, \lambda_k) = \max\left[\overline{\widehat{L}}(x_i, \lambda_{k-1}), \widehat{L}(x_i, \lambda_k)\right], \\
&\underline{\widehat{U}}(x_i, \lambda_k) = \min\left[\underline{\widehat{U}}(x_i, \lambda_{k-1}), \widehat{U}(x_i, \lambda_k)\right], \\
&k = 1, 2, \ldots, K \quad \text{and} \quad \overline{\widehat{L}}(x_i, \lambda_0) = \underline{\widehat{U}}(x_i, \lambda_0) = 0.
\end{aligned}$$

The largest value for these k for which $\underline{\widehat{U}}(x_i, \lambda_k) \geq \overline{\widehat{L}}(x_i, \lambda_k)$ gives k^*, and it yields a bandwidth λ_k^*, that is the required optimal ICI bandwidth.

In other words, denoting $\mathcal{I}_j = \bigcap_{j=k}^{K} \widehat{I}(x_i, \lambda_j)$ for $k = 1, 2, \ldots, K$, we choose k^* such that

$$\begin{cases}
\mathcal{I}_j \neq \emptyset, & \forall j \geq k^*, \\
\mathcal{I}_{k^*-1} = \emptyset.
\end{cases}$$

As the bandwidth λ_k is increased, the standard deviation of $\widehat{\psi}(x_i, \lambda_k)$, and hence $\|\boldsymbol{s}(x_i, \lambda_k)\|$, decreases. The confidence intervals become narrower. If λ_k is increased too far, the estimate $\widehat{\psi}(x_i, \lambda_{k)}$ will become heavily biased,

and the confidence intervals will become inconsistent in the sense that the intervals constructed at different bandwidths have no common intersection. The optimal bandwidth λ_{k^*} is the largest k when $\underline{\widehat{U}}(x_i, \lambda_k) \geq \overline{\widehat{L}}(x_i, \lambda_k)$ is still satisfied, i.e. when $\mathcal{I}_j \neq \emptyset$.
Because the optimal bandwidth is decided by c, this parameter plays a crucial part in the performance of the algorithm. When c is large, the segment $\widehat{I}(x_i, \lambda_k)$ becomes wide and it leads to a larger value of λ_k^*. This results in over-smoothing. On the contrary, when c is small, the segment $\widehat{I}(x_i, \lambda_k)$ would become narrow and it leads to a small value of λ_k^* so that the volatility can not be removed effectively. In theory, we could apply the criteria presented in Section 3.4 to determine a reasonable value c. However, because of practical constraints, the choice of c is done subjectively.

4.3.2 Local bandwidth factor methods

Instead of having a pointwise procedure, other types of adaptive approaches could be performed by using a global criterion. We could incorporate additional information into a global procedure by allowing the bandwidth to vary according to the reliability of the data, such as the variable kernel estimator proposed in Gavin *et al.* (1995, pp.190-193). We can calculate a different bandwidth for each age at which the curve has to be estimated. The local bandwidth at each age is simply the global bandwidth multiplied by a local bandwidth factor to allow explicit dependence on this information. As we already obtained the local bandwidth factors, the process of using a global criterion decides the global value at which the bandwidth curve is located.

The aim is to allow the bandwidth to vary according to the reliability of the data, and to take into account the nature of the risk considered. The local bandwidth factors could depend on the exposure or the number of deaths per attained age, in case of annuities and death benefits, respectively. For regions in which the exposure is large, a low value for the bandwidth results in an estimate that more closely reflects the crude rates. On the other hand, for regions in which the exposure is small, such as long duration, a higher value for the bandwidth allows the estimate of the true forces of mortality to progress more smoothly. This means that for long duration we are calculating local averages over a greater number of observations, which reduces the variance of the graduated rates but at the cost of a potentially higher bias.

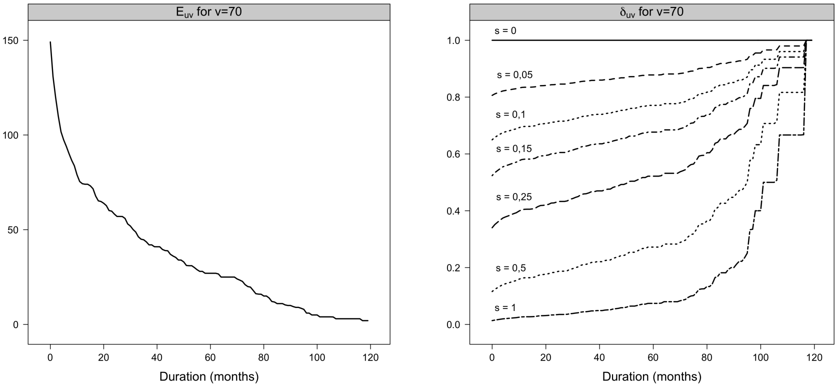
The local bandwidth at each age is the global bandwidth multiplied by a local bandwidth factor, $h_i = h \times \delta_i^s$ for $i = 1, \ldots, n$. The variation in exposure or in deaths within a dataset can be enormous. To dampen the effect of this variation we choose

$$\delta_i^s \propto \widehat{\xi}_i^{-s}, \qquad \text{for } i = 1, \ldots, n \text{ and } 0 \leq s \leq 1, \tag{4.1}$$

where s is a sensitivity parameter and

$$\widehat{\xi}_i = \begin{cases} E_i / \sum_{j=1}^{n} E_j & \text{for } i = 1, \ldots, n \quad \text{in case of annuities,} \\ d_i / \sum_{j=1}^{n} d_j & \text{for } i = 1, \ldots, n \quad \text{in case of death benefits.} \end{cases} \tag{4.2}$$

Choosing $s = 0$ reduces both models to the fixed parameter case, while $s = 1$ may result in very large smoothness variation depending on the particular dataset. We choose the reciprocal of $\max\{\xi_i^{-s};\ i = 1, \ldots, n\}$ as the constant of proportionality in (4.1), so that $0 < \delta_i^s \leq 1$, for $i = 1, \ldots, n$. The observed exposure, or the observed deaths, decides the shape of the local bandwidth factor but the sensitivity parameter s determines the magnification of that shape, becoming more pronounced as s tends to 1. Figure 4.9a shows the exposure for the age of occurrence 70 and Figure 4.9b displays the resulting smoothness tuning parameter for values of the sensitivity parameter of $0, 0.05, 0.1, 0.15, 0.25, 0.5$ and 1.



(a) *Exposure* $E_{u,70}$

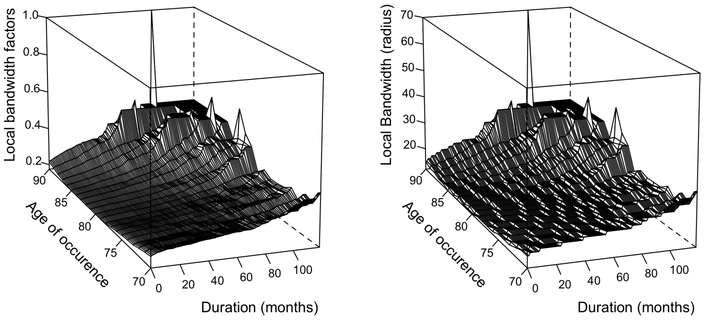
(b) *Local bandwidth factors* $\delta_{u,70}$

Figure 4.9: $E_{u,70}$*, and values of* $\delta_{u,70}$*, for various sensitivity parameters.*

For $s = 0.15$, the minimum smoothness tuning parameter is about 0.5, at duration 0. This means that the bandwidth at the longest duration is about twice that at the shortest duration.

Figure 4.10 presents the value of $\delta_{u,v}$ for $s = 0.15$ and local bandwidth values (radius) derived. If there is a small exposure, then $\delta_{u,v}^s$ is large. It increases the smoothness tuning parameter and allows to apply more smoothing. The other way around if the amount of exposure is large.

Similarly to the global approach, we can apply the criteria presented in Section 3.4 to select the optimal constellation of smoothing parameters. As



(a) *Local bandwidth factors* $\delta_{u,v}$ **(b)** *Local bandwidths (radius)*

Figure 4.10: $\delta_{u,v}$ *for* $s = 0.15$ *and the resulting local bandwidths.*

we already obtained the shape and the magnification of the local bandwidth factors, this process decides the global value at which the bandwidth curve is located.

4.4 Application

To illustrate the adaptive local kernel-weighted log-likelihood approaches, we discuss an application concerning the mortality of individuals having a long-term care (LTC) insurance contract. LTC is a mix of social and health care provided on a daily basis, formally or informally, at home or in institutions, to people suffering from a loss of mobility and autonomy in their activity of daily living. Although loss of autonomy may occur at any age, its frequency rises with age. LTC insurance contracts are individual or collective and guarantee the payment of a fixed allowance, in the form of monthly cash benefit, possibly proportional to the degree of dependency, see Kessler (2008) and Courbage and Roudaut (2011) for studies on the French LTC insurance market.

Most of the actuarial publications on this topic focus on the construction of models of projected benefits, see Gauzère *et al.* (1999) and Deléglise *et al.* (2009). Here we are concerned about the construction of the survival distribution of LTC insurance policyholders. The pricing and reserving as well as the management of LTC portfolios are very sensitive to the choice of the mortality table adopted. In addition, the construction of such table is a difficult exercise due to the following features:

i. French LTC portfolios are relatively small and the estimation of crude death rates is very volatile;

ii. because of the strong link between the age at subscription of LTC insurance policy and the related pathology, it is usual to construct a mortality table based on both age of occurrence of the pathologies, *which is an explanatory variable*, and duration of the care (or seniority), *which is the duration variable.* Hence, it is necessary to construct a mortality surface;

iii. mortality rates decrease very rapidly with the duration of the care. In consequence, the first year is often difficult to integrate into the usual (parametric) models.

Thus practitioners often use empirical methods that rely heavily on experts opinion. We therefore propose, in this chapter, more rigorous methods for the graduation of mortality tables not depending on experts advice.

4.4.1 Analysis of the changes in mortality

We analyze the changes in mortality of individuals subscribing LTC insurance policies as a function of both the duration of the care and the age of occurrence of the pathology. Let $T_u(v)$ be the remaining lifetime of an individual when the pathology occurred at age v, for the duration of the care u, with v and u being integers. We are working with two temporal dimensions u and v, however, they do not have the same status: v is a variable denoting the heterogeneity while u represents the variable linked with the duration. The distribution function of $T_u(v)$ is denoted as ${}_\tau q_u(v) = \Pr\left[T_u(v) \leq \tau\right] = 1 - {}_\tau p_u(v)$. The force of mortality at duration $u+\tau$ for the age of occurrence v, denoted by $\varphi_{u+\tau}(v)$ is defined by

$$\varphi_{u+\tau}(v) = \lim_{\Delta\tau \to 0^+} \frac{\Pr\left[\tau < T_u(v) \leq \tau + \Delta\tau | T_u(v) > \tau\right]}{\Delta\tau} = \frac{1}{{}_\tau p_u(v)} \frac{\partial}{\partial\tau} {}_\tau q_u(v),$$

$$\text{and,}\quad {}_\tau p_u(v) = \exp\left(-\int_0^\tau \varphi_{u+\xi}(v+\xi)\, d\xi\right).$$

We assume that the duration-specific forces of mortality are piecewise constant in each unit square, but allowed to vary from one unit square to the next, $\varphi_{u+\tau}(v+\xi) = \varphi_u(v)$ for $0 \leq \tau < 1$ and $0 \leq \xi < 1$. Under this assumption, $p_u(v) = \exp(-\varphi_u(v)) \Leftrightarrow \varphi_u(v) = -\log(p_u(v))$.

We define the exposure-to-risk $(E_{u,v})$, measuring the time during which individuals are exposed to the risk of dying. It is the total time lived by these individuals. Assume that we have $L_{u,v}$ individuals at duration u and age of occurrence v. Using the notation of Gschlössl *et al.* (2011), we associate to each of these $L_{u,v}$ individuals the dummy variable

$$\delta_i = \begin{cases} 1 & \text{if individual } i \text{ dies,} \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, 2, \dots, L_{u,v}$. We define the time lived by individual i before $(u+1)$st duration when the pathology occurred at age v by τ_i. We assume that we have at our disposal iid observations (δ_i, τ_i) for each of the $L_{u,v}$ individuals. The contribution of individual i to the likelihood equals $\exp(-\tau_i \varphi_u(v))(\varphi_u(v))^{\delta_i}$. Finally we define

$$\sum_{i=1}^{L_{u,v}} \tau_i = E_{u,v} \quad \text{and} \quad \sum_{i=1}^{L_{u,v}} \delta_i = D_{u,v}.$$

Under these assumptions, the likelihood becomes

$$\mathcal{L}(\varphi_u(v)) = \prod_{i=1}^{L_{u,v}} \exp(-\tau_i \varphi_u(v))(\varphi_u(v))^{\delta_i} = \exp(-E_{u,v}\ \varphi_u(v))(\varphi_u(v))^{D_{u,v}}.$$

The associated log-likelihood is $\ell(\varphi_u(v)) = \log \mathcal{L}(\varphi_u(v)) = -E_{u,v}\ \varphi_u(v) + D_{u,v} \log \varphi_u(v)$. Maximizing the log-likelihood $\ell(\varphi_u(v))$ gives $\widehat{\varphi}_u(v) = D_{u,v}/E_{u,v}$ which coincides with the central death rates $\widehat{m}_u(v)$. Then it is apparent that the likelihood $\ell(\varphi_u(v))$ is proportional to the Poisson likelihood based on $D_{u,v} \sim \text{Poisson}(E_{u,v}\varphi_u(v))$ and it is equivalent to work on the basis of the *true* likelihood or on the basis of the Poisson likelihood, as recalled in Gschlössl *et al.* (2011). Thus, under the assumption of constant forces of mortality between non-integer values of u and v, we consider

$$D_{u,v} \sim \text{Poisson}(E_{u,v}\varphi_u(v)), \tag{4.3}$$

to take advantage of the Generalized Linear Models (GLMs) framework.

4.4.2 Bi-dimensional local likelihood

We present briefly the generalization to two predictors. Suppose we have n independent realizations $y_1, y_2, \dots, y_n$ of the random variable Y with

$$Y_i \sim f(Y|\theta(x_i)), \quad \text{for } i = 1, 2, \dots, n,$$

where $f(\cdot|\theta(x_i))$ is a probability mass/density function in the exponential dispersion family and $\theta(x_i)$, the natural parameter in the GLMs framework, is an unspecified smooth function $\psi(x_i)$. For simplicity, we use $x_i = (u_i, v_i)$ to denote the vector of the predictor variables. The bivariate local likelihood fits a polynomial model locally within a bivariate smoothing window. Suppose that the function ψ has a $(p+1)$st continuous derivative at the point $x_i = (u_i, v_i)$. For data point $x_j = (u_j, v_j)$ in a neighborhood of $x_i = (u_i, v_i)$ we approximate $\psi(x_j)$ via a Taylor expansion by a polynomial of degree p. If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares and their cross-products. For example, a local quadratic approximation is:

$$\begin{aligned}\psi(x_j) =& \psi(u_j, v_j) \approx \beta_0(x_i) + \beta_1(x_i)\,(u_j - u_i) + \beta_2(x_i)\,(v_j - v_i) \\ &+ \frac{1}{2}\beta_3(x_i)(u_j - u_i)^2 + \beta_4(x_i)(u_j - u_i)(v_j - v_i) + \frac{1}{2}\beta_5(x_i)(v_j - v_i)^2.\end{aligned}$$

The local log-likelihood can be written as

$$L(\boldsymbol{\beta}|\lambda, x_i) = \sum_{j=1}^{n} l\left(y_j, \boldsymbol{x}^T\boldsymbol{\beta}\right) w_j, \tag{4.4}$$

where, in the case of locally quadratic fitting, $\boldsymbol{x} = (1, u_j - u_i, v_j - v_i, (u_j - u_i)^2, (v_j - v_i)(u_j - u_i), (v_j - v_i)^2)^T$, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)^T$.

The weights are defined on the bivariate space. The non-negative weight function, $w_j = w_j(x_i)$, depends on the distance $\rho(x_i, x_j)$ between the observations $x_j = (u_j, v_j)$ and the fitting point $x_i = (u_i, v_i)$ and in addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the radius of the neighborhood of x_i.

Maximizing the local log-likelihood (4.4) with respect to $\boldsymbol{\beta}$ gives the vector of estimators $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_5)^T$. Estimator $\psi(x_i)$ is given by $\widehat{\psi}(x_i) = \widehat{\beta}_0$. We proceed by forming the local likelihood as in (4.4) and estimate the coefficients $\boldsymbol{\beta}$ based on data in the neighborhood $x_j = (u_j, v_j)$ of the target point $x_i = (u_i, v_i)$.

Since we want to maximize the log-likelihood, we look for a solution of the set of normal equations to be fulfilled by the maximum likelihood parameter estimates $\boldsymbol{\beta}$. In case of locally quadratic fitting,

$$\frac{\partial L\left(\beta_v|\boldsymbol{y}, w_j, \phi\right)}{\partial \beta_v} = 0 \quad \text{for } v = 0, 1, \ldots, 5.$$

These equations are usually non-linear, so the solution must be obtained through iterative methods. One way to solve those is to use Fisher's scoring method.
The derivatives of the local Poisson log-likelihood function with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned}
&\frac{\partial}{\partial\beta_0}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}; &&\frac{\partial}{\partial\beta_1}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}(u_j - u_i);\\
&\frac{\partial}{\partial\beta_2}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}(v_j - v_i); &&\frac{\partial}{\partial\beta_3}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}(u_j - u_i)^2;\\
&\frac{\partial}{\partial\beta_4}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}(v_j - v_i)(u_j - u_i); &&\frac{\partial}{\partial\beta_5}L = \sum_{j=1}^{n} w_j \frac{d_j - \mu_j}{\mu_j}\frac{\partial\mu_j}{\partial\eta_j}(v_j - v_i)^2.
\end{aligned} \tag{4.5}$$

The Fisher information for $\boldsymbol{\beta}$ is given, in matrix notation, by

$$\mathcal{I}_{vk} = \left\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{X}\right\}_{vk},$$

where $\mathcal{I}$ denotes the Fisher information matrix, $\boldsymbol{X}$ is the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & u_1 - u_i & v_1 - v_i & (u_1 - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2 \\ 1 & u_2 - u_i & v_2 - v_i & (u_2 - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_n - u_i & v_n - v_i & (u_n - u_i)^2 & (u_1 - u_i)(v_1 - v_i) & (v_1 - v_i)^2 \end{bmatrix}$$

and $\boldsymbol{\Omega}$ is the matrix of the working weights just as in (3.11), while $\boldsymbol{W}$ is a diagonal matrix, with entries $\{w_j\}_{j=1}^n$, such that

$$w_j = \begin{cases} W(\rho(x_i, x_j)/h) & \text{if } \rho(x_i, x_j)/h \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

$W(.)$ denotes a non-negative weight function depending on the distance $\rho(x_i, x_j)$. A common choice is the Euclidean distance,

$$\rho(x_i, x_j) = \sqrt{(u_j - u_i)^2 + (v_j - v_i)^2}.$$

In addition, it contains a smoothing parameter $h = (\lambda - 1)/2$ which determines the radius of the neighborhood of x_i. The two components of the Euclidean distance can be scaled in order to apply more smoothing in one direction than the other.

Following the general Fisher scoring procedure, see Section 3.2.4, we obtain the estimates.

When modeling experience data from life-insurance, we wish generally to take into account the exposure in the setting. Specifically, we are looking for a smooth estimate of the observed forces of mortality and from equation (4.3) the linear predictor η_j can be written as

$$\eta_j = \log\left(\mathbb{E}\left[Y|X = x_j\right]\right) = \log(\mu_j) = \log(E_j \varphi_j) = \log(E_j) + \log(\varphi_j)$$

The term E_j called the offset can be easily incorporated.

4.4.3 *p*-splines framework for count data

In this section, we present the essential background material on p-splines methodology for count data. Descriptions of the p-splines method can be found in the seminal paper of Eilers and Marx (1996), as well as in Marx and Eilers (1998), Eilers and Marx (2002), and in Currie and Durbán (2002). Currie *et al.* (2006) present a comprehensive study of the methodology. Applications covering mortality can be found in Currie *et al.* (2004), Richards *et al.* (2006), Kirkby and Currie (2010) and in the Ph.D. thesis of Camarda (2008). Planchet and Winter (2007) use the same framework to discuss an application concerning sick leave retentions.

Again, we suppose that the data can be arranged as a column vector, $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y}) = (y_1, y_2, \ldots, y_n)^T$. Let $\boldsymbol{B}_u = \boldsymbol{B}(\boldsymbol{u})$ and $\boldsymbol{B}_v = \boldsymbol{B}(\boldsymbol{v})$, be regression matrices, of dimensions $n_u \times k_u$ and $n_v \times k_v$, of B-splines based on the duration $\boldsymbol{u}$ and age of occurrence $\boldsymbol{v}$, respectively, with k denoting the number of internal knots.
Specifically, B-splines are bell-shaped curves composed of smoothly joint polynomial pieces. Polynomials of degree 3 are used in the following. The positions on the horizontal axis where the pieces come together are called knots. We use equally spaced knots. The numbers of columns of $\boldsymbol{B}_u$ and $\boldsymbol{B}_v$ are related to the number of knots chosen for the B-splines. Details on B-splines can be found in de Boor (2001).
The regression matrix for our two dimensional model is the Kronecker product

$$\boldsymbol{B} = \boldsymbol{B}_u \otimes \boldsymbol{B}_v$$

The matrix $\boldsymbol{B}$ has an associated vector of regression coefficients $\boldsymbol{a}$ of length $k_u\ k_v$. As in the GLM framework, the linear predictors $\boldsymbol{\eta}$ is linked to the expectation of $\boldsymbol{y}$ by a link function $g(.)$.

$$\boldsymbol{\eta} = g(\mathbb{E}[\boldsymbol{y}]) = \log(\boldsymbol{\mu}) = \boldsymbol{B}\ \boldsymbol{a} = (\boldsymbol{B}_u \otimes \boldsymbol{B}_v)\ \boldsymbol{a}, \tag{4.7}$$

The elements of $\boldsymbol{a}$ can be arranged in a $k_u \times k_v$ matrix $\boldsymbol{A}$, where $\boldsymbol{a} = \mathrm{vec}(\boldsymbol{A})$. The columns and rows of $\boldsymbol{A}$ are then given by $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_u)$ and $\boldsymbol{A}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_v)$. Then instead of computing equation (4.7) as a vector, it can be written as

$$\log(\mathbb{E}[\boldsymbol{y}]) = \log(\boldsymbol{M}) = \boldsymbol{B}_u\ \boldsymbol{A}\ \boldsymbol{B}_v^T. \tag{4.8}$$

From the definition of the Kronecker product, the linear predictor of the columns of $\boldsymbol{Y}$ can be written as linear combinations of k_v smooths in the duration u. The linear predictors corresponding to the jth column of $\boldsymbol{Y}$ can be expressed as

$$\sum_{k=1}^{k_v} b^v_{jk}\ \boldsymbol{B}_u\ \boldsymbol{a}_k,$$

where $\boldsymbol{B}_v = b^v_{ij}$. We apply a roughness penalty to each of the columns of $\boldsymbol{A}$. The penalty is given by

$$\sum_{j=1}^{k_v} \boldsymbol{a}_j^T\ \boldsymbol{D}_u^T\ \boldsymbol{D}_u\ \boldsymbol{a}_j = \boldsymbol{a}^T \left(\boldsymbol{I}_{k_v} \otimes \boldsymbol{D}_u^T\ \boldsymbol{D}_u\right)\ \boldsymbol{a},$$

where $\boldsymbol{D}_u$ is the second order difference matrix acting on the columns of $\boldsymbol{A}$. Similarly by considering the linear predictor corresponding to the ith row of $\boldsymbol{Y}$,

$$\sum_{i=1}^{k_u} \boldsymbol{a}_i^T\ \boldsymbol{D}_v^T\ \boldsymbol{D}_v\ \boldsymbol{a}_i = \boldsymbol{a}^T \left(\boldsymbol{D}_v^T\ \boldsymbol{D}_v \otimes \boldsymbol{I}_{k_u}\right)\ \boldsymbol{a},$$

where $\boldsymbol{D}_v$ is the second order difference matrix acting on the rows of $\boldsymbol{A}$. The penalized log-likelihood to be maximized can be written as

$$\ell^* = \ell(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y}) - \frac{1}{2}\boldsymbol{a}^T \boldsymbol{P}\boldsymbol{a}. \tag{4.9}$$

where $\ell(\boldsymbol{a}; \boldsymbol{B}, \boldsymbol{y})$ is the usual log-likelihood for a GLM and the penalty term $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \lambda_u \left(\boldsymbol{I}_{k_v} \otimes \boldsymbol{D}_u^T \ \boldsymbol{D}_u\right) + \lambda_v \left(\boldsymbol{D}_v^T \ \boldsymbol{D}_v \otimes \boldsymbol{I}_{k_u}\right),$$

where λ_u and λ_v are the smoothing parameters used for the duration and the age of occurrence respectively, $\boldsymbol{I}_{k_u}$ and $\boldsymbol{I}_{k_v}$ being identity matrices of dimension k_u and k_v respectively. More details can be found in Currie *et al.* (2004).
Then maximizing equation (4.9) gives the penalized likelihood equations

$$\boldsymbol{B}^T(\boldsymbol{y} - \boldsymbol{M}) = \boldsymbol{P}\boldsymbol{a},$$

which can be solved by a penalized version of the IRWLS algorithm,

$$\left(\boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{B} + \boldsymbol{P}\right)\boldsymbol{a} = \boldsymbol{B}^T\boldsymbol{\Omega}\boldsymbol{z}, \tag{4.10}$$

where $\boldsymbol{\Omega}$ is the matrix of the working weights similar to (3.11). Again in case of Poisson errors, $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\mu})$. The working dependent variable $\boldsymbol{z}$ is defined by

$$\boldsymbol{z} = \boldsymbol{B}\ \boldsymbol{a} + \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\boldsymbol{\mu}}.$$

Hence, a maximum likelihood estimate of $\boldsymbol{a}$ is found by a penalized version of IRWLS algorithm:

> Repeat $\boldsymbol{a}^* := \boldsymbol{B}\big(\boldsymbol{B}^T\boldsymbol{\Omega}\ \boldsymbol{B} + \boldsymbol{P}\big)^{-1}\boldsymbol{B}^T\boldsymbol{\Omega}\ \boldsymbol{z}$;
>
> using $\boldsymbol{a}^*$, update the working weights $\boldsymbol{\Omega}$, as well as the working dependent variables $\boldsymbol{z}$ until convergence.

Again when modeling mortality data, we may take into account the exposure in the setting. The linear predictor $\boldsymbol{\eta}$ can be written as

$$\boldsymbol{\eta} = g(\mathbb{E}[\boldsymbol{y}]) = \log(\boldsymbol{\mu}) = \log(\boldsymbol{e}) + \log(\boldsymbol{\varphi}) = \log(\boldsymbol{e}) + \boldsymbol{B}\ \boldsymbol{a} = \log(\boldsymbol{e}) + (\boldsymbol{B}_u \otimes \boldsymbol{B}_v)\ \boldsymbol{a},$$

where $\boldsymbol{e}$ denotes the vector of exposure. Similarly to Section 4.4.2, the offset can be easily incorporated in the regression system (4.10).

The penalized IRWLS would be efficient only in moderate-sized problems. For our application, the parameter vector $\boldsymbol{a}$ has length 2520 and this required the usage of 2520 × 2520 matrices. The size is moderate, but for larger dimensional matrices the penalized IRWLS algorithm can run into storage and computational difficulties. Currie *et al.* (2006) and Eilers *et al.* (2006) proposed an algorithm that takes advantage of the special structure of both the data as a rectangular array and the model matrix as a tensor product. The idea of this algorithm can be seen in the computation of the mean $\boldsymbol{\mu} = \text{vec}(\boldsymbol{M})$ in two dimensions, as in (4.8). This avoids having to construct a large Kronecker product basis, saving time and space. For the presentation of this algorithm we refer to the mentioned articles Currie *et al.* (2006) and Eilers *et al.* (2006).

The smoothing parameters for p-splines method are chosen according the Bayesian information criterion (BIC) which penalizes heavily the model complexity particularly when n is large,

$$BIC = \sum_{i=1}^{n} D(y_i, (\theta(\widehat{\mu}_i))) + \log(n)\ v.$$

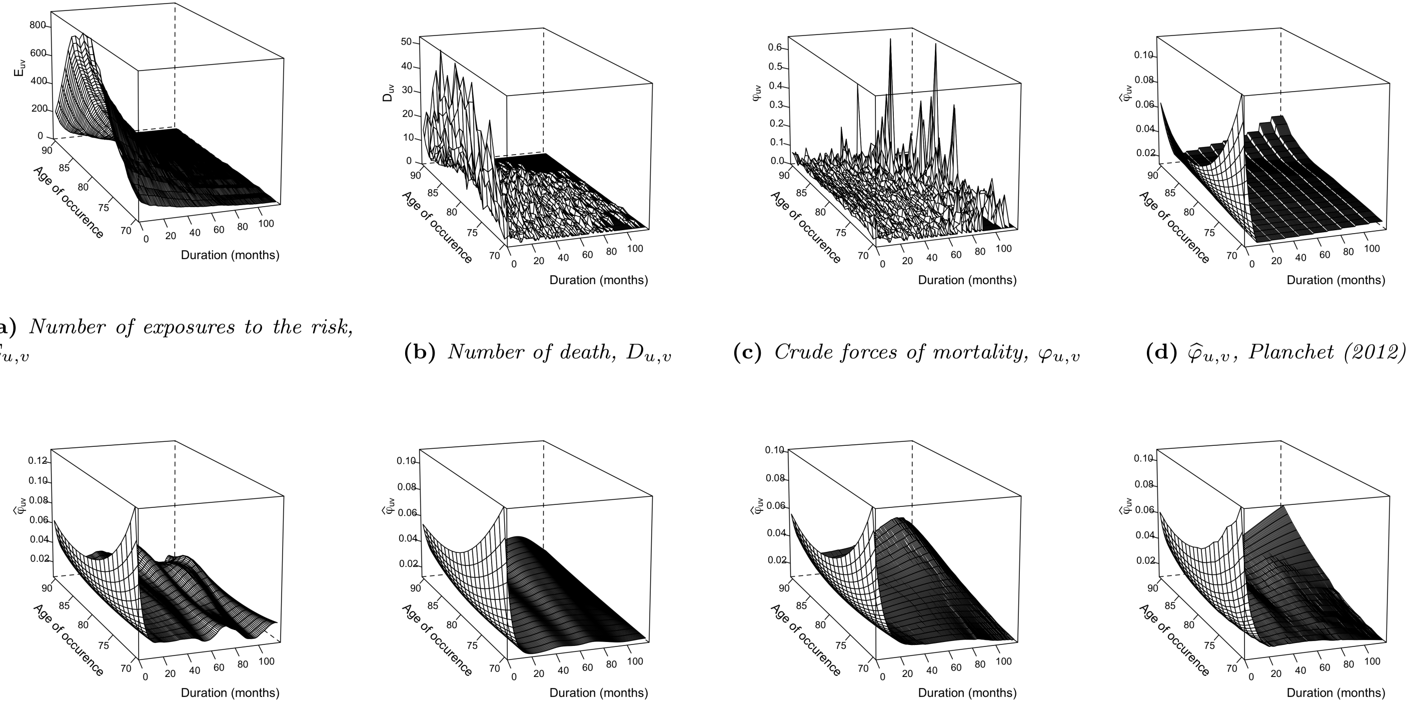
4.4.4 The data

The data come from observations of individuals subscribing to LTC insurance policies originating from a portfolio of a French insurance company. For these applications, we focus on measuring the forces of mortality as a function of the age v of occurrence of the pathologies and the duration u of the care.
The range of ages of occurrence is $70 - 90$ and the maximum duration of the pathologies is 119 months. The period of observation stretches from 01/01/1998 to 31/12/2010. The data have been aggregated according to the age of occurrence and the duration. The pathologies are composed, among others, by dementia, neurological illness and terminal cancer. The data consist for 2/3 of women and 1/3 of men. Figures 4.11a, 4.11b, and 4.11c display the observed statistics of the dataset.

Moreover, we have at our disposal the adjusted surface obtained from the technical report Planchet (2012), Figure 4.11d. It gives an idea about the desirable shape that we aim to retain, and the adjusted forces of mortality will be useful when assessing the comparisons of the models. This surface has been obtained by treating separately the first month of duration from the others and applying a Whittaker-Henderson model to adjust the crude surface.

4.4.5 Smoothed surfaces and fits

Figure 4.11e presents the smoothed surface obtained with the local likelihood model with an Epanechnikov weight function, a polynomial of degree 2 and a bandwidth (radius) of 13 observations. The corresponding degrees of freedom v are 29.25. The order of polynomial and the bandwidth have been chosen by minimizing the AIC criterion. The surface is relatively wiggly showing an inappropriate variance.



(a) *Number of exposures to the risk, $E_{u,v}$*

(b) *Number of death, $D_{u,v}$*

(c) *Crude forces of mortality, $\varphi_{u,v}$*

(d) $\widehat{\varphi}_{u,v}$, *Planchet (2012)*

(e) $\widehat{\varphi}_{u,v}$, *local likelihood*

(f) $\widehat{\varphi}_{u,v}$, *p-splines*

(g) $\widehat{\varphi}_{u,v}$, *ICI*

(h) $\widehat{\varphi}_{u,v}$, *local bandwidth factors*

Figure 4.11: *Observed statistics: $E_{u,v}$, $D_{u,v}$, $\varphi_{u,v}$ and smoothed forces of mortality $\widehat{\varphi}_{u,v}$ according to Planchet (2012), local likelihood, p-splines, ICI rule and local bandwidth factors methods*

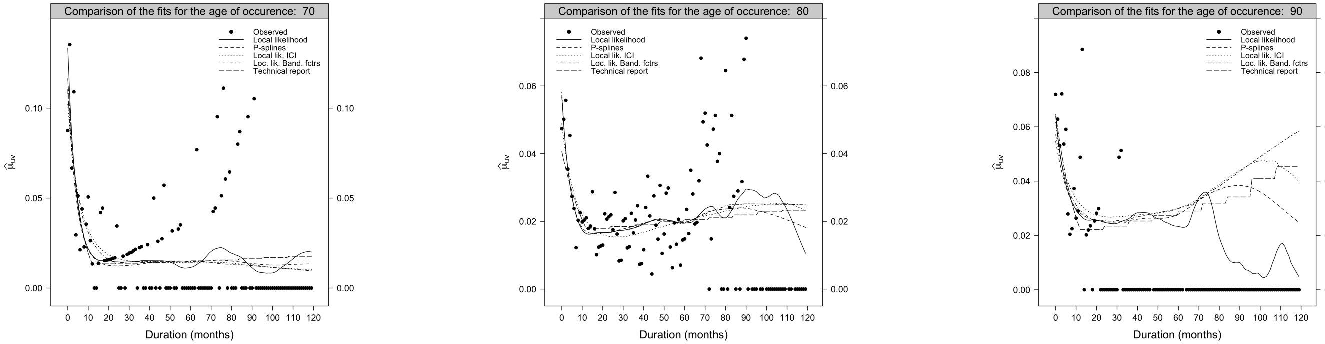
Figure 4.11f displays the smoothed surface obtained when fitting p-splines. The smoothing parameters $\lambda_u = 31.6$, $\lambda_v = 31.6$, have been chosen by minimizing the BIC criterion. It leads to $k_u = 24$, $k_v = 4$ for $\upsilon = 18.11$. The surface seems satisfactory, though the increase in the upper right corner (highest age of occurrence and longest duration) is not present as in the surface adjusted from Planchet (2012).

Figures 4.11g and 4.11h present the smoothed surface obtained with the adaptive local likelihood methods. For these applications, only the bandwidth is varying. The order of polynomial is still fixed at 2 and we use an Epanechnikov weight function. The fitted degrees of freedom υ are 10.05, 10.76 and 16.16 respectively.

In general, only for the first months of the duration, the graduations are similar. After that, we obtain very different shapes according to the models. The ICI rule and the local bandwidth factors seem the most satisfying methods in modeling the monotone phenomenon at the extreme ages, Figures 4.11g and 4.11h. The fitted degrees of freedom for the local bandwidth factors are larger than the ones obtained by the ICI rule indicating that the model is slightly more flexible and shows more features. The bandwidth values depend on the amount of exposure to represent effectively the remaining life expectancy in the regions where the amount of exposure is high. The corresponding bandwidths, in the left region, are relatively low, and they increase as the amount of exposure decreases. For regions in which the amount of exposure is low, a large value for the bandwidth results in an estimate that progress more smoothly. As we already obtained the shape and the magnification of the local bandwidth factors, we used the AIC criterion to decide the global value at which the bandwidth curve is located. The sensitivity parameter s for the local bandwidth factors as well as the value c for the ICI rule have been chosen arbitrarily to be $0,15$ and 0.1 respectively. For higher value of s spurious features started to appear showing unacceptable variance, while for higher c, bias tends to show up.

Figures 4.12 and 4.13 present the smooth fits obtained from the different models for various ages of occurrence and durations.

The approaches produce relatively similar graduations for regions having a low amount of noise, Figure 4.12b. However, when the data are more volatile the benefits of the adaptive approaches become apparent. The fit obtained from global local likelihood, and not as strongly the p-splines, present an unacceptably high variance. It shows the inapplicability to model such datasets with global methods or to select the smoothing parameters by relying explicitly on a criterion. The local bandwidth factors method has the capability to model the forces of mortality in the first months of duration relatively well, Figure 4.12d, and the sharp increase at the highest extremes of the age of occurrence and duration, Figures 4.12c and 4.12f. The ICI rule

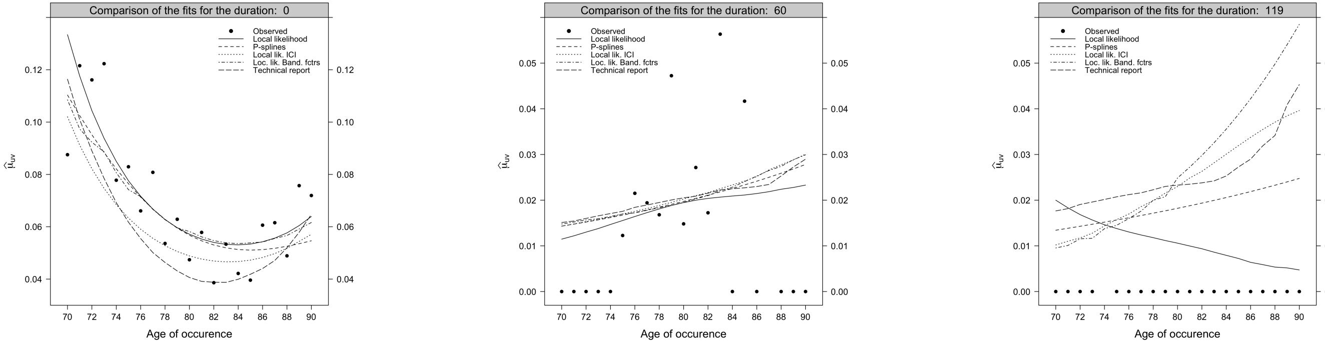


(a) *Age of occurrence* $v = 70$

(b) *Age of occurrence* $v = 80$

(c) *Age of occurrence* $v = 90$

Figure 4.12: *Observed forces of mortality and smooth fits for various ages of occurrence.*



(d) *Duration* $u = 0$

(e) *Duration* $u = 60$

(f) *Duration* $u = 119$

Figure 4.13: *Observed forces of mortality and smooth fits for various durations.*

and p-splines fail to model these features. However, all the models miss the slow increase at the age of occurrence 70 present in the fit obtained from Planchet (2012), Figure 4.12a.

4.4.6 Analysis of the residuals

Figure 4.14 presents the residuals of the 5 models for the age of occurrence 70 as well as the ones obtained from the adjusted surface from Planchet (2012).

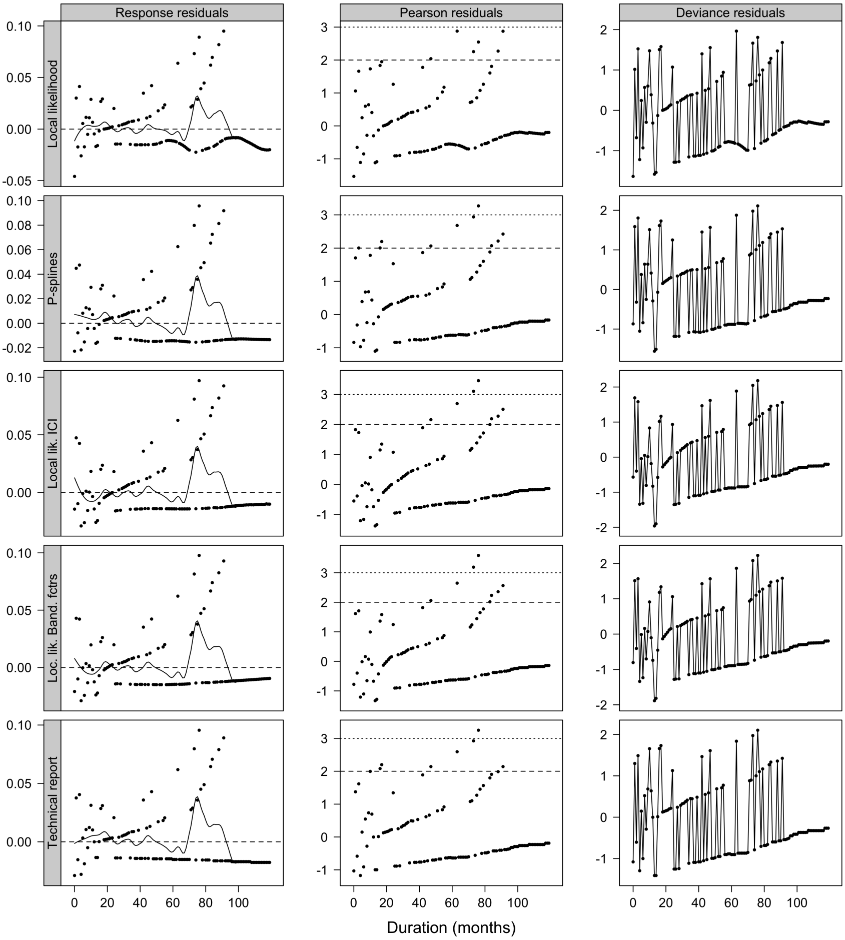


Figure 4.14: *Response, Pearson and deviance residuals for the age of occurrence* 70

The pattern of the residuals displayed for each model is roughly similar. We superimposed a *loess* smooth curve on the response and Pearson residuals. These smooths help search for clusters of residuals that may indicate lack of fit. By reducing the noise, our attention may be more readily drawn to features that have been missed or not properly modeled by the smooth. Here the process is not to judge a fit adequate if a smooth curve on its residuals plot is flat. A flat curve means simply that no systematic, reproducible lack of fit has been detected. The fit may well be too noisy, and stays to close to an interpolation since trends in small parts of the data are interpreted as more widespread trends. Then for small datasets, the fit is very nearly interpolating the data which results in unacceptably high variance. Strong patterns appear in the response residuals in Figure 4.14. It indicates a lack of fit in this region. However, this is not surprising as most of the deaths at the longest durations are zero for the age of occurrence 70.

The Pearson residuals are mainly in the interval $[-2, 2]$, which indicates that the models adequately capture the variability of the dataset.

The deviance residuals present, for the longest durations, several successive residuals having the same sign. It illustrates that the forces of mortality are over-smoothed locally. As the sign is negative, from 80 to 119 months, we strongly overestimate the forces of mortality. However, we would have excepted such a pattern as we observe zero deaths at the highest extreme of the duration of the care.

4.5 Comparisons

4.5.1 Tests to compare graduations

We continue the comparisons by applying the tests proposed by Forfar *et al.* (1988, p.56-58) and Debón *et al.* (2006, p.231). We have also obtained the values of the mean absolute percentage error $MAPE$ and R^2 used in Felipe *et al.* (2002). We compare the crude mortality rates to the graduated series to see whether the approaches lead to similar graduation. Table 4.3 presents the results.

The approaches display different results. The global local likelihood approach, having the highest degrees of freedom, has the capacity to show many features in the data. Therefore, the values of the various tests are the *best*. It has the lowest deviance, lowest number of standardized residuals exceeding the thresholds 2 and 3, highest number of runs, best mix of the residuals between positive and negative signs, highest value for the run test. In addition, the approach results in the minimum χ^2 and $MAPE$. Conversely, the adaptive local likelihood using the ICI rule yields the smallest degrees of freedom. As a consequence, the results of the various tests and values of the χ^2 and $MAPE$ are the *worst*.

The results for the p-splines and the adaptive approaches using the local bandwidth factors are similar, even though we have seen that the adaptive method has a better ability to model the mortality patterns (high mortality for the first month of duration of the care and increase at the extreme highest

		Local lik.	p-splines	Adapt. lik. ICI	Adapt. band. factors	Planchet (2012)
Fitted DF υ		29.25	18.11	10.76	16.16	NA
Deviance		2259.91	2291.89	2440.81	2311.04	2409.99
Standardised	> 2	108	117	127	115	129
sesiduals	> 3	25	31	38	32	42
Signs	$+(-)$	910(1610)	907(1613)	891(1629)	900(1620)	908(1612)
test	p-value	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
Runs	Nb of runs	1028	1016	971	1019	1013
test	Value	-6.25	-6.71	-8.19	-6.47	-6.64
	p-value	$4.05e-10$	$1.98e-11$	$2.57e-16$	$9.69e-11$	$3.10e-11$
Kolmogorov	Value	0.4401	0.5008	0.4829	0.4786	0.5167
Smirnov test	p-value	0.00	0.00	0.00	0.00	0.00
χ^2		2433.56	2487.09	2616.08	2458.71	2644.09
R^2		0.2406	0.2340	0.2433	0.2476	0.2234
$MAPE$ (%)		46.11	46.62	48.18	46.99	47.38

Table 4.3: *Comparisons between the smoothing approaches.*

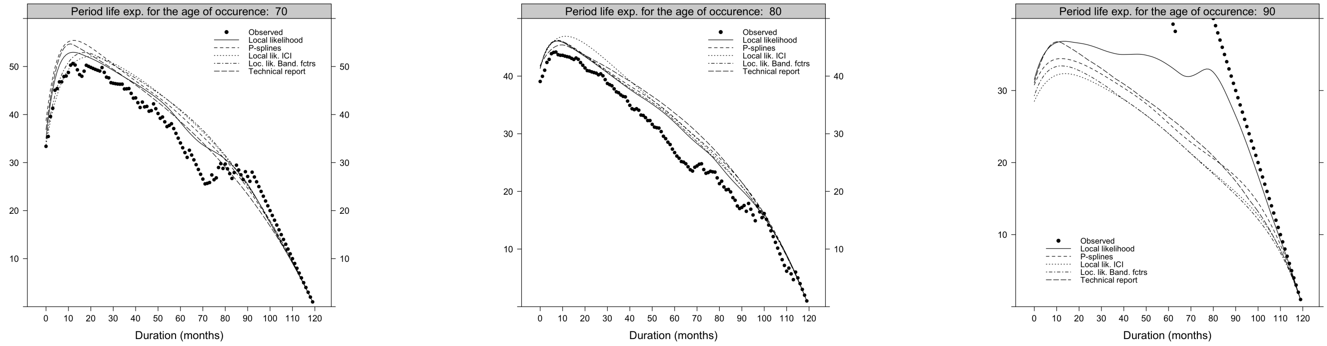
of the duration and age of occurrence), even though the p-splines model has higher degrees of freedom.

4.5.2 Comparing figures summarizing the lifetime probability distribution

We end these comparisons by presenting some figures summarizing the lifetime probability distribution. Figures 4.15 and 4.16 display the life expectancy obtained from the different models for various ages of occurrence and durations.

At age of occurrence 70, with the exception of the adaptive local likelihood using the ICI rule and local bandwidth factors, the models are over-estimating the period life expectancy for the first months of duration (until 10 months), Figure 4.15a. This is particularly visible for the p-splines and the adjusted surface obtained in Planchet (2012). The over-estimation is general at age of occurrence 80, Figure 4.15b. The shapes of the life expectancy differ much at age of occurrence 90, Figure 4.15c, where the global local likelihood tends to estimate a more rectangular shape.

The shape and trend of the life expectancies are similar when we observe a large amount of exposure (first months of duration of the care), Figure 4.15d. The high correlation of the pathologies with the age of occurrence can explained the concave shape observed for the life expectancies during

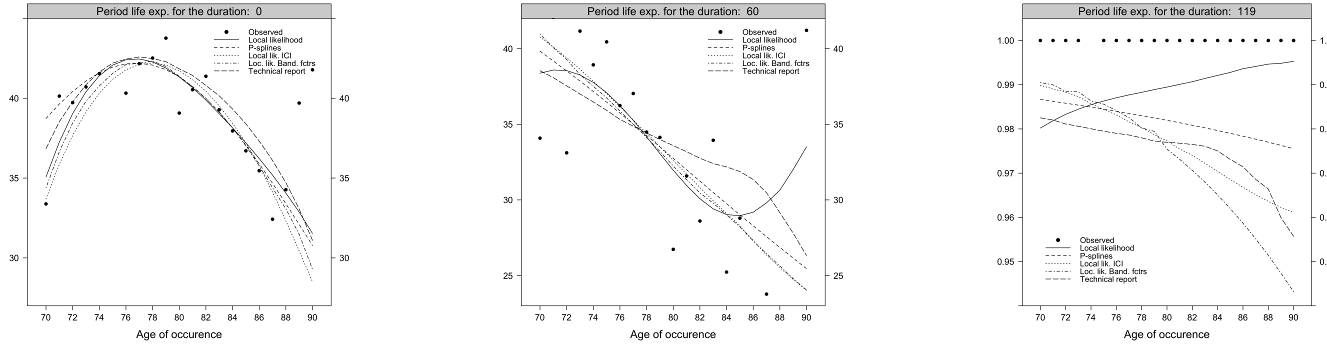


(a) *Age of occurrence $v = 70$*

(b) *Age of occurrence $v = 80$*

(c) *Age of occurrence $v = 90$*

Figure 4.15: *Observed and predicted life expectancy for various ages of occurrence.*



(d) *Duration $u = 0$*

(e) *Duration $u = 60$*

(f) *Duration $u = 119$*

Figure 4.16: *Observed and predicted life expectancy for various duration.*

the first months of the care. The lowest ages of occurrence are marked by a relatively high mortality mainly due to the death of the individuals suffering from terminal cancer, while the highest ages concern principally the dementia. At the 60th month of the care, the life expectancy is decreasing rapidly, Figure 4.15e. However, while the ICI rule and local bandwidth factors produce similar patterns, the shapes and trends given by the other models diverge markedly, the local likelihood predicting a rise of the life expectancy for the highest age of occurrence. This pattern is also present, although less markedly, in Figure 4.15f.

Figure 4.17 shows the median month at death, Figure 4.17a, standard deviation of the random life time, Figure 4.17b and entropy, Figure 4.17c, as a function of the age of occurrence of the pathology.

In Figure, 4.17a displaying the median month at death as a function of the age of occurrence of the pathology, we observe a concave shape similar to Figure 4.15d. This phenomenon shows, once more, the correlation between the age of occurrence and the pathologies. The adaptive local likelihood using the ICI rule, having the lowest degrees of freedom, mostly underestimates the median month at death compared to the others models.

After a steady increase, the standard deviation of the random lifetime is slowing down at age of occurrence 82, and decreases until 90 years old, Figure 4.17b. It is explained by the fact that we observe most of the deaths at the lowest age of occurrence and duration, while the number of deaths is zero, and thus stable, at the highest age and duration.

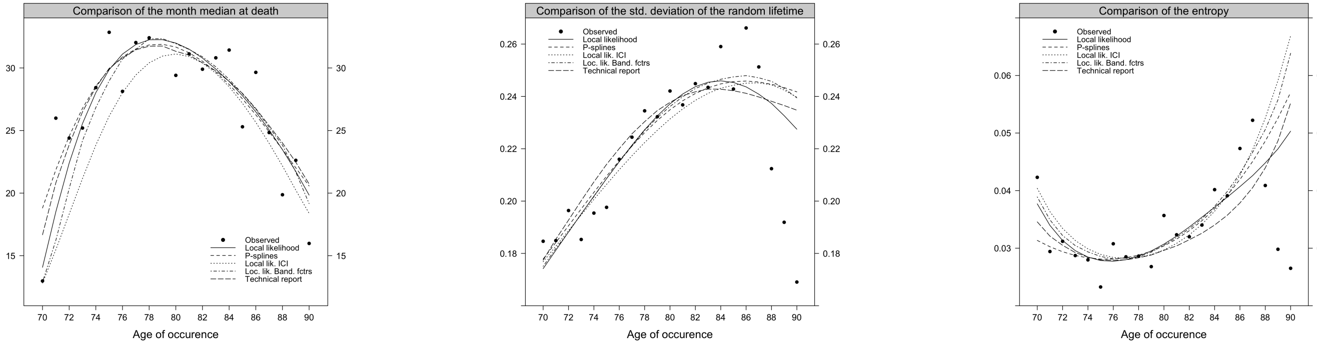
Figure 4.17c shows the entropy. The values decline as the deaths become more concentrated. We observe that the deaths predicted by the adaptive local likelihood models (ICI rule and bandwidth factors) are the most stretched. Conversely, the adjusted number of deaths obtained by Planchet (2012) are more concentrated.

Table 4.4 summarizes the indices. For the life expectancy, ${}_{0|120}e_{70}$, ${}_{0|120}e_{80}$, and ${}_{0|120}e_{90}$, the observations suggest an increase with the age, which, based on our knowledge, is unrealistic. We are more likely to look for a concave shape, predicted by the models as displayed in Figure 4.15d. On average, the models agree on the same life expectancy, around 38 months and underestimate slightly the observed ${}_{0|120}e$.

The median month at death, $\mathrm{Med}(T_0)$, estimated by the models varies in average slightly from 25 to 27 months. However, for a particular age of occurrence, such as $\mathrm{Med}(T_0(70))$, the difference between the models (p-splines and ICI rule) can grow until 6 months.

All the models sensibly estimated the same average standard deviation of the random life time, σ_0, which corresponds to the observed standard deviation, around 0.22.

Finally, all the models agree on the estimated average entropy $H(T_0)$,



(a) *Median month at death*

(b) *Std. dev. of life time*

(c) *Entropy*

Figure 4.17: *Median month at death, standard deviation of the random life time and entropy with the age of occurrence of the pathology.*

	Observed	Local lik.	p-splines	Adapt. lik. ICI	Adapt. band. factors	Planchet (2012)
${}_{0\|120}e_{70}$	33.38	35.07	38.73	33.71	34.38	36.84
${}_{0\|120}e_{80}$	39.06	41.33	41.34	41.72	41.37	41.84
${}_{0\|120}e_{90}$	41.78	31.53	30.78	28.50	29.29	31.09
${}_{0\|120}e$	39.44	38.59	38.80	37.88	38.12	39.23
$\text{Med}(T_0(70))$	12.98	14.07	18.81	12.81	12.89	16.67
$\text{Med}(T_0(80))$	29.41	31.99	31.67	31.12	31.94	31.33
$\text{Med}(T_0(90))$	15.98	19.84	20.57	18.38	19.16	20.73
$\text{Med}(T_0)$	27.07	26.83	27.29	24.99	26.40	27.18
$\sigma_0(70)$	0.1846	0.1742	0.1778	0.1768	0.1751	0.1774
$\sigma_0(80)$	0.2421	0.2372	0.2349	0.2314	0.2361	0.2377
$\sigma_0(90)$	0.1691	0.2274	0.2417	0.2397	0.2394	0.2347
σ_0	0.2196	0.2231	0.2250	0.2228	0.2251	0.2250
$H(T_0(70))$	0.0423	0.0377	0.0313	0.0404	0.0389	0.0346
$H(T_0(80))$	0.0357	0.0307	0.0304	0.0297	0.0305	0.0296
$H(T_0(90))$	0.0265	0.0503	0.0570	0.0669	0.0638	0.0552
$H(T_0)$	0.0337	0.0350	0.0353	0.0374	0.0370	0.0341

Table 4.4: *Single figure indices to summarize the lifetime probability distributions*

between 0.035 to 0.037. The entropy estimated from the adjusted surface obtained in Planchet (2012) suggests that the estimated deaths are less stretched than the models predictions.

4.6 Summary and Outlook

In this chapter, we illustrate how adaptive local likelihood methods can be used to graduate mortality tables in two dimensions. Tests and single indices summarizing the lifetime distributions are used to compare the graduated forces of mortality obtained from adaptive local likelihood to global non-parametric methods such as local likelihood and p-splines models.

Using locally adaptive parameters instead of a global smoothing one may be advantageous for several reasons. The estimator can adapt to the structure of the regression function and to the reliability of the data, smoothing more when the volume of observations is low and less when it is high.

The intersection of confidence intervals (ICI) rule has been introduced as a locally adaptive pointwise method. The critical value controls the bias-variance tradeoff. Because a larger class of estimators is available, it may in turn affect the variability. Hence, the set of window sizes contains relatively large bandwidths. The choice of the set of window sizes is done subjectively,

based on the the mechanism generating the data and on the performance of the smoother used in the fitting. Another drawback in applying such methods is that they require more computer time than a global procedure. Specifically, the computational effort is multiplied by the number of observations.
A technique closely related to the ICI rule is the Lepski method. This procedure uses the standard deviation of the difference $\widehat{\psi}_{\lambda_1}(x_i) - \widehat{\psi}_{\lambda}(x_i)$ for some $\lambda \leq \lambda_1$ until a significance difference is found. Chichignoud (2010, Section 1.5) provides an extensive discussion of the technique in his recent Ph.D Thesis. The discussion and the implementation of the Lepski method for graduating experience data originating from life insurance is a topic of ongoing research.

The bandwidth correction factors method allows the estimated forces of mortality to include explicitly the extra information provided by the changing amounts of exposure. The observed exposure decided the shape of the smoothness parameter. The magnification of that shape has been determined by a sensitivity parameter which we chose subjectively for practical reasons. The global bandwidth parameter is used to control the absolute level of the bandwidth curve. We used a global criterion instead of pointwise methods. It appears that the procedure has the ability to model relatively well the mortality pattern where the other models fail to model these features.

In global procedures as well as for locally adaptive procedures, there is no deterministic method to obtain the constellation of smoothing parameters with the classical selectors. Residual analysis and goodness of fit diagnostics are just as important for locally adaptive procedures as they are for global procedures. It is important to use appropriate residuals diagnostics to look for lack of fit. The purpose for which the mortality table is required must be kept clearly in mind, and the final choice of graduation is always a matter of judgment.

The methodologies proposed adapt neatly to the complexity of mortality surface, clearly because of the appropriate data-driven choice of the adaptive smoothing parameters. The adaptive local likelihood method using the bandwidth factors models well the high mortality during the first months of duration and the increase at the extreme high duration and age of occurrence compared to the other methods. Having 13 degrees of freedom less than the local likelihood model, the adaptive bandwidth factors model is less flexible although the tests presented in Table 4.3 show relatively good results. Rather than treating the first months separately, having an adaptive model can be a benefit. However, the relative merit of the procedures would depend on the purpose for which the mortality table has been computed. If we are essentially exploring the data, then additional information derived might not justify the effort. However, the potential uses of adaptive approaches suggest that they have much to offer as part of the actuarial toolkit.

Chapter 5

Entity specific prospective mortality tables

5.1 Introduction

It is now well documented that the human mortality has globally declined over the 20th century. Life expectancy is greater than ever before and increasing rapidly, see Pitacco *et al.* (2009, Ch.3). These mortality improvements pose a challenge for the pricing and reserving in life insurance and for the management of public pension regimes. In a pension plan, the longevity risk is transferred from the policyholder to the insurer. The latter has to evaluate his liability with appropriate mortality tables. It is in this context that since 1993 the French regulatory tables for annuities have been prospective mortality tables taking in account the increase of the life expectancy.

Prospective mortality tables allow to determine the remaining lifetime for a group, not according to the conditions of the moment, but given the future developments of living conditions. However, applying exogenous tables to the group considered may result in under-provisioning the annuities, when the mortality of the group is lower than of the reference population. With the international regulations *Solvency II* and *IFRS* insurers are required to evaluate their liabilities from realistic assumptions leading to an evaluation of the *best estimate*. In consequence, for pensions regimes and more generally due to the longevity risk, insurers have to build specific mortality tables, taking into account the expected evolution of the mortality of their insured population, see Planchet and Kamega (2011).

Probably because it was not understood initially in which respects demographic sciences differed from natural sciences, it was believed that mortality laws similar to those discovered in astronomy and physics could be found. However, none were found and it is far from certain that there are any. As a consequence it is impossible to predict the evolution of mortality as the expected movements of the stars, as notes Henry (1987).

Not being able to predict from laws, but being forced to forecast, Henry (1987) suggests that it is in the experience that we should seek the best means to do it. His view is not radically different from the one expressed by de Laplace (1829) where ignorance is temporary and research will increase our understanding and help formulating accurate forecasts. *« Imagine [...] an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings which compose it [...], to it, nothing would be uncertain and the future, as the past, would be present in its eyes »*.

Laws can be replaced by assumptions about the future characteristics of a population to deduce future perspectives, in numbers and structure, of this population. In the absence of laws, we observe some regularities and patterns in mortality, either permanent or specific to a certain period, from which we can produce forecasts that most of the time are sufficient for our needs. It is generally accepted that the demographic phenomenon of inertia is sufficient (apart from periods of war) for extrapolation of past trends, see Booth (2006).

In this chapter, we present a two steps approach to build entity specific mortality tables. From portfolios of several insurance companies, the first step consists in constructing global prospective mortality tables by gender. For clarity, only results about the male population are presented. By reasoning globally, this table summarizes the male mortality experience of these portfolios. The heterogeneity present between the portfolios is taken into account in a second step. The male prospective table is then used to adjust the mortality specifically to each male insured portfolio. The computations are carried out with the help of R, R Development Core Team (2012).

When the size of the group is sufficiently large, we can construct a prospective mortality table with the intention of identifying the behavior of the insured population that would differ from the regulatory tables or more generally from the national standard. However, in practice the size of the group may be limited and the past experience is observed over a short period. As mentioned in Planchet and Lelieur (2007), two approaches can be proposed to smooth the crude data and project the future mortality using past observations. We distinguish

i. Endogenous approaches, which consist of exploiting the information contained in the crude forces of mortality to obtain a smooth surface representing the data correctly, and yield a *realistic* projection. In case of a small volume of data, these techniques could lead to biased estimations of the mortality trend.

ii. Models using an external reference mortality table (exogenous approaches) that present a solution to overcome the difficulties associated with having a small volume of data. The idea is to adjust a reference table to the experience of a given set of data.

Considering the limited volume of data available, our attention, in the first step of our methodology, is focused on the second class of models even though comparisons with the first approach are presented.

As a part of the construction of such tables, it is necessary to describe the risks we are facing according to their nature: poolable (hazard on different strata of the population) or systematic (the financial impact can be potentially more important for the insurer or the pension regime). More precisely Planchet and Kamega (2011), similar to Alho (1990), classified the risks into four different but related sources:

i. The risk that can be pooled, originating from random variations of the empirical expectancy around the mathematical expectation due to the sampling variations.

ii. The systematic risk of parameters estimation, originating from a wrong estimation of the model parameters given the sampling variations.

iii. The systematic risk of errors in expert judgment when taking into account external information.

iv. The systematic risk of model due to model misspecification or a change in the trend over time.

The poolable risk, referring to the random character of individual deaths, is not treated here. Extensive studies have discussed the issue of systematic risk of parameter estimation due to the sampling fluctuations. The variance and covariance of parameter estimates are derived either by standard estimation or by bootstrapping, resampling from the original data to create replicate datasets from which the variance and covariance can be estimated. See Planchet and Kamega (2011) for an application of parametric bootstrap.

In this chapter, we focus on the model risk and, to a lesser extent, on the risk of expert judgment related to the choice of the external references used. There is a need for awareness of model risk when assessing longevity-related liabilities, especially for annuities and pensions regimes. The problem is that one can quantify uncertainty within a given model, but one cannot quantify the uncertainty about the model itself. If recent studies, Sibberstein *et al.* (2008) or Richards and Currie (2009), suggest a rather general analytical framework for the pricing of financial derivatives, the establishment of a standard framework for mortality and longevity models remains to be done. The model risk is particularly difficult to measure, because we cannot measure the uncertainty on a number, as with a price. We have to measure the uncertainty on a much more complex object, which is the survival distribution. A pragmatic approach is to define a set of possible models and measure the differences on variables of interest.

In our first step, we do not take into account the heterogeneity between the different portfolios. The mortality of the entire male population is not specific to any male subpopulation. The second step of our approach is then to build entity specific male prospective mortality tables by adjusting the reference table validated in the first step to the mortality of each male portfolios. For this purpose, we use a Poisson generalized linear model including interactions with age and calendar year.

The chapter is organized as follows. Section 5.2 specifies the notation and assumptions used in this chapter. It also briefly describes the data and presents our approach to construct specific prospective mortality tables. Section 5.3 covers the extrapolation method for the surfaces obtained by local likelihood smoothing. The extrapolation is performed by identifying the mortality components and their importance over time using functional data analysis. Time series methods are used to extrapolate the time-varying coefficients. The construction of a global prospective reference table for the male population is illustrated in Section 5.4 with the assessment of model risk. Section 5.5 illustrates the construction of entity specific prospective tables. A Poisson GLM including interactions with age and calendar year gives a solution to this problem. Finally, Section 5.6 summarizes the conclusions drawn in the chapter.

5.2 Notation, assumption, data and approach

5.2.1 Notation

We analyze the changes in mortality as a function of both the attained age x and the calendar year t. The force of mortality at attained age x for the calendar year t, is denoted by $\varphi_x(t)$. We denote $D_{x,t}$ the number of deaths recorded at attained age x during calendar year t from an exposure-to-risk $E_{x,t}$ that measures the time during individuals are exposed to the risk of dying. It is the total time lived by these individuals during the period of observation.

5.2.2 Piecewise constant forces of mortality

We assume that the age-specific forces of mortality are constant within bands of time, but allowed to vary from one band to the next, $\varphi_{x+\tau}(t+\xi) = \varphi_x(t)$ for $0 \leq \tau < 1$ and $0 \leq \xi < 1$.

We denote by $p_x(t)$ the probability that an individual aged x in calendar year t reaches age $x+1$, and by $q_x(t) = 1-p_x(t)$ the corresponding probability of death. The expected remaining lifetime of an individual reaching age x during calendar year t is denoted by $e_x(t)$. Under the assumption of piecewise constant forces of mortality, we have for integer age x and calendar year t,

$$\begin{aligned}
p_x(t) &= \exp(-\varphi_x(t)) \Leftrightarrow \varphi_x(t) = -\log(p_x(t)),\\
e_x(t) &= \frac{1-\exp(-\varphi_x(t))}{\varphi_x(t)}\\
&\quad + \sum_{k\geq 1}\left\{\prod_{j=0}^{k-1}\exp(-\varphi_{x+j}(t+j))\right\}\frac{1-\exp(-\varphi_{x+k}(t+k))}{\varphi_{x+k}(t+k)}.
\end{aligned}$$

5.2.3 The data

Data are originating from 8 portfolios of various French insurance companies, denoted P1, P2 ..., P8. Table 5.1 displays the observed statistics of the male data.

					Period of observation	
Portfolios	Mean Age In	Mean Age Out	Mean Expo	Mean Age at death	Beginning	End
P1	38.36	43.42	5.06	53.47	01/01/1996	31/12/2007
P2	44.28	45.76	1.48	51.68	01/01/2005	31/12/2007
P3	43.18	45.44	2.26	76.98	01/07/2004	30/06/2007
P4	51.43	61.74	10.31	77.92	01/01/1996	31/12/2007
P5	42.48	44.60	2.12	54.42	01/01/2003	31/12/2007
P6	47.42	51.15	3.73	71.84	01/01/1996	31/12/2007
P7	55.77	56.78	1.01	72.44	01/01/2006	31/12/2007
P8	53.65	55.06	1.41	62.28	01/01/2005	31/12/2007

Table 5.1: *Observed statistics by portfolios.*

The second column of Table 5.1 displays the mean age at entrance of the period of observation for the male population while the third column is the mean age at exit. The fourth column is the average exposure to the risk, and the fifth presents the average age at death. The period of observation of each portfolio is displayed in the sixth and seventh columns. It illustrates that we are facing two difficulties. On one hand the period of observation is small, spreading over 12 years. On the other hand, the structure of the heterogeneity is changing over time as the portfolios are not observed during the same period. Figure 5.1 displays the difference between the portfolios.

From Figure 5.1, the differences in structure by age between the portfolios are apparent. Portfolios P3, P6 and P7, top left corner, are marked by a

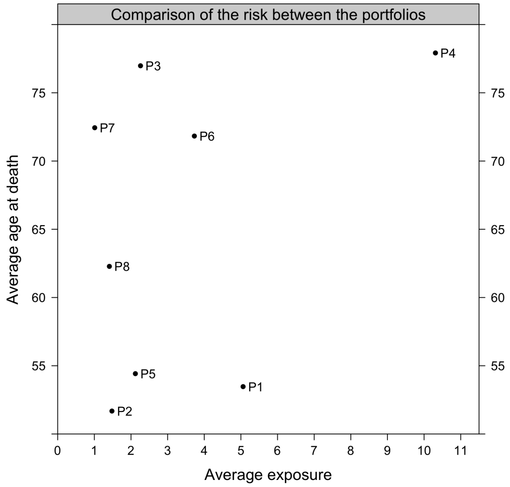


Figure 5.1: *Comparisons of the level of the risk between the portfolios.*

relative low average exposure and high average age at death, while portfolios P1, P2 and P5 have a lower average age at death. But a low average age at death does not necessarily mean a higher mortality because observations are censored and truncated.

5.2.4 The approach

With the notation of Section 5.2.2 and under the assumption of a piecewise constant force of mortality, the likelihood becomes

$$\mathcal{L}(\varphi_x(t)) = \exp(-E_{x,t}\ \varphi_x(t))(\varphi_x(t))^{D_{x,t}}.$$

The associated log-likelihood is

$$\ell(\varphi_x(t)) = \log \mathcal{L}(\varphi_x(t)) = -E_{x,t}\ \varphi_x(t) + D_{x,t} \log \varphi_x(t).$$

Similarly to Section 4.4.1, maximizing the log-likelihood $\ell(\varphi_x(t))$ gives $\widehat{\varphi}_x(t) = D_{x,t}/E_{x,t}$ which coincides with the central death rates $\widehat{m}_x(t)$. Then it is apparent that the likelihood $\ell(\varphi_x(t))$ is proportional to the Poisson likelihood based on

$$D_{x,t} \sim \text{Poisson}(E_{x,t}\varphi_x(t)), \tag{5.1}$$

and it is equivalent to work on the basis of the *true* likelihood or on the basis of the Poisson likelihood, as recalled in Gschlössl *et al.* (2011). Thus, under the assumption of constant forces of mortality between non-integer values of x and t, we consider (5.1) to take advantage of the Generalized Linear Models (GLMs) framework.

Our approach to construct entity specific mortality tables is in two steps. From our collection of portfolios originating from several insurance companies, the first step consists in constructing global prospective mortality tables by gender. For clarity, only results about the male population are presented. By reasoning globally, the male mortality table summarizes the mortality experience of the male portfolios. For this purpose, following Hyndman and Ullah (2007) and Hyndman and Booth (2008), we use principal component analysis (PCA) of functional data combined with a preliminary smoothing, and fit time series models to each component coefficient to obtain forecasts of the forces of mortality. The preliminary smoothing reduces some of the inherent randomness in the observed data. For this purpose we compare the following models described in Table 5.2.

			Estimation method	
Model	Formula	Ref. table	Local lik.	Max. lik.
M1	$D_{x,t} \sim$ Poisson $(E_{x,t}\ \exp(f_1(x,t)))$		M1	
M2	$D_{x,t} \sim$ Poisson $\left(E_{x,t}\ \exp(f_2(\log(\varphi_x^{\text{ref}}(t))))\right)$	INSEE	M2.A	
		TG05	M2.B	
M3	$D_{x,t} \sim$ Poisson $\left(E_{x,t}\ \varphi_x^{\text{ref}}(t)\ \exp(f_1(x,t))\right)$	INSEE	M3.A	
		TG05	M3.B	
M4	$D_{x,t} \sim$ Poisson $\left(E_{x,t}\ \exp(f_1(x,t) + f_2(\log(\varphi_x^{\text{ref}}(t))))\right)$	INSEE	M4.A	
		TG05	M4.B	
M5	logit $\varphi_x(t) = \alpha + \beta$ logit $\varphi_x^{\text{ref}}(t) + \epsilon_{x,t}$	INSEE		M5.A
		TG05		M5.B

Table 5.2: *Description of the models and estimation method used in the first step.*

The functions f_1 and f_2 are unspecified smooth functions of attained age x and calendar year t, and forces of mortality according to a reference table $\varphi_x^{\text{ref}}(t)$, respectively. Model M1 is an endogenous non-parametric approach. Model M2 is an exogenous non-parametric relational model. Models M3 and M4 are mixtures of endogenous and exogenous approaches. Model M3 includes the expected number of deaths $E_{x,t}\ \varphi_x^{\text{ref}}(t)$ according to an external reference table. In model M4, f_1 targets cells for which enough exposure is available (and $f_2 \approx 0$) whereas f_2 allows to borrow strength from the reference prospective table when the exposure is too limited ($f_1 \approx 0$). Finally model M5 is a semi-parametric Brass-type relational model.

The models M1, M2, M3 and M4 are estimated by non-parametric methods. We considered local kernel-weighted log-likelihood methods presented in Chapter 3 to estimate the smooth functions $f_1(x,t)$ and $f_2(\log(\varphi_x^{\text{ref}}(t)))$ for $x \in [x_1, x_n]$ and $t = 1, \ldots, m$. The extrapolation, for $t = m+1, \ldots, m+h$, relies only on the information contained in the smoothed surface. It is performed by identifying the mortality components and their importance over time using functional PCA. Then time series methods are used to extrapolate the time-varying coefficients. In model M5, the logits of the crude forces of mortality are regressed on the logits of the forces of mortality according to a reference table. The estimation is done by minimizing a weighted distance between the estimated and observed forces of mortality. We refer to Planchet and Thérond (2011, Ch.7) for details. Moreover, M5 has the advantage of integrated estimation and forecasting, as the parameters α and β are constant.

Finally, we consider Model (6) in Thatcher (1999, p9) to complete the tables until age 120: logit $\varphi_x(t) \approx \log(\alpha_t) + \beta_t\, x$. It is a robust model that has been found to give good results when fitted to data below age 100 and then extrapolated to higher ages.

We consider two external prospective tables for the first step of our approach as references for the relational models. One is the national demographic projections for the French population over the period 2007-2060, provided by the French National Office for Statistics, INSEE, Blanpain and Chardon (2010). These projections are based on assumptions concerning fertility, mortality and migrations. We choose the baseline scenario among a total of 27 scenarios. The baseline scenario is based on the assumption that until 2060, the total fertility rate is remaining at a very high level (1.95). The decrease in gender-specific and age-specific mortality rates is greater for men over 85 years old. The baseline assumption on migration consists in projecting a constant annual net-migration balance of $100,000$ inhabitants. We complete this table by adding the years 1996-2006 from a previous INSEE table. The tables being relatively wiggly, we smoothed the forces of mortality of the completed table using local kernel weighted log-likelihood. The second external reference table, denoted TG05, is a market table built for the entire French market provided by the French Institute of Actuaries, Planchet (2006). Originally the table is generational and covers the period 1900-2005. We adapted it to our needs and to cover the period 1996-2035. It is worth to mention that this table was constructing using mortality trends originating from the INSEE table where a prudence has been added. As a consequence, this table is not fully faithful to the data but incorporates prudence in an arbitrary manner.

In our first step, we do not take into account the heterogeneity between the different portfolios. The mortality of the entire male population is not specific to any male subpopulation. The second step of our approach is then to build entity specific male mortality tables by adjusting the reference table, validated in the first step, to the mortality of each male portfolio. A Poisson GLM including interactions with age and calendar year gives a solution to this problem.

5.3 Extrapolative method

Stochastic methods of mortality forecasting have received considerable attention, see Booth (2006) and Booth and Tickle (2008) for recent reviews. The most widely used are those involving some forms of extrapolation often using time series methods. Extrapolative methods assume that future trends will essentially be a continuation of the past. In mortality forecasting, this is usually a reasonable assumption because of historical regularities. Functional data methods fall into this category, but they have only recently been adopted in mortality forecasting, see Hyndman and Ullah (2007) and Hyndman and Booth (2008).

Lee-Carter or its variants are now the dominant methods of mortality forecasting in actuarial sciences. The Lee-Carter method, Lee and Carter (1992), has a number of advantages, among them simplicity. The Lee-Carter method involves using the first principal component of the log-mortality matrix. In contrast to parametric approaches which specify the functional form of the age pattern of mortality in advance, principal components approaches estimate the age pattern from the data. Improvements to the Lee-Carter estimation basis have been proposed. A Poisson log-likelihood approach has been developed in Brouhns *et al.* (2002b), Brouhns *et al.* (2002a) and Renshaw and Haberman (2003) to remedy to some of the drawbacks of the Lee-Carter approach, such as for instance the assumed homoskedasticity of the errors. Cosette *et al.* (2007) use a binomial maximum likelihood, and a negative binomial version of the Lee-Carter model has been developed by Delwarde *et al.* (2007) to take into account the over-dispersion phenomenon. The methodology proposed by Hyndman and Ullah (2007) and Hyndman and Booth (2008) can be considered as a successor to the Lee-Carter estimation in that it also involves a principal component decomposition of the mortality surface. However the approach differs in that it uses the functional data paradigm, see Ramsay and Silverman (2005).

Semi-parametric relational models such as M5 have the advantage of integrated estimation and forecasting. This section covers the extrapolation method for the smooth surfaces obtained by local likelihood for models M1, M2, M3 and M4. The extrapolation is performed by identifying the mortality components and their importance over time using functional data analysis, see Ramsay and Silverman (2005, CH.8) and Hyndman and Ullah (2007). Time series methods are used to extrapolate the time-varying coefficients. It can be summarized as follows:

i. Smooth the aggregated data using non-parametric local kernel-weighted log-likelihood to estimate $\varphi_x(t)$ for $x \in [x_1, x_n]$ and $t = 1, \ldots, m$.

ii. Decompose the smoothed surfaces via a basis function expansion using the following model:

$$y_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\ \phi_k(x) + \varepsilon_t(x) \text{ with } \varepsilon_t(x) \sim \text{Normal}\,(0, \upsilon(x))\,, \tag{5.2}$$

where $y_t(x) = \log \widehat{\varphi}_x(t)$, μ_x is the mean of $\log \widehat{\varphi}_x(t)$ across years and $\{\phi_{k,x}\}$ is a set of orthonormal basis functions.

iii. Fit ARIMA models to each of the coefficients $\{\beta_{t,k}\}$, $k = 1, \dots, K$.

iv. Extrapolate the coefficients $\{\beta_{t,k}\}$, $k = 1, \dots, K$, for $t = m+1, \dots, m+h$ using the fitted time series models.

v. Use the forecast coefficients with (5.2) to obtain forecasts of $y_t(x)$, $t = m+1, \dots, m+h$, and hence of $\varphi_x(t)$.

A smoothed version of principal component analysis for functional data is discussed in Silverman (1996). Following the approach of Ramsay and Dalzell (1991) and Hyndman and Ullah (2007), we prefer smoothing the observed data first rather than smoothing the principal component directly to place relevant constraints on the smoothing more easily.

5.3.1 Functional principal components analysis

The decomposition using an orthonormal basis (step ii.) is obtained via functional principal components analysis developed by Ramsay and Dalzell (1991). In the following, we proceed similarly to Hyndman and Ullah (2007). A more general presentation can be found in Ramsay and Silverman (2005, Ch.8).

We want to find a set of K orthonormal functions $\phi_k(x)$ so that the expansion of each curve in terms of the basis functions approximates the curve as closely as possible. For a given K, the optimal orthonormal basis functions $\{\phi_k(x)\}$ minimize the mean integrated squared error

$$\text{MISE} = \frac{1}{n} \sum_{t=1}^{m} \int \varepsilon_t^2(x)\ dx$$

This basis set provides informative interpretation and coefficients $\{\beta_{t,k}\}$ that are uncorrelated, simplifying the forecasting method as multivariate time series models are not required.

The parameter $\mu(x)$ is estimated as the mean of $\log \widehat{\varphi}_x(t)$ across years. Then we estimate $\{\beta_{t,k}\}$ and $\{\phi_k(x)\}$ using a principal components decomposition of $\widehat{y}_t^{\,*}(x) = \widehat{y}_t(x) - \widehat{\mu}(x)$. Our aim is to find the functions $\phi_k(x)$ that maximize the variance of the scores

$$z_{t,k} = \int \phi_k(x) \widehat{y}_t^{\,*}(x)\ dx,$$

subject to the constraints

$$\int \phi_k^2(x)\,dx = 1 \quad \text{and} \quad \int \phi_k(x)\phi_{k-1}(x)\,dx = 0 \text{ if } k \geq 2.$$

These are defined iteratively for $k = 1, \ldots, K$ where $k \leq m-1$. The number K of basis functions depends on many consideration, as explained by Ramsay and Silverman (2005). It depends on the number of discrete points m in the original data, whether some level of smoothing is imposed by using $K < m$, on the efficiency of the basis functions in reproducing the behavior of the original functions, and so on. For our application, 12 sampling points are available per curve and actually for these data a value of K as small as 3 captures most of the interesting variation in the original data.

Assume that we can rewrite each smoothed function $\widehat{y}_t^{\,*}(x)$ in an alternative basis expansion

$$\widehat{y}_t^{\,*}(x) = \sum_{j=1}^{p} a_{t,j}\xi_j(x).$$

We denote $\boldsymbol{A}$ the $m \times p$ matrix of the coefficients $a_{t,j}$. Let $\boldsymbol{J}$ be a $p \times p$ matrix with (i,k)th element $J_{ik} = \int \xi_i(x)\xi_k(x)\,dx$. We find the Choleski decomposition $\boldsymbol{J} = \boldsymbol{U}^T\boldsymbol{U}$ and define

$$\phi_k(x) = \left(\boldsymbol{U}^{-1}\boldsymbol{g}^{(k)}\right)^T \boldsymbol{\xi}(x),$$

where $\boldsymbol{g}^{(k)}$ is the kth normalized eigenvector of $(\boldsymbol{U}^{-1})^T\boldsymbol{JSJ}^T\boldsymbol{U}^{-1}$, $\boldsymbol{S} = (m-1)^{-1}\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{\xi}(x) = \left(\xi_1(x), \ldots, \xi_p(x)\right)^T$. Now, if $\boldsymbol{\Phi}$ denotes an $n \times (m-1)$ matrix with (i,k)th element $\phi_k(x_i)$, and $\boldsymbol{Y}$ is a $m \times n$ matrix with (t,i)th element $\widehat{y}_t^{\,*}(x)$, then $\widehat{\beta}_{t,k}$ is the (t,k)th element of $\boldsymbol{\beta} = \boldsymbol{Y\Phi}$.

This procedure is a simplified version of the approach presented in Hyndman and Ullah (2007). In addition, the authors propose a robust method to avoid difficulties with outlying years. For the presentation of their approach, we refer to the mentioned article.

5.3.2 Extrapolation of the time-varying coefficients

The estimated $\beta_{t,k}$'s can be extrapolated using Box-Jenkins time series methods. The Box-Jenkins approach is one of the most powerful forecasting techniques available and it can be tailored to analyze almost any set of data.

We need to forecast $\beta_{t,k}$ for $k = 1, \ldots, K$ and $t = m+1, \ldots, m+h$. For $K > 1$ this is a multivariate time series problem. However, as mentioned previously, because of the way the basis functions $\phi_k(x)$ have been chosen, the coefficients $\widehat{\beta}_{t,k}$ and $\widehat{\beta}_{t,l}$ are uncorrelated for $k \neq l$. As a consequence, univariate time series methods are adequate for forecasting each series $\{\widehat{\beta}_{t,k}\}$. It is expressed through the development of an ARIMA(p,d,q) model where p, d, and q are integers, greater than or equal to zero and refer to the order of the autoregressive, integrated and moving average parts of the model.

Given the time series $\{\widehat{\beta}_{t,k}\}$, where t is an integer index, an ARIMA (p,d,q) model is described by

$$(1-B)^d\phi(B)\widehat{\beta}_{t,k} = \theta(B)Z_t, \quad \text{and} \quad \{Z_t\} \sim \text{White Noise } (\sigma^2), \qquad (5.3)$$

where B is the backshift operator, $B\ \widehat{\beta}_{t,k} = \widehat{\beta}_{t-1,k}$, expressing the length of the previous data that the model uses to provide the forecasts, and $\phi()$ and $\theta()$ are polynomials of degrees p and q respectively.
The parameter d controls the level of differencing. If $d = 0$, the ARIMA is equivalent to an ARMA model. If $d \geq 1$, we can add an arbitrary polynomial trend of degree $(d-1)$ to $\{\widehat{\beta}_{t,k}\}$, without violating the difference equation (5.3). Therefore, ARIMA models are useful for representing data with trend. The AR stands for autoregressive and describes a stochastic process that can be described by a weighted sum of its previous values and a white noise error, while MA stands for moving average and describes a stochastic process that can be described by a weighted sum of a white noise error and the white noise error from the previous periods.

We consider a full range of ARIMA(p,d,q) models with $d = 0, 1, 2$ and $p, q = 0, 1, 2, 3, 4$ as candidates for the period effects. The Bayes information criterion (BIC) is calculated for each ARIMA model and, on the basis of this information, the parameters p, d and q are selected. We refer the reader to Brockwell and Davis (2002) for a useful theoretical introduction to time series methods and to Delwarde and Denuit (2003) for an exhaustive application to the Lee-Carter model.

Then, extrapolated forces of mortality are derived using estimated $\mu(x)$ and $\mathbf{\Phi}$, the set of the basis functions, and extrapolated $\{\beta_{t,k}\}$. Then conditioning on the observed data $\mathcal{J} = \{\varphi_t(x_i);\ t = 1, \ldots, m; i = 1, \ldots, n\}$ and on the set of the basis function $\mathbf{\Phi}$, we deduce h-step ahead forecasts of $\varphi_{m+h}(x)$

$$\widehat{\varphi}_{m,h}(x) = \mathbb{E}\left[\varphi_{m+h}(x)|\mathcal{J}, \mathbf{\Phi}\right] = \widehat{\mu}(x) + \sum_{k=1}^{K} \tilde{\beta}_{m,h,k}\ \widehat{\phi}_k(x),$$

where $\tilde{\beta}_{m,h,k}$ denotes the h-step ahead forecasts of $\beta_{m+h,k}$ using the estimated series $\widehat{\beta}_{1,k}, \ldots, \widehat{\beta}_{m,k}$. Hyndman and Ullah (2007) and Hyndman and Booth (2008) provide a procedure to approximate the forecast variance. We refer to the mentioned articles for the presentation of their method.

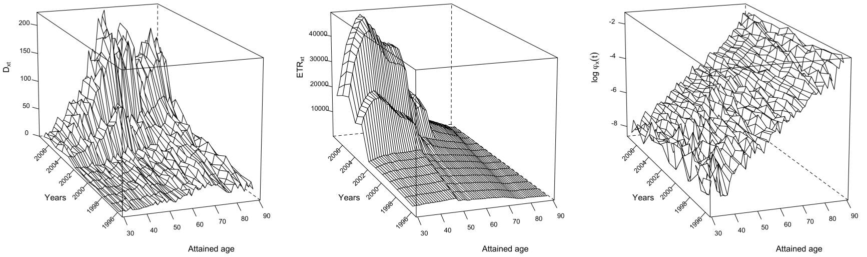
5.4 Construction of a global prospective table

From our collection of portfolios originating from several insurance companies presented in Section 5.2.3, the first step consists in constructing global prospective mortality tables by gender.

By reasoning globally, these tables summarize the mortality experience of these portfolios. We focus on the measurements of the forces of mortality as a function of the attained age x and the calendar year t.

5.4.1 The aggregated data

We aggregate the portfolios by attained age x and calendar year t. The range of attained ages is 30-90 and the observations cover the period 01/01-/1996-31/12/2007. Figures 5.2 displays the observed statistics of the aggregated datasets for the male population.



(a) $D_{x,t}$ **(b)** $E_{x,t}$ **(c)** $\log \varphi_x(t)$

Figure 5.2: *Observed surfaces of the aggregated datasets, male population.*

For years 1996 to 2002 only portfolios P1, P4 and P6 are contributing to the surface. After the year 2002, we observe an increase of the number of deaths and exposures due to the aggregation of the other portfolios. As a consequence, the structure of the heterogeneity is changing over time. It may impact the estimation of the mortality trend over the years and ideally we should have stuck to the same structure of the heterogeneity. By aggregating the portfolios, we are therefore making a trade-off between the constitution of a relatively long history and a situation where the structure of the heterogeneity would be stable.

5.4.2 Comparisons of the fits

We fitted the models presented in Table 5.2. Figure 5.3 displays the fits in the log scale for the 9 models over the years for several ages. It gives us the opportunity to visualize the similarities and differences between the smoothed surfaces obtained by the models.

Figure 5.3a shows the smoothed fits at attained age 30. It is apparent that the relational models using the table TG05 as reference lead to higher forces of mortality at this age while the models using the national population table originating from INSEE produce smoothed fits in the neighborhood of the endogenous model M1.

In Figure 5.3b, the decreasing trend of the forces of mortality over the years is sharper at age 40 for models using the national population table. Moreover, compared to the smoothed fits at attained age 30, we observe that the models using the national population table lead to higher forces of mortality than the models having the table TG05 as reference. This fact remains true for ages 50, Figure 5.3c, and 60, Figure 5.3d.

At attained age 50, Figure 5.3c, the fully exogenous non-parametric models M2 and semi-parametric models M5 lead to similar graduation when using the market national population table. Similar remarks can be made with respect to the reference table used, for the models M3 and M4, being mixtures between endogenous and exogenous modeling. Because the models having an exogenous component rely on the general shape of the reference table, the decreasing trend observed for models M2, M3, M4 and M5 is mostly linear. But for the fully endogenous model M1, we observe a non-linear trend.

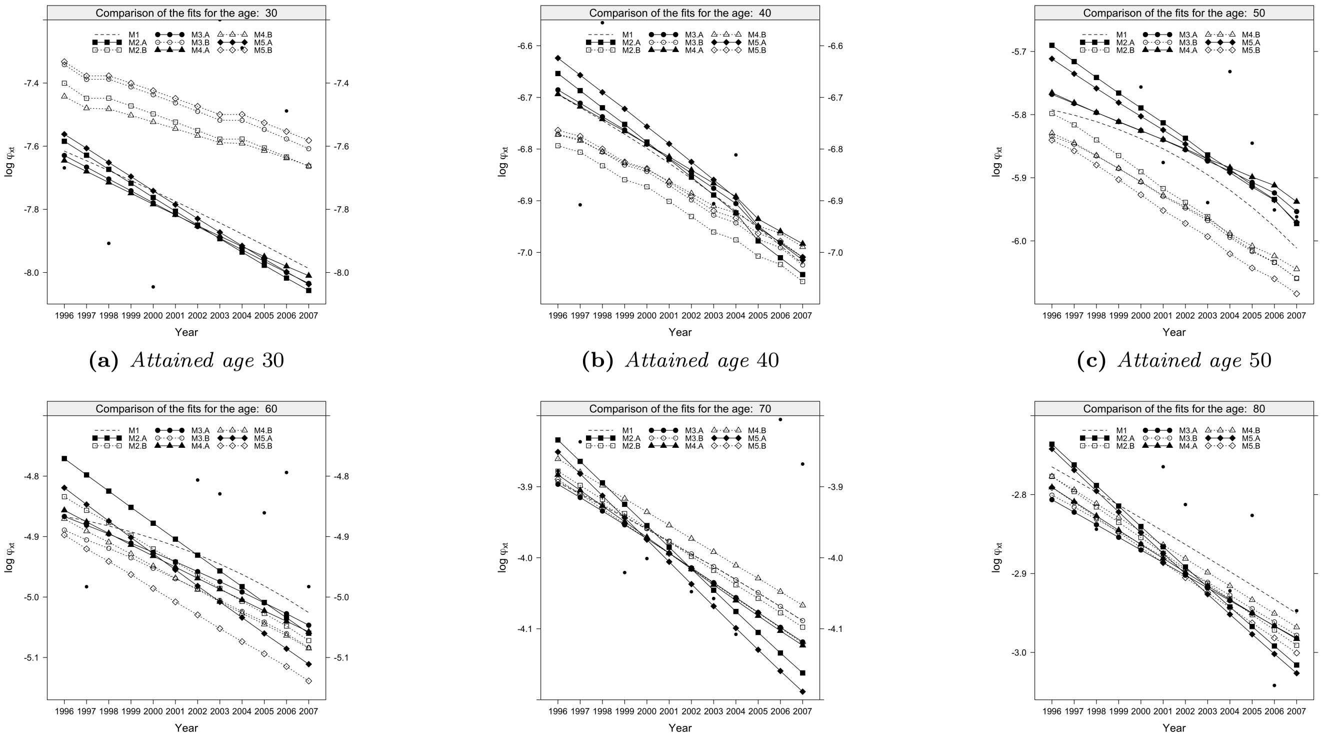
In Figures 5.3d, 5.3e and 5.3f, the decreasing trend of the forces of mortality is steeper for the fully exogenous models M2 and M5 than models having an endogenous component.

We observe that the models have the following features in common. The overall level of mortality has been declining over time and these improvements have been greater at lower ages than at higher ages. However the models diverge in the speed of the improvement. The fully exogenous models M2 and M5 estimate a steeper decrease of the forces of mortality than models M1, M3 and M4 using an endogenous component. The models using a mixture of endogenous and exogenous modeling M3 and M4 behave similarly with respect to the reference table used. At the extreme ages, the models using the market table lead to higher estimated forces of mortality, while for ages in the center, the models using the national population table yield higher estimates. It gives us a first indication of the degree of model risk. In the following section, these visual comparisons are supplemented by a range of quantitative diagnostics which will increase our confidence in some models and question the suitability of others for our purposes.

5.4.3 Tests and quantities to compare graduations

We now carry out a number of tests to assess the impact of model choice. We apply the tests proposed by Forfar *et al.* (1988, p.56-58) and Debón *et al.* (2006, p.231). We have also obtained the values of the mean absolute percentage error $MAPE$ and R^2 used in Felipe *et al.* (2002). In addition, we compute the relative difference between the observed number of deaths and the expected number of deaths obtained by the models $\mathrm{SMR}-1$, where the standardized mortality ratio (SMR) is defined by

$$\mathrm{SMR} = \frac{\sum_{(x,t)} E_{x,t}\, \widehat{\varphi}_x(t)}{\sum_{(x,t)} E_{x,t}\, \widehat{\varphi}_x^{ref}(t)} = \frac{\sum_{(x,t)} D_{x,t}}{\sum_{(x,t)} E_{x,t}\, \widehat{\varphi}_x^{ref}(t)}, \tag{5.4}$$



(a) *Attained age* 30

(b) *Attained age* 40

(c) *Attained age* 50

(d) *Attained age* 60

(e) *Attained age* 70

(f) *Attained age* 80

Figure 5.3: *Comparisons of the Fits for several attained ages, log scale, male population.*

for (x, t) in the set of ages and calendar years of interest. We compare the crude forces of mortality rates to the graduated series to see whether the approaches lead to similar graduation. Table 5.3 presents the results.

		M1	M2.A	M2.B	M3.A	M3.B	M4.A	M4.B	M5.A	M5.B
Fitted DF		7.56	4.02	3.94	4.88	4.88	4	4	NA	NA
Deviance		1302.53	1302.83	1328.04	1296.58	1351.66	1313.93	1366.80	1355.92	1417.08
Standardised	> 2	86	84	92	86	93	87	94	86	100
residuals	> 3	24	22	24	23	24	23	26	26	31
Signs	$+(-)$	327(405)	319(413)	336(396)	345(387)	343(389)	334(398)	337(395)	333(399)	368(364)
test	p-value	0.0043	0.0005	0.0291	0.1296	0.0962	0.0198	0.0350	0.0162	0.9117
Runs	Nb of runs	346	326	322	328	302	332	306	334	319
test	Value	-1.55	-3.00	-3.46	-3.17	-5.02	-2.77	-4.82	-2.61	-3.91
	p-value	0.1188	0.0026	$5.28e-4$	0.0014	$5.25e-7$	0.0056	$1.46e-6$	0.0089	$8.86e-5$
Kolmogorov	Value	0.0327	0.0286	0.0601	.0286	0.0642	0.0286	0.0614	0.0300	0.0683
Smirnov test	p-value	0.8262	0.9239	0.1419	0.9239	0.0978	0.9239	0.1257	0.8955	0.0657
χ^2		1400.19	1402.53	1418.89	1405.81	1445.53	1421.85	1466.93	1473.63	1545.42
R^2		0.9326	0.9256	0.9325	0.9312	0.9325	0.9302	0.9306	0.9221	0.9306
$MAPE$ (%)		25.86	26.81	26.41	25.86	26.69	26.01	26.51	26.84	26.63
SMR $-$ 1 (%)		-0.79	0.29	0.21	0.11	0.30	$1.22e-12$	$-7.77e-13$	1.87	2.87

Table 5.3: *Comparisons between the smoothing approaches.*

The approaches display different results. Model M1, having the highest degrees of freedom and being fully endogenous, has the capacity to reveal many features in the data. Therefore, it has the highest number of runs, lowest χ^2 and $MAPE$ and highest R^2. We observe that M1 is the only model to lead to a higher number of expected deaths than observed. Conversely, the fully exogenous semi-parametric models M5, and to a lesser degree the non-parametric M2, lead to higher deviance, higher χ^2, lower R^2, higher number of standardized residuals exceeding the thresholds 2 and 3 and higher relative difference between expected and observed number of deaths.

We observe that the fully exogenous models M2 and M5 do not behave similarly. The non-parametric models M2, being more flexible, perform better than the semi-parametric models M5. With respect to the reference table used, models M2 have a lower deviance, lower number of standardized residuals exceeding the thresholds 2 and 3, lower χ^2 and $MAPE$, and higher R^2. Also, the expected and observed number of deaths are closer.

The mixtures of endogenous and exogenous modeling M3 and M4 have similar results with respect to the reference table used. Nevertheless, models M3, including the expected number of deaths according to the reference table INSEE or TG05, perform better than models M4. Models M3 have a better spread of the residuals between positive and negative signs, higher value for the sign test, lower deviance, χ^2 and $MAPE$. However, models M4 have the smallest relative difference between expected and observed number of deaths.

We observe that, in general, models incorporating the national population table originating from INSEE (models A) produce graduations that are closer to the data than models using the market table TG05 (models B) as reference. Using the market table leads to higher deviance, higher χ^2, lower number of runs and higher number of standardized residuals exceeding the thresholds 2 and 3 compared to models incorporating the national population table. The market table TG05 is derived on mortality trends originating from the INSEE table where a prudence has been added. As a consequence, this table is not fully faithful to the data but incorporates prudence in a arbitrary manner.

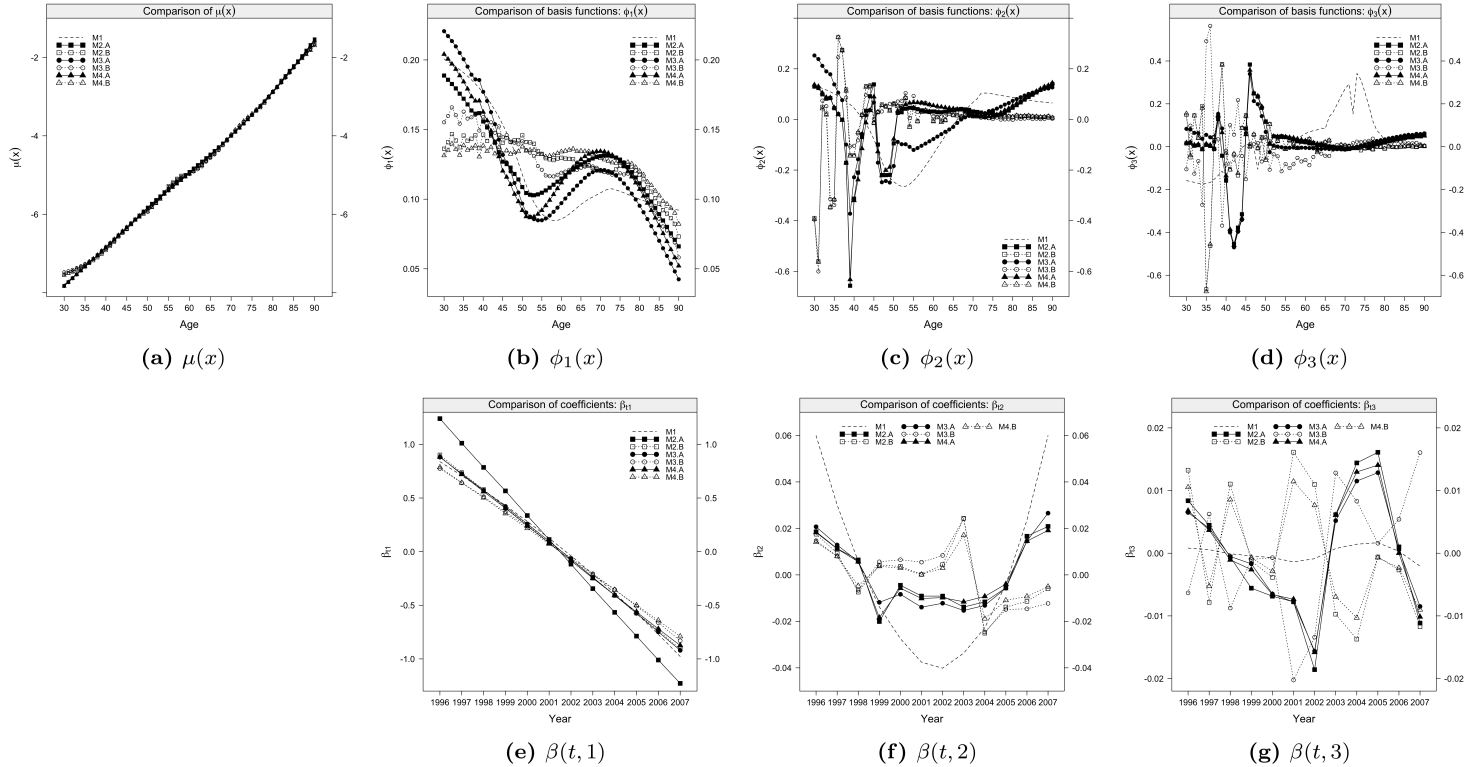
The tests and quantities carried out in Table 5.3 show the strengths and weaknesses of each model to adjust the observed mortality. The choice between the models is only a matter of judgment and depends on the purpose for which the prospective mortality table would be used. It is up to potential users of the table to decide the weights they place on the different criteria. However, regarding the wide ranging set of model selection criteria, we can eliminate some models. We have seen that the non-parametric models, due to their flexibility, ensure a good fit. Hence models M2 would be preferred to M5. Within the mixture of endogenous and exogenous models, M3 would be preferred to M4. Compared to the fully endogenous model and to the fully exogenous models, relying partly on the national population table is beneficial according to the various tests and quantities used in assessing the adjustment of observed mortality, hence model M3.A would be preferred to models M1 and M2.

5.4.4 Extrapolation of the smoothed surfaces and completed tables

The extrapolation of the smoothed surface of models M1, M2, M3 and M4 is performed by identifying the mortality components and their importance over time using functional principal component analysis presented in Section 5.3.1. Then time series methods are used to extrapolate the time-varying coefficients. Model M5 has the advantage of integrated estimation and forecasting.

Figure 5.4 displays the basis functions and associated coefficients using (5.2) for the models M1, M2, M3 and M4. A decomposition of order $K = 3$ has been used.

The average log-mortality at attained ages is similar for the models over time except at the extreme ages, Figure 5.4a. Models using the market mortality table TG05 as reference lead to higher mortality around age 30 compared to models using the national population mortality table, as observed previously in Figure 5.3a.



(a) $\mu(x)$

(b) $\phi_1(x)$

(c) $\phi_2(x)$

(d) $\phi_3(x)$

(e) $\beta(t,1)$

(f) $\beta(t,2)$

(g) $\beta(t,3)$

Figure 5.4: *Basis functions and associated coefficients with $K = 3$, for models M1, M2, M3 and M4, male population.*

Figure 5.4b shows the first basis function for all models. The first term accounts for at least 99.6 % of the variation in mortality. The coefficient, Figure 5.4e, indicates a fairly steady decline in mortality over time. The models leads to more or less the same results except for model M2.A that gives the steepest decrease. The models M3 and M4 using a mixture of endogenous and exogenous modeling produce similar results with respect to the reference table used. The basis function $\phi_1(x)$ indicates that the decline has been faster for the young adults and at ages $60-80$ for the models using the national population table originating from INSEE (models A) as well as a fully endogenous model M1. But for models using the market table TG05 (models B) the decrease has been steady for ages $30-80$. We observe that models M3 lead to the fastest and slowest improvement of the mortality for the young adults and individuals above 80, respectively.

The basis function $\phi_2(x)$, displayed in Figure 5.4c, models the differences between the young adults and those over 75. The coefficients in Figure 5.4f shows that this difference in mortality has falling from the beginning of the period of investigation to 2002 - starting date of observation of additional portfolios - and increasing since 2002 to the end of the period of investigation.

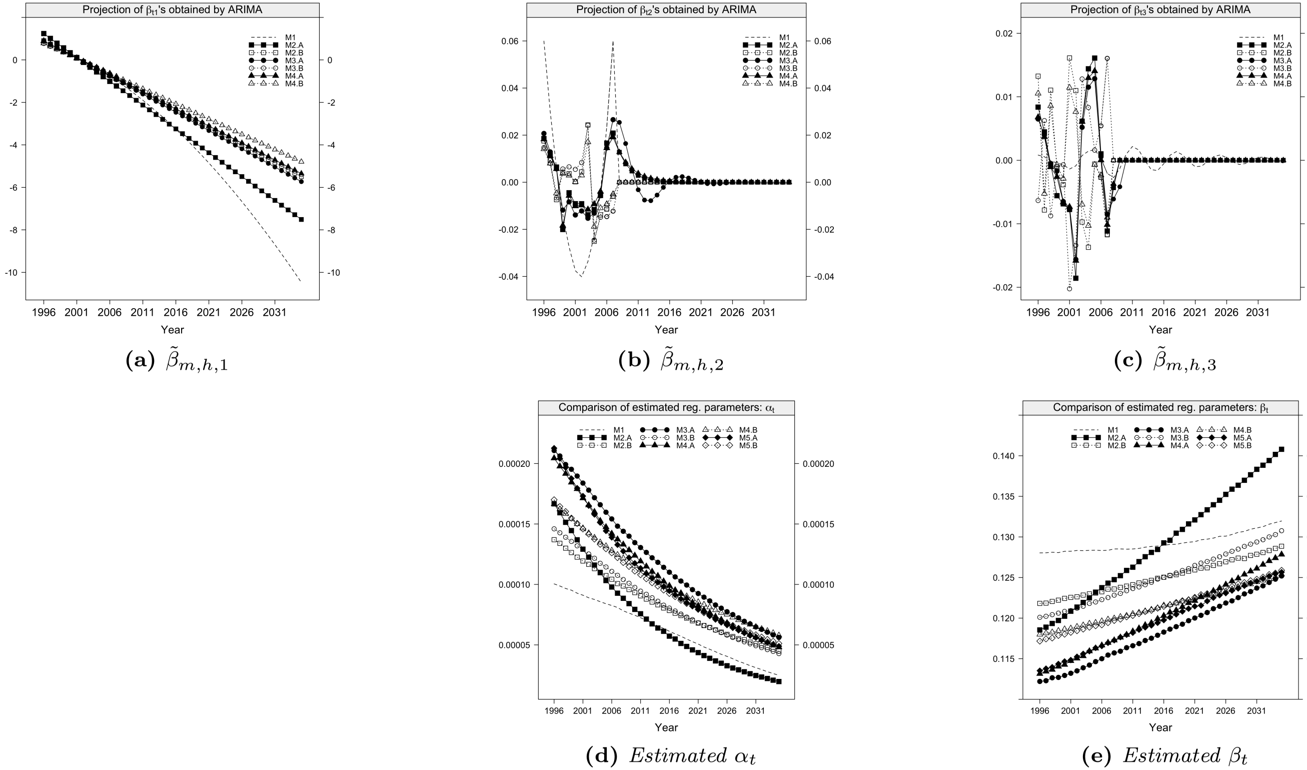
Similarly, Figure 5.4d displays difference between the young adults (up to 50) and those over 80. However, the shape of associated coefficient Figure 5.4g is more irregular than $\beta_{t,2}$. Again we observed that the choice of the reference table used leads to a different pattern of the basis functions and associated coefficients.

The time-varying coefficients are forecast using univariate time series methods. Table 5.4 summarizes the ARIMA models, introduced in Section 5.3.2. For each of the models M1, M2, M3 and M4, we considered a full range of ARIMA(p,d,q) models with $d = 0, 1, 2$ and $p, q = 0, 1, 2, 3, 4$ as candidates for the period effects. The Bayes information criterion (BIC) was calculated for each ARIMA model and, on the basis of this information, the parameters p, d and q have been selected. Figure 5.5 displays the resulting projections for models M1, M2, M3 and M4 for $h = 28$, that is until year 2035. For clarity, the confidence intervals are omitted.

We notice that the coefficients $\tilde{\beta}_{m,h,2}$ and $\tilde{\beta}_{m,h,3}$ in Figures 5.5b and 5.5c are rapidly constant. As a consequence, we could have performed a decomposition using the first principal component as in the original Lee-Carter method. However, it may not be the case for other datasets, as illustrated in Hyndman and Ullah (2007) and Hyndman and Booth (2008). The use of several components is the main difference between this approach and the Lee-Carter method, which uses only the first component and also involves an adjustment. The extra principal components allow more accurate forecasting of age-specific forces of mortality, though in our application at least 99.6 % of the variation is explained by the first component.

Model & component		Model for the $\beta_{t,k}$	
M1	$k = 1$	ARIMA(1,2,1)	$\beta_t = 2\,\beta_{t-1} - \beta_{t-2} + \mu + \phi(\beta_{t-1} - 2\,\beta_{t-2} + \beta_{t-3} - \mu) + Z_t + \theta\, Z_{t-1}$
M1	$k = 2$	ARIMA(0,0,0) with non-zero mean	$\beta_t = Z_t$
M1	$k = 3$	ARIMA(2,0,1) with zero mean	$\beta_t = \mu + \phi_1(\beta_{t-1} - \mu) + \phi_2(\beta_{t-2} - \mu) + Z_t + \theta\, Z_{t-1}$
M2.A & M4.A	$k = 1$	ARIMA(0,1,0) with drift	$\beta_t = \beta_{t-1} + d + Z_t$
M2.A & M4.A	$k = 2$	ARIMA(1,0,0) with zero mean	$\beta_t = \mu\phi(\beta_{t-1} - \mu) + Z_t$
M2.A & M4.A	$k = 3$	ARIMA(0,0,1) with zero mean	$\beta_t = Z_t + \theta\, Z_{t-1}$
M2.B & M4.B	$k = 1$	ARIMA(1,1,0) with drift	$\beta_t = \beta_{t-1} + \mu + d + \phi(\beta_{t-1} - \beta_{t-2} - \mu) + Z_t$
M2.B & M4.B	$k = 2, 3$	ARIMA(0,0,0) with zero mean	$\beta_t = Z_t$
M3.A	$k = 1$	ARIMA(0,2,0)	$\beta_t = 2\,\beta_{t-1} - \beta_{t-2} + Z_t$
M3.A	$k = 2$	ARIMA(2,0,0) with zero mean	$\beta_t = \mu + \phi_1(\beta_{t-1} - \mu) + \phi_2(\beta_{t-2} - \mu) + Z_t$
M3.A	$k = 3$	ARIMA(0,0,2) with zero mean	$\beta_t = Z_t + \theta_1\, Z_{t-1} + \theta_2\, Z_{t-2}$
M3.B	$k = 1$	ARIMA(1,2,0)	$\beta_t = 2\,\beta_{t-1} - \beta_{t-2} + \mu + \phi(\beta_{t-1} - 2\,\beta_{t-2} + \beta_{t-3} - \mu) + Z_t$
M3.B	$k = 2, 3$	ARIMA(0,0,0) with zero mean	$\beta_t = Z_t$

Table 5.4: *Description of the models for the time-varying coefficients, male population.*



(a) $\tilde{\beta}_{m,h,1}$

(b) $\tilde{\beta}_{m,h,2}$

(c) $\tilde{\beta}_{m,h,3}$

(d) *Estimated* α_t

(e) *Estimated* β_t

Figure 5.5: *Projections of the estimated coefficients* $\beta_{t,k}$ *for the models M1, M2, M3 and M4 obtained by ARIMA, and estimated regression parameters of model* 6 *of Thatcher (1999), male population.*

The next step is to obtain completed tables until age 120. For this, we apply Model (6) in Thatcher (1999, p9) to the forces of mortality to extrapolate the data: logit $\varphi_x(t) \approx \log(\alpha_t) + \beta_t\ x$. It is a robust model that has been found to give good results when fitted to data below age 100 and then extrapolated to higher ages. Figures 5.5d and 5.5e show the estimated regression parameters α_t and β_t, respectively. All models estimate a linear effect of time on the forces of mortality at high ages, Figure 5.5e. We observe that models using the national population table (models A) lead to a steeper increase of the linear component β_t over time, Figure 5.5e, than models using the market table as reference (models B). As a consequence, those models lead to a more rapid increase of the forces of mortality at the highest ages (70-90), which in turn results in a more rapid decrease of the forces of mortality at lower ages. The mixture of endogenous and exogenous modeling models M4 and fully exogenous semi-parametric model M5 produce very similar results, while the fully endogenous model M1 and fully exogenous model M2.A differ largely from the other models.

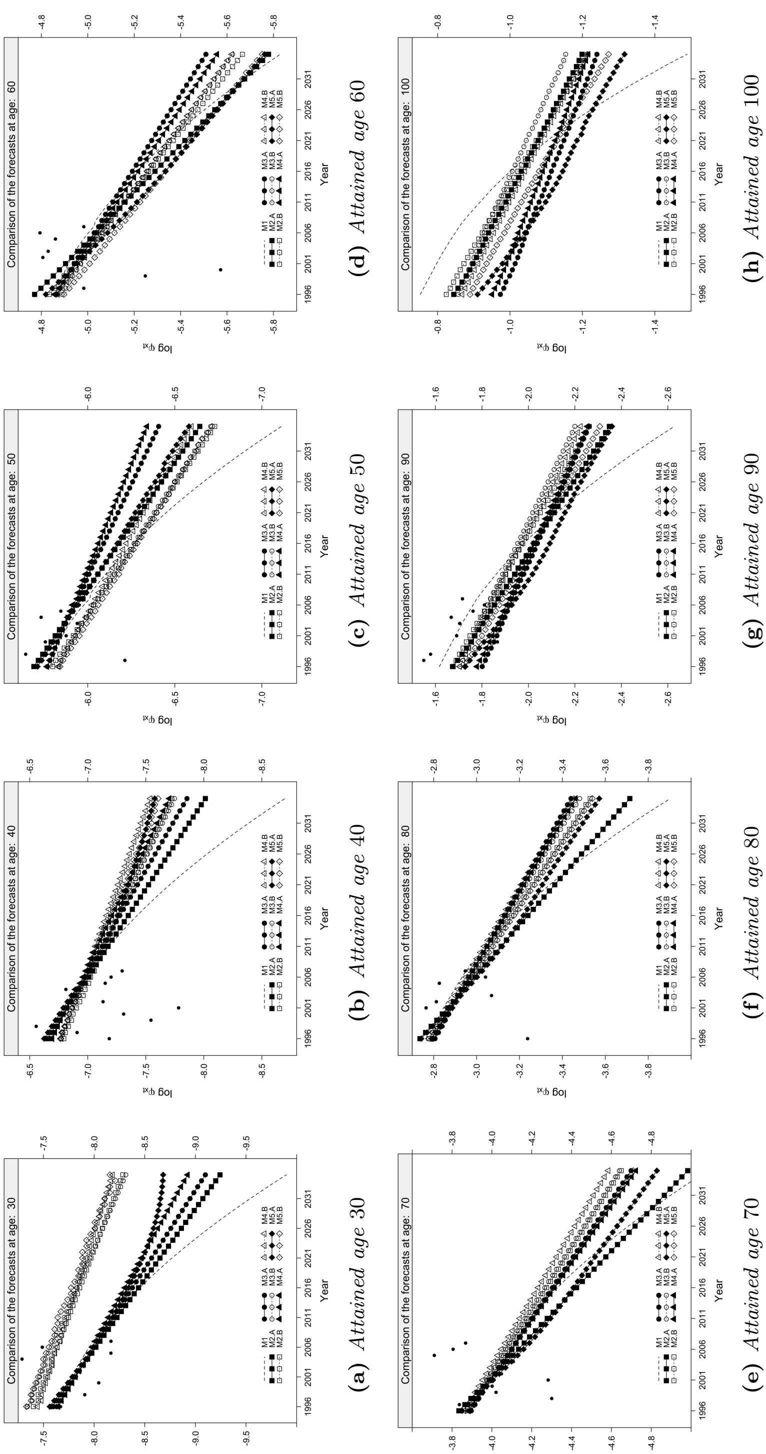
Figure 5.6 displays the fits in the log scale for the 9 models over the years for several ages. For clarity, the confidence intervals are omitted. The forecasts produced here are based on the first three principal components. The additional components may serve to incorporate relatively recent changes in pattern. The use of smoothing prior to modeling results in forecast age patterns that are relatively smooth.

As visualized in Figure 5.3, the overall level of mortality is declining over time and these improvements are greater at lower ages than at higher ages. However the models diverge in the level and speed of the improvement. At the extreme ages, the models using the market table (models B) lead to higher estimated forces of mortality, while for ages in the center, the models using the national population table (models A) yield higher estimates. The fully exogenous models M2 and M5 produce a steeper decrease of the forces of mortality than models M3 and M4. Model M1 stands out, leading to a non-linear decline of the forces of mortality and inducing the sharpest decrease.

5.4.5 Model risk and validation of the final table

We have seen in Figure 5.6 that models diverge in the level and speed of the improvement of the level of mortality across the age. It gives us a first indication of the degree of model risk.

Figure 5.7 shows the survival indexes at several ages computed from the completed tables obtained with the different models. It represents the survival indexes of cohorts aged 30, 40, 50, 60, 70 and 80 in 1996 over 40 years. This measures the proportion from a group of males aged 30, 40, 50, 60, 70 or 80 at the start of 1996 who remain alive for the next 40 years.



(a) *Attained age* 30

(b) *Attained age* 40

(c) *Attained age* 50

(d) *Attained age* 60

(e) *Attained age* 70

(f) *Attained age* 80

(g) *Attained age* 90

(h) *Attained age* 100

Figure 5.6: *Comparisons of the fits and forecasts for several attained ages, log scale, male population.*

It can be seen that these survival indexes are affected by the choice of the modeling, endogenous, exogenous or mixture of the two and to a lesser degree by the choice of the reference table used. Endogenous model M1, and exogenous models M2 and M5 with respect to the reference table used lead to higher survival indexes for cohorts aged 30, 40 and 50 in 1996, Figures 5.7a, 5.7b and 5.7c. For a cohort aged 60, 70 and 80 in 1996, Figures 5.7d, 5.7e and 5.7f, the survival indexes are consistent within the models.

We observe substantial differences in using the market table TG05 or the national population table as reference in the exogenous or mixture models. For cohort aged 30 in 1996, Figure 5.7a, the incorporation of the market table (models B) leads to a higher survival index for the models using of the national population table (models A). Conversely, when incorporating the national population table for cohort aged 40 and 50 in 1996, in Figures 5.7b and 5.7c the survival indexes are higher with respect to the models.

As a second example, we calculate some single figures summarizing the lifetime probability distribution for cohorts at several ages in 1996. Table 5.5 displays the indices.

The mixtures of endogenous and exogenous modeling, models M3 and M4, lead to the smallest partial life expectancies ${}_{40}e_{30}$, ${}_{40}e_{40}$, ${}_{40}e_{50}$ and ${}_{40}e_{60}$ for cohorts aged 30, 40, 50 and 60 in 1996. The fully endogenous model M1 yields the highest partial life expectancies ${}_{40}e_{30}$ and ${}_{40}e_{40}$ but leads to the smallest for cohorts aged 80 in 1996.

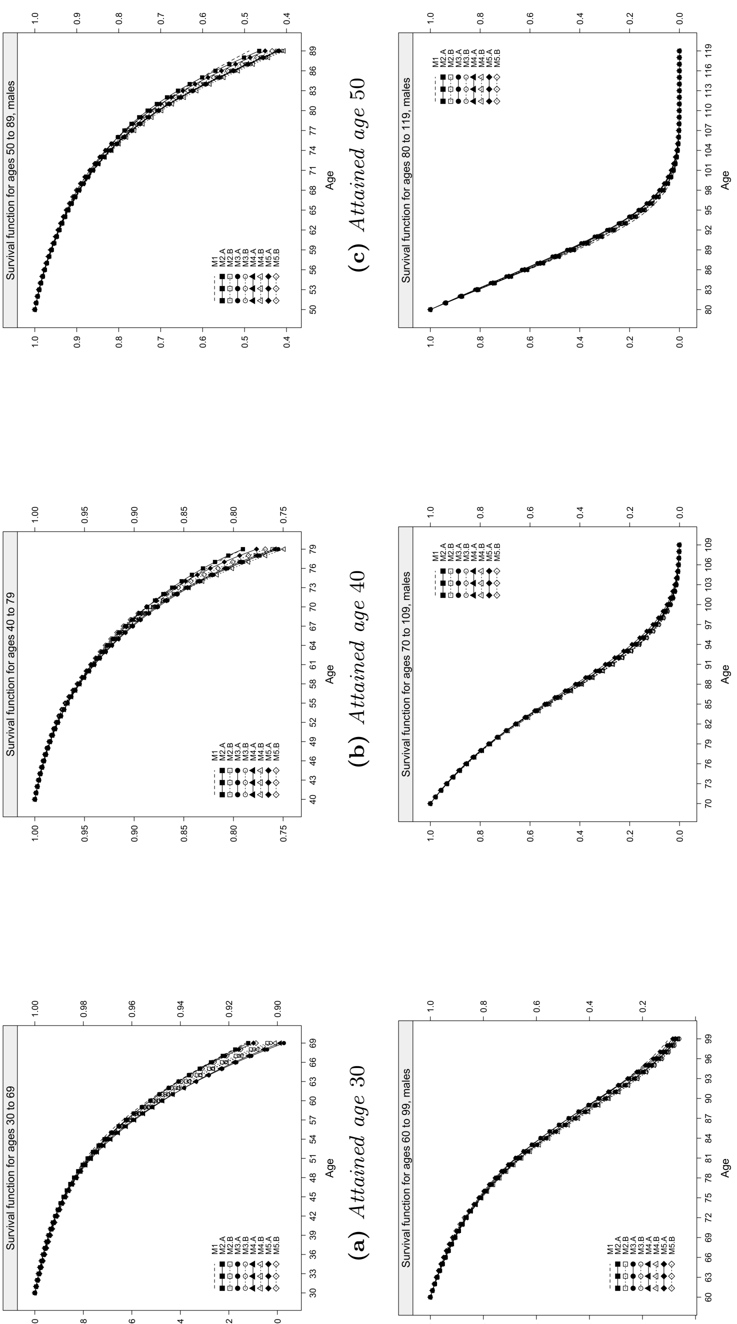
The semi-parametric models M5 produce higher partial life expectancies than the non-parametric models M2 except for ${}_{40}e_{50}$ and ${}_{40}e_{60}$. Similarly, the mixture models M4 yield higher partial life expectancies than models M3 incorporating the expected number of deaths according to a reference table except for ${}_{40}e_{80}$.

We observe, once more, that the choice of the reference table affects the quantities. Using the national population table leads to higher life expectancy than incorporating the market table.

These results can be seen in the median age at death, $\mathrm{Med}({}_{40}T)$. The exogenous models M2 and M5 produce close estimates, and the mixture models M3 and M4 lead to more or less similar results.

The mixture M3 and M4 models stand out as having a much higher standard deviation of the random life time, ${}_{40}\sigma$, than the exogenous model, which would be expected. However it suggests that model risk might be an issue. For example, the price of a financial option that has the survival index as its underlying quantity is strongly dependent on its standard deviation; everything else being equal, the higher the variance, the higher the value of the option, as recalled in Cairns *et al.* (2009).

The entropy $\mathrm{H}({}_{40}T)$ obtained with the exogenous models M2 and M5 is similar, also there is not much difference in $\mathrm{H}({}_{40}T)$ for the mixture M3 and M4 models. However, we notice that using the market table leads to



(a) *Attained age 30*

(b) *Attained age 40*

(c) *Attained age 50*

(d) *Attained age 60*

(e) *Attained age 70*

(f) *Attained age 80*

Figure 5.7: *Survival indexes for cohorts at several ages at the start of* 1996 *over* 40 *years, male population.*

	M1	M2.A	M2.B	M3.A	M3.B	M4.A	M4.B	M5.A	M5.B
${}_{40}e_{30}$	38.80	38.78	38.72	38.65	38.69	38.65	38.69	38.77	38.75
${}_{40}e_{40}$	37.00	36.83	36.62	36.73	36.80	36.77	36.76	36.98	36.95
${}_{40}e_{50}$	32.71	32.74	32.24	32.20	32.16	32.26	32.07	32.65	32.47
${}_{40}e_{60}$	24.43	24.47	24.12	24.12	23.99	24.12	23.89	24.62	24.37
${}_{40}e_{70}$	15.07	15.26	15.11	15.34	15.03	15.28	14.98	15.52	15.28
${}_{40}e_{80}$	7.81	8.01	7.97	8.40	8.02	8.31	8.02	8.28	8.17
Med(${}_{40}T_{50}$)	39.64	39.04	37.89	37.74	37.61	37.83	37.43	38.62	38.17
Med(${}_{40}T_{60}$)	27.03	27.39	26.81	26.77	26.65	26.80	26.48	27.41	27.02
Med(${}_{40}T_{70}$)	16.51	16.83	16.66	16.83	16.59	16.79	16.49	17.02	16.79
Med(${}_{40}T_{80}$)	8.63	8.77	8.75	9.17	8.82	9.08	8.78	8.95	8.93
${}_{40}\sigma_{30}$	0.0258	0.0260	0.0278	0.0300	0.0282	0.0297	0.0285	0.0262	0.0264
${}_{40}\sigma_{40}$	0.0619	0.0604	0.0687	0.0699	0.0695	0.0688	0.0713	0.0630	0.0658
${}_{40}\sigma_{50}$	0.1502	0.1519	0.1664	0.1688	0.1702	0.1676	0.1729	0.1577	0.1636
${}_{40}\sigma_{60}$	0.2934	0.2978	0.3068	0.3047	0.3112	0.3053	0.3101	0.2963	0.3038
${}_{40}\sigma_{70}$	0.3557	0.3549	0.3569	0.3550	0.3578	0.3554	0.3558	0.3524	0.3551
${}_{40}\sigma_{80}$	0.3125	0.3117	0.3127	0.3158	0.3139	0.3150	0.3127	0.3121	0.3138
H(${}_{40}T_{30}$)	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
H(${}_{40}T_{40}$)	0.0020	0.0019	0.0021	0.0021	0.0021	0.0021	0.0021	0.0019	0.0020
H(${}_{40}T_{50}$)	0.0061	0.0060	0.0067	0.0067	0.0068	0.0067	0.0069	0.0061	0.0064
H(${}_{40}T_{60}$)	0.0243	0.0248	0.0265	0.0263	0.0274	0.0264	0.0275	0.0242	0.0254
H(${}_{40}T_{70}$)	0.1166	0.1177	0.1209	0.1136	0.1246	0.1157	0.1223	0.1079	0.1147
H(${}_{40}T_{80}$)	0.5452	0.5280	0.5302	0.4685	0.5337	0.4826	0.5191	0.4680	0.4941

Table 5.5: *Single figure indices to summarize the lifetime probability distributions for cohorts at several ages in* 1996.

less concentrated deaths, while incorporating the national population table yields more stretched deaths.

To have a clear picture of the contribution of model risk to forecast uncertainty, we can make use of the first two robust principal component scores of quantities of interests such as the partial life expectancies of cohorts at several ages in 1996 with the Highest Density Regions (HDR) boxplots of Hyndman (1996). Hyndman and Shang (2010) have proposed this method with identification of outliers in mind. Our idea is to use this graphical method on single figure indices summarizing the lifetime probability distributions, such as the partial life expectancies for cohorts at several ages in 1996, to visualize similarity between the models and outliers and thus model risk in forecast uncertainty.

The bivariate HDR boxplot displays the mode, the highest density point, along with the 50 % inner and 99 % outer highest density regions. All points excluded from the outer HDR are outliers. Figure 5.8 displays the bivariate HDR boxplot of the first two robust principal component scores of the partial life expectancies for cohorts at several ages in 1996.

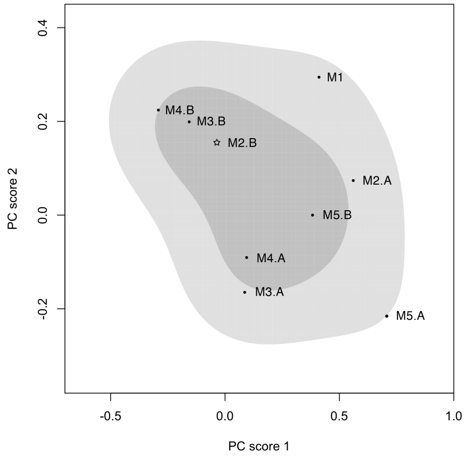


Figure 5.8: *Bivariate HDR boxplot of the first two robust principal component scores of the partial life expectancies for cohorts at several ages in* 1996*, male population.*

The dark and light gray regions show the 50 % HDR and the outer HDR, respectively. The points outside the outer regions are identified as outliers, as model M5.A. The asterisk in Figure 5.8 marks the mode of the bivariate robust principal component scores, corresponding to model M2.B. It shows clearly that the non-parametric models are grouped more by the reference table used and less by the kind of modeling (non-parametric, semi-parametric, endogenous, exogenous and so on).

We have concentrated here on the contribution of model risk in extrapolating the future mortality. However, it is appropriate to allow for parameter uncertainty to provide a more complete picture of the level of risk on the valuations of an insurer, such as provisioning and capital requirement.

The overall model risk associated with a prospective mortality table should ideally take into account two factors,

i. the adjustment according to the past mortality, and

ii. the extrapolation of the future mortality.

From Section 5.4.3, we can eliminate some models, regarding the wide ranging set of model selection criteria. We have seen that the non-parametric models, due to their flexibility, ensure a good fit. Hence models M2 would be preferred to M5. Within the mixture of endogenous and exogenous models, M3 would be preferred to M4. Compared to the fully endogenous model and to the fully exogenous models, relying partly on the national population table is beneficial according to the various tests and quantities used in assessing the adjustment of observed mortality in Table 5.3, hence model M3.A would be preferred to models M1 and M2.

This choice could be refined by analyzing the extrapolated future mortality. We can apply the concept of *biological reasonableness* which was first proposed in Cairns *et al.* (2006) as an aid in assessing the forecasts. This concept is not based on hard scientific, biological or medical facts. It is rather subjective and asks the question where the data are originating from and based on this knowledge, what mixture of biological factors, medical advances and environmental changes would have to happen to cause this particular set of forecasts.
For instance, in Figure 5.6, the projections for model M1, look rather more optimistic than the set of projections of the other models. If we cannot think about any good reason why this might happen, then we must disqualify the model on the basis of *biological reasonableness.* The projections of Model M3.A seem reasonable, in accordance with the set of projections with the other models. Hence, in the following section we adjust the entity specific portfolio experience to the baseline mortality surface obtained by the mixture of endogenous and exogenous modeling M3.A.

5.5 Adjustment to entity specific mortality experience

5.5.1 Entity specific mortality experience

In our first step, we do not take into account the heterogeneity between the different portfolios. The mortality of the aggregated male population is not specific to any male portfolio. We compare the mortality experiences of the 8 portfolios presented in Table 5.1 to the validated table constructed in the first step. The standardized mortality ratio (SMR), as defined in expression (5.4), appears to be a useful index. The observed deaths in a particular portfolio are compared with those that would be expected if the mortality validated in the first step applied. Table 5.6 displays the SMR of the 9 portfolios with the national population reference table originating from INSEE, the market table TG05 and the validated table obtained by the mixture of endogenous and exogenous modeling M3.A.

	P1	P2	P3	P4	P5	P6	P7	P8	mean(|SMR − 1|)
INSEE	−36.93	−57.36	−2.03	−21.22	−29.80	−35.41	−47.11	−42.75	34.08
TG05	52.31	18.09	59.62	34.47	88.43	9.80	0.18	30.39	37.91
M3.A	10.85	−18.67	32.41	10.73	30.98	−5.25	−22.66	−7.50	17.38

Table 5.6: *Relative difference between expected and observed number of deaths by portfolios, ($SMR - 1 \quad (\%)$), male population.*

Table 5.6 illustrates the heterogeneity between the portfolios. We observe that the table validated in the first step under-estimates the number of deaths for portfolios P1, P3, P4 and P5, while it over-estimates the number of deaths for the other portfolios. It should be noted that the national population table constantly over-estimates the number of deaths, but the market table leads to an under-estimation. The relative difference between the observed and expected number of deaths obtained with model M3.A is similar for portfolios P1 and P4, P3 and P5 and P6 and P8, respectively. Relative differences are smaller when using the national population table for P3, or using the market table for P7. However, on average the validated table originating from M3.A leads to the smallest difference in absolute value, illustrating the usefulness of the first step of our approach.

5.5.2 Poisson GLM with age and calendar year interactions

In a Poisson regression, we include the portfolio dummies as a covariate and allow interactions with age and calendar year. We assume that the number of deaths for a portfolio i at attained age x and calendar year t is determined by

$$D_{x,t,i} \sim \text{Poisson}(E_{x,t,i} \; \varphi_x(t,i)),$$

with

$$\log \varphi_x(t,i) = \alpha + \beta \log \widehat{\varphi}_x^{\text{ref}}(t) + \sum_{j=1}^{n} \gamma_j \; \mathbb{I}_i + \sum_{j=1}^{n} \delta_j \; x \; \mathbb{I}_i + \sum_{j=1}^{n} \kappa_j \; t \; \mathbb{I}_i + \sum_{j=1}^{n} \lambda_j \; x \; t \; \mathbb{I}_i \tag{5.5}$$

where $\widehat{\varphi}_x^{\text{ref}}(t)$ is the baseline force of mortality derived in our first step, the $\mathbb{I}_i$'s are binary variables coding the portfolios and n represents the number of portfolios.

If we do not allow for interactions, we will observe parallel shifts of the forces of mortality according to the baseline mortality for each dimension. This view is certainly unrealistic and interactions need to be incorporated. We take the first portfolio P1 as reference level. The relative mortality of the portfolios is expressed with respect to this reference level P1.

We start by incorporating all interactions. We remove the calendar year effect for portfolios P2, P7 and P8, having less than 4 years of observation.

With a parsimonious principle in mind, we progressively exclude the insignificant interactions by computing the drop in deviance test (or likelihood-ratio test) for models with and without the interaction considered.

The final model is the following

$$\begin{aligned}\varphi_x(t,i) = \alpha &+ \beta \log \widehat{\varphi}_x^{\,\mathrm{ref}}(t) \\ &+ \delta_1\, x + \kappa_1\, t + \lambda_1\, x\, t \\ &+ \delta_2\, x\, \mathbb{I}_{i=2} \\ &+ \delta_3\, x\, \mathbb{I}_{i=3} + \kappa_2\, t\, \mathbb{I}_{i=3} \\ &+ \kappa_3\, t\, \mathbb{I}_{i=4} \\ &+ \kappa_4\, t\, \mathbb{I}_{i=5} \\ &+ \gamma_1\, \mathbb{I}_{i=6} + \delta_4\, x\, \mathbb{I}_{i=6} + \kappa_5\, t\, \mathbb{I}_{i=6} \\ &+ \gamma_2\, \mathbb{I}_{i=7} \\ &+ \gamma_3\, \mathbb{I}_{i=8} + \delta_5\, x\, \mathbb{I}_{i=8}.\end{aligned} \tag{5.6}$$

The main effects and interactions included in the final model (5.6) are presented in Table 5.7.

Regression coef.	Parameter est.	Std. error	z value	p value
α	136.2	21.54	6.322	$2.59e-10$
β	1.648	$9.610e-02$	17.153	$< 2e-16$
γ_1	-36.11	9.510	-3.797	0.0001
γ_2	-0.4028	$3.910e-02$	-10.301	$< 2e-16$
γ_3	0.6307	$9.675e-02$	6.519	$7.08e-11$
δ_1	-1.783	0.2929	-6.088	$1.14e-09$
δ_2	$-2.168e-03$	$8.188e-04$	-2.648	0.008
δ_3	$-2.585e-02$	$1.673e-03$	-15.452	$< 2e-16$
δ_4	$-5.658e-03$	$1.105e-03$	-5.122	$3.03e-07$
δ_5	$-1.178e-02$	$1.466e-03$	-8.034	$9.45e-16$
κ_1	$-6.477e-02$	$1.089e-02$	-5.951	$2.67e-09$
κ_2	$1.012e-03$	$6.345e-05$	15.945	$< 2e-16$
κ_3	$-8.897e-05$	$1.743e-05$	-5.104	$3.33e-07$
κ_4	$1.585e-05$	$1.517e-05$	10.445	$< 2e-16$
κ_5	$1.810e-02$	$4.754e-03$	3.807	0.0001
λ_1	$8.638e-04$	$1.467e-04$	5.887	$3.94e-09$

Table 5.7: *Results from the Poisson regression model (5.5), male population.*

Model (5.5) is estimated over the observation period 1996-2007 and for the age range 30-90. The specific prospective mortality tables are now easily derived by incorporating the entire mortality table $\widehat{\varphi}_x^{\,\mathrm{ref}}(t)$ obtained in the first step of our approach. For instance, for portfolio P1, the forces of mortality are given by

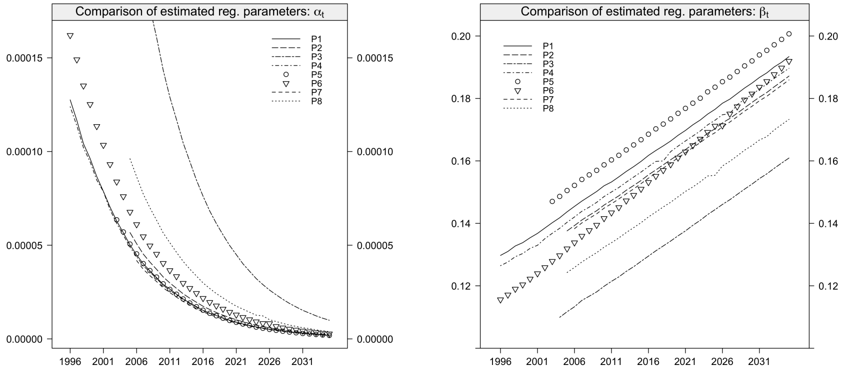
$$\widehat{\varphi}_x(t,1) = \exp\left(\widehat{\alpha} + \widehat{\beta}\log\widehat{\varphi}_x^{\,\mathrm{ref}}(t) + \widehat{\delta}_1\; x + \widehat{\kappa}_1\; t + \widehat{\lambda}_1\; x\; t\right),$$

and for portfolio P6,

$$\widehat{\varphi}_x(t,6) = \exp\left(\widehat{\alpha} + \widehat{\gamma}_1 + \widehat{\beta}\log\widehat{\varphi}_x^{\,\mathrm{ref}}(t) + (\widehat{\delta}_1 + \widehat{\delta}_4)\; x + (\widehat{\kappa}_1 + \widehat{\kappa}_5)\; t + \widehat{\lambda}_1\; x\; t\right).$$

We observe that the final model (5.6) only includes the baseline age calendar year mixed effect, meaning that there is no significant difference of the age calendar year mixed effect between the portfolios. Portfolio P2 differs significantly from P1 only by the age pattern of the forces of mortality. The time trends are then similar to P1. Conversely, Portfolios P4 and P5 have similar age pattern but differ significantly from P1 by the time trends. P3 and P6 behave differently than P1 in age and calendar year, while the behavior of P7 is similar and only the overall level of mortality is significantly different. Similarly, the overall level of mortality is significantly different for portfolios P6, and P8. In addition, the age effect is also significant for P8.

The derivation of the portfolio specific prospective tables can sometimes lead to unrealistic estimates at the highest ages for long-term projections. Therefore, in a similar fashion as the reference table obtained in the first step, we apply Model (6) in Thatcher (1999, p9) to the forces of mortality to adapt the data at the highest ages. Figure 5.9 shows the estimated regression parameters α_t, Figure 5.9a, and β_t, Figure 5.9b.



(a) *Estimated* α_t **(b)** *Estimated* β_t

Figure 5.9: *Estimated regression parameters of model* 6 *of Thatcher (1999), male population.*

The linear component, Figure 5.9b, is much higher for portfolio P5 indicating that the forces of mortality increase more rapidly than the other portfolios at the highest ages. Conversely, portfolios P3 and P8 have smaller estimated β_t's illustrating that those portfolios lead to a less pronounced increase.

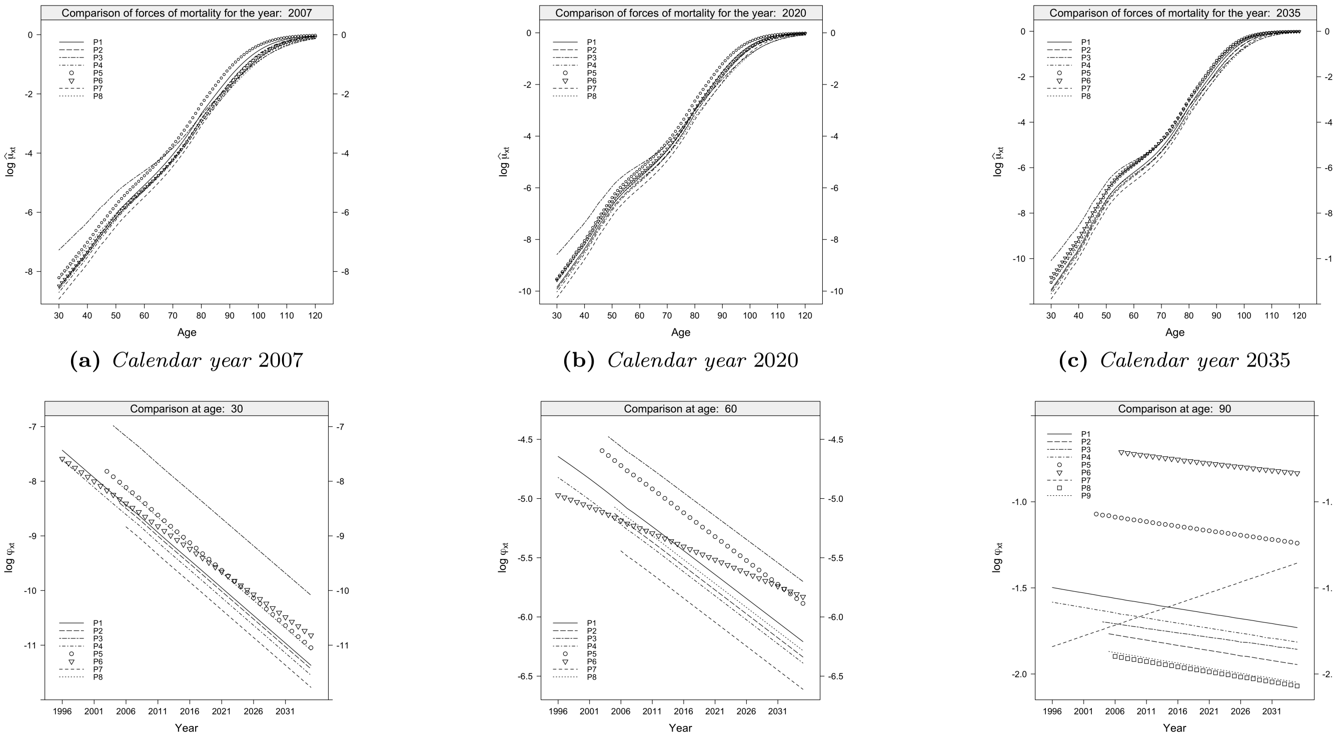
Figure 5.10 displays the forces of mortality derived for each portfolio by age and calendar years. Since we have incorporated interactions in the model, we see that the portfolio specific prospective mortality tables show different patterns with age and calendar year.
As noted in Table 5.6, P3, P5 and P1 yield the highest mortality experience, while P7 and P2 lead to the lowest mortality experience.

5.6 Summary and outlook

In this chapter, we illustrated the construction and the validation of entity specific prospective mortality tables by a two steps approach. From portfolios of several insurance companies we constructed, in a first step, a global prospective reference table summarizing the mortality experience of these portfolios. We used a non-parametric method, the local kernel-weighted log-likelihood, and semi-parametric relational models to graduate and extrapolate the surfaces. The extrapolations of the smoothed surface, obtained by local likelihood methods, were performed by identifying the mortality components and their importance over time using functional principal components analysis. Then time series methods were used to extrapolate the time-varying coefficients, while semi-parametric relational models had the advantage of integrated estimation and forecasting.

We investigated the divergences in the mortality surfaces generated by a number of proposed models. We found that the model risk is present. The overall model risk associated with a prospective mortality table was assessed by taking into account two factors, the adjustment according to the past mortality and the extrapolation of the future mortality. We have carried out a number of tests to assess the impact of model choices on the adjustment of the past mortality. We find that even for those models satisfying our criteria, there are significant differences among the smoothed forces of mortality at different ages. Moreover, selecting models purely on the basis of how well they fit historical data is dangerous, because the model may lead to a good fit to the historical data, and still give inadequate forecasts.

To measure the divergence in the extrapolation of the future mortality, we used single figure indices summarizing the lifetime probability distribution that utilize those forecasts, such as the survivor index, or partial life expectancy (which is, in turn, derived from the survivor index). We visualized those differences by a bivariate HDR boxplot of the first two robust principal



(a) *Calendar year* 2007

(b) *Calendar year* 2020

(c) *Calendar year* 2035

(d) *Attained age* 30

(e) *Attained age* 60

(f) *Attained age* 90

Figure 5.10: *Comparisons of the forces of mortality by age and calendar years, log scale, male population.*

component scores of the partial life expectancies for cohorts at several ages in 1996.

We found that the models have the following features in common: the overall level of mortality has been declining over time and these improvements have been greater at lower ages than at higher ages. However the models diverge in the level and speed of the improvement.

We therefore need to weigh the strengths and weaknesses of each model to validate the mortality table. It is up to potential users of the table to decide the weights they place on the different criteria. The validation of the mortality table involved many judgmental decisions. It has been driven by the trade-off between how the model smooths the historical data and the concept of *biological reasonableness* leading us to question the plausibility of the forecasts produced.

Then, we switched our attention to the construction of a portfolio specific prospective mortality table. The validated table is used in a second step to adjust the mortality to each portfolios by a Poisson generalized linear model including age and calendar year interactions. The estimated baseline forces of mortality are used in the regression analysis as if they were known with certainty. This approach has shown to be very simple and convenient in practical applications.

Another approach would be to use a generalized additive model (GAM) with p-splines to perform the mortality analysis in a one step approach. A GAM combines both continuous and categorical model components in one model and p-splines would have the advantage of integrated estimation and forecasting.

References

Alho, J. M. (1990). Stochastic methods in population forecasting. *International Journal of Forecasting*, **6**, 521–530.

Alistair, N. (1989). *Life contingencies.* Heinemann professional Publishing Ltd.

Benjamin, B. and Pollard, J. (1980). *The analysis of mortality and other actuarial statistics.* William Heinemann Ltd. London.

Bizley, M. T. (1958). A measure of smoothness and some remarks on a new principle of graduation. *Journal of the Institute of Actuaries*, **84**, 125–165.

Blanpain, N. and Chardon, O. (2010). Projections de populations 2007-2060 pour la France métropolitaine: méthode et principaux résultats. Série des Documents de Travail de la direction des statistiques Démographiques et Sociales F1008, Institut National de la Statistique et des Études Économiques.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, **22**(3), 547–581.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: a review of methods. *Annals of Actuarial Science*, **3**(1/2), 3–43.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting.* Springer-Verlag New-York, Inc., second edition.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002a). Measuring the longevity risk in mortality projections. *Bulletin of the Swiss Association of Actuaries*, **2**, 105–130.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002b). A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics & Economics*, **31**, 373–393.

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). Pricing death: Frameworks for the valuation and securization of mortality risk. *ASTIN Bulletin*, **36**, 79–120.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 556.

Camarda, C. G. (2008). *Smoothing methods for the analysis of mortality development*. Ph.D. thesis, Universidad Carlos III de Madrid.

Charpentier, A. (2007). Ajuster les tables de mortalité, le rôle des actuaires. *Risques*, **72**, 127–130.

Chichignoud, M. (2010). *Performances statistiques d'estimateurs non-linéaires*. Ph.D. thesis, Université de Provence - U.F.R Mathématiques.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**(368), 829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cleveland, W. S. and Loader, C. R. (1996). Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49. W. Härdle and M. G. Schimek, eds.

Cleveland, W. S., Devlin, S. J., and Grosse, E. (1988). Regression by local fitting. *Journal of Econometrics*, **37**, 87–114.

Cook, D. R. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18.

Copas, J. B. and Haberman, S. (1983). Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries*, **110**, 135–156.

Cosette, H., Delwarde, A., Denuit, M., Guillot, F., and Marceau, E. (2007). Pension plan valuation and dynamic mortality tables. *North American Actuarial Journal*, **11**(2), 1–34.

Courbage, C. and Roudaut, N. (2011). Long-term care insurance: The French example. *European Geriatric Medecine*, **2**(1), 22–25.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with *p*-splines: a unified approach. *Statistical Modelling*, **2**(333-349).

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

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimentional smoothing. *Journal of the Royal Statistical Society*, **68**(Part 2), 259–280.

Daw, R. H. (1980). Johann Heinrich Lambert (1728-1777). *Journal of the Institute of Actuaries*, **107**, 345–363.

de Boor, C. (2001). *A practical guide to splines.* New York: Springer Verlag, (revised ed.) edition.

de Laplace, P. S. (1812-1829). *Œuvres Vol. VII - Théorie Analytique des Probabilités.*

Debón, A., Montes, F., and Sala, R. (2006). A comparison of nonparametric methods in the graduation of mortality: Application to data from the Valencia region (Spain). *International statistical Review*, **74**(2), 215–233.

Deléglise, M.-P., Hess, C., and Nouet, S. (2009). Tarification, provisionnement et pilotage d'un portefeuille dépendance. *Bulletin Français d'Actuariat*, **9**(17), 70–108.

Delwarde, A. and Denuit, M. (2003). Importance de la période d'observation et des âges considérés dans la projection de la mortalité selon la méthode de Lee-Carter. *Belgian Actuarial Bulletin*, **3**(1), 1–21.

Delwarde, A., Kachkhdze, D., Olie, L., and Denuit, M. (2004). Modèles linéaires et additifs géneralisés, maximum de vraisemblance local et méthodes relationelles en assurance sur la vie. *Bulletin Français d'Actuariat*, **6**(12), 77–102.

Delwarde, A., Denuit, M., and Partrat, C. (2007). Negative binomial version of the Lee-Carter model for mortality forecasting. *Applied Stochastic Models in Buisiness and Industry*, **23**(5), 385–401.

Diewert, W. E. and Wales, T. J. (2006). A "new" approach to the smoothing problem. In *Money measurement and computation*, pages 104–144. Palgrave Macmillan.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donselaar, J., Attema, J. W., Van Broekhoven, H., Roodenburg-Berkhout, L., Willemse, W. J., and Zijp, P. (2007). On mortality and life expectancy. *Dutch Actuarial Association (Actuarieel Genootschap).*

Dupâquier, J. (1985). Leibniz et la table de mortalité. *Annales. Histoire, Sciences Sociales*, **40**(1), 136–143.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *b*-splines and penalties. *Statistical Science*, **11**(2), 89–102.

Eilers, P. H. C. and Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–783.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, **50**, 61–76.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer Series in Statistics. New York: Springer Verlag, second edition.

Fan, J. and Gijbels, I. (1995a). Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, **4**(3), 213–227.

Fan, J. and Gijbels, I. (1995b). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society*, **57**(2), 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability 66. Chapman and Hall.

Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi likelihood functions. *Journal of the American Statistical Association*, **90**(429), 141–150.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficieny. *Annals of the Institute of Statistical Mathematics*, **49**(1), 79–99.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society*, **60**(3), 591–608.

Felipe, A., Guillén, M., and Pérez-Marín, A. (2002). Recent mortality trends in the Spanish population. *British Actuarial Journal*, **8**(4), 757–786.

Forfar, D., McCutcheon, J., and Wilkie, A. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, **115**(part I(459)), 643–652.

Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society*, **47**(2), 238–252.

Gasser, T., Kneip, A., and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, **86**(415), 643–652.

Gauzère, F., Commenges, D., Barberger-Gateau, P., Letenneur, L., and Dartigues, J.-F. (1999). Maladie et dêpendance: description des évolutions par des modèles multi-états. *Population*, **54**(2), 205–222.

Gavin, J. B., Haberman, S., and Verrall, R. J. (1993). Moving weighted graduation using kernel estimation. *Insurance: Mathematics & Economics*, **12**(2), 113–126.

Gavin, J. B., Haberman, S., and Verrall, R. J. (1995). Graduation by kernel and adaptive kernel methods with a boundary correction. *Transactions of the Society of Actuaries*, **47**, 173–209.

Goldenshulger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Mathematical methods of statistics*, **6**(2), 135–170.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of london*, **115**, 513–583.

Gompertz, B. (1871). On one uniform law of mortality from birth to extreme old age, and on the law of sickness. *Journal of the Institute of Actuaries and Assurance Magazine*, **16**(5), 329–344.

Gschlössl, S., Schoenmaekers, P., and Denuit, M. (2011). Risk classification in life insurance: methodology and case study. *European Actuarial Journal*, **1**(1), 23–41.

Haberman, S. (1996). Landmarks in the history of actuarial science (up to 1919). *Actuarial Research Paper No. 84, Dept. of Actuarial Science and Statistics, City University, London.*

Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society*, **45**(4), 407–436.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference.* Cambridge University Press.

Hald, A. (1990). *A history of probability and statistics and their applications before 1750.* Probability and mathematical statistics. Wiley.

Halley, E. (1693). An estimate of the degrees of the mortality of manking, drawn from curious tables of births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. *Philosophical Transactions of the Royal Society of london*, **17**, 596–610.

Hannerz, H. (2001). An extension of relational methods in mortality estimation. *Demographic Research*, **4**, 337–368.

Härdle, W. (1990). *Applied nonparametric regression.* Econometric Society Monogrpahs. Cambridge University Press.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83**(401), 86–95.

Hastie, T. and Loader, C. R. (1993). Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, **2**, 120–143.

Heligman, L. and Pollard, J. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, **107**, 49–80.

Henderson, R. (1916). Note on graduation by adjusted average. *Transactions of the Actuarial society*, **17**, 43–48.

Henry, L. (1987). Perspectives et prévision. In *Les projections démographiques*, pages 3–11. Tome 1, Actes du VIII colloque national de démographie, Presses Universitaires de France.

Human Mortality Database (2012). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded June 2011).

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society*, **60**(2), 271–293.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, **50**(2), 120–126.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, **24**(3), 323–342.

Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, **19**(1), 29–45.

Hyndman, R. J. and Ullah, M. (2007). Robust forecasting or mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, **51**(10), 4942–4956.

Kaas, R., Goovaerts, M. J., Dhaene, J., and Denuit, M. (2008). *Modern Actuarial Risk Theory – Using R.* Berlin Heidelberg: Springer Verlag, second edition.

Katkovnik, V. (1999). A new method for varying adaptive bandwidth selection. *IEEE Transactions on signal processing*, **47**(9), 2567–2571.

Katkovnik, V., Foi, A., Egiazarian, K. O., and Astola, J. T. (2005). Anisotropic local likelihood approximations: Theory, algorithm, applications. In E. R. Dougherty, J. T. Astola, and K. O. Egiazarian, editors, *Proceedings of the SPIE - Image processing algorithms and systems IV*, volume 5672, pages 181–192.

Kessler, D. (2008). The long-term care insurance market. *The Geneva Papers on Risk and Insurance - Issues and Practice*, **33**(1), 33–40.

Kirkby, J. G. and Currie, I. D. (2010). Smooth models of mortality with period shocks. *Statistical Modelling*, **10**(2), 177–196.

Le Bras, H. (2000). *Naissance de la mortalité*. Seuil / Gallimard, collection hautes Études edition.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, **24**(4), 1602–1618.

Loader, C. R. (1999a). Bandwidth selection: classical or plug-in? *The Annals of Statistics*, **27**(2), 415–438.

Loader, C. R. (1999b). *Local Regression and Likelihood*. Statistics and Computing Series. New York: Springer Verlag.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, **15**(4), 661–675.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive smoothing with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Boca Raton: Chapman & Hall / CRC Press, second edition.

McLain, D. H. (1974). Drawing contours from arbitrary data. *Computer Journal*, **17**(4), 318–324.

Meslé, F. (2006). Progrès récents de l'espérance de vie en France, les hommes comblent une partie de leur retard. *Population*, **61**, 437–462.

Müller, H. G. (1987). Weighted local regression and kernel method for nonparametric curve fitting. *Journal of the American Statistical Association*, **82**, 231–238.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **135**, 370–384.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**(409), 66–72.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**(3), 1065–1076.

Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling longevity dynamics for pensions and annuity business.* Oxford University Press.

Planchet, F. (2006). Tables de mortalité d'expérience pour les portefeuilles de rentiers (tables TGH05 et TGF05). Technical report, Institut des Actuaires.

Planchet, F. (2012). Analyse de la survie des dépendants. Confidentiel Version 1.9, Institut de Science Financière et d'Assurances - Université Claude Bernard Lyon 1, 50 Avenue Tony Garnier - 69366 Lyon Cedex 07 - France.

Planchet, F. and Kamega, A. (2011). Construction de tables de mortalité prospectives sur un groupe restreint: Mesure du risque d'estimation. *ISFA Lab SAF Working paper*, pages 1–28.

Planchet, F. and Lelieur, V. (2007). Utilisation des méthodes de Lee-Carter et log-Poisson pour l'ajustement de tables de mortalité dans le cas de petits échantillon. *Bulletin Français d'Actuariat*, **7**(14), 118–146.

Planchet, F. and Thérond, P. (2011). *Modélisation statistique des phénomènes de durée - Applications actuarielles.* Assurance Audit Actuariat. Economica Paris.

Planchet, F. and Winter, P. (2007). L'utilisation des splines bidimensionnels pour l'estimation de lois de maintien en arrêt de travail. *Bulletin Français d'Actuariat*, **13**(7), 83–106.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, **9**(4), 705–724.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ramsay, J. O. and Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society*, **53**(3), 539–572.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis.* Springer-Verlag New-York, Inc., second edition.

Renshaw, A. E. (1991). Actuarial graduation practice and generalized linear and non-linear models. *Journal of Institute of Actuaries*, **118**, 295–312.

Renshaw, A. E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics & Economics*, **33**(2), 255–272.

Rice, J. A. (1984). Bandwidth choice for non-parametric regression. *Annals of Statistics*, **12**(4), 1215–1230.

Richards, S. J. and Currie, I. D. (2009). Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal*, **15**(2), 317–343.

Richards, S. J., Kirkby, J. G., and Currie, I. D. (2006). The importance of year of birth in two dimensional mortality data. *British Actuarial Journal*, **12**(1), 5.

Rohrbasser, J.-M. and Véron, J. (1998). Leibniz et la mortalité: mesure des "apparences" et calcul de la vie moyenne. *Population*, **53**(1-2), 29–44.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**(3), 832–837.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**(3), 1346–1370.

Schucany, W. R. (1989). On nonparametric regression with high-order kernels. *Journal of Statistical Planning and Inference*, **23**, 141–151.

Seal, H. L. (1982). Graduation by piecewise cubic polynomials: a historical review. *Blätter der Deutschen Gesellschaft für Versicherungsmathematik*, **15**, 89–114.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, **53**(3), 683–690.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, pages 517–524. Association for Computing Machinery - New York.

Sibberstein, P., Stahl, G., and Luedtke, C. (2008). Measuring model risk. *The Journal of Risk Model Validation*, **2**(4), 65–81.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**(1), 1–24.

Silverman, W. S. (1985). Some aspects of spline smoothing approaches to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, **47**, 1–52.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *Annals of Statistics*, **5**(4), 595–645.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**(6), 1348–1360.

Stone, C. J. (1982). Optimal rates of convergence for nonparametric regression. *The Annals of Statistics*, **10**(4), 1040–1053.

Taylor, G. (1992). A bayesian interpretation of Whittaker-Henderson graduation. *Insurance: Mathematics & Economics*, **11**(1), 7–16.

Thatcher, A. R. (1999). The long term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society*, **162**, 5–44.

Tibshirani, R. J. and Hastie, T. J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**(398), 559–567.

Tomas, J. (2011). A local likelihood approach to univariate graduation of mortality. *Bulletin Français d'Actuariat*, **11**(22), 105–153.

Tomas, J. (2012a). Univariate graduation of mortality by local polynomial regression. *Bulletin Français d'Actuariat*, **12**(23), 5–58.

Tomas, J. (2012b). Essays on boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models. *Insurance and Risk Management*, pages 1–31. Forthcoming.

Tomas, J. and Planchet, F. (2012). Multidimensional smoothing by adaptive local kernel-weighted log-likelihood with application to long-term care insurance. *ISFA - Laboratoire SAF Working paper - Submitted to Insurance: Mathematics & Economics*, (2012.8), 1–28.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local approximation method. *Problems of Information Transmission*, **22**(2), 133–146.

Vandeschrick, C. (2001). The Lexis diagram, a misnomer. *Demographic Research*, **4**(3), 97–124.

Véron, J. and Rohrbasser, J.-M. (2000). Lodewijk et Christiaan Huygens: La distinction entre vie moyenne et vie probable. *Mathématiques et Sciences Humaines*, **38**(149), 7–21.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, **26**(4), 359–372.

Whittaker, E. T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, 62–75.

Samenvatting (Summary in Dutch)

Overlevingstafels worden gebruikt om de eenjarige sterftekansen te beschrijven in een welomschreven populatie, in functie van de bereikte leeftijd en het kalenderjaar. Zulke kansen spelen een belangrijke rol bij het bepalen van premies en voorzieningen bij levensverzekeringen. De ruwe schattingen waarop overlevingstafels gebaseerd zijn kunnen worden gezien als een steekproef uit een grotere populatie en zijn daarom onderhevig aan stochastische fluctuaties. Meestal echter wil de actuaris deze grootheden gladmaken om de karakteristieken van de sterfte van de beschouwde groep, waarvan hij vermoedt dat deze redelijk regelmatig zijn, beter uit te lichten.

Dit proefschrift beoogt een uitputtende en gedetailleerde beschrijving te geven van verdelingsvrije afrondingsmethoden van de sterfte-ervaring in levensverzekering. De term verdelingsvrij verwijst naar de flexibele functionele vorm van de regressiecurve. Net als parametrische methoden neigen ook deze methoden naar onzuivere schattingen, maar zodanig dat het mogelijk is een grotere onzuiverheid op te laten wegen tegen een lagere steekproefvariatie. De oneffenheden van de ruwe data worden afgevlakt, alsof men een weg aanlegt over ruw terrein. Afronden is echter meer dan gladmaken. Gladgemaakte kansen moeten de onderliggende data goed weergeven, en afronding zal vaak uitdraaien op een compromis tussen de best mogelijke fit en optimale gladheid.

Regressie met lokale polynomen en lokale kernel-gewogen log-likelihood komen uitgebreid aan de orde. Belangrijke kwesties over de keuze van de parameters voor het gladmaken, statistische eigenschappen van de schatters, criteria gebruikt bij modelselectie, constructie van betrouwbaarheidsintervallen en vergelijking van de modellen worden besproken en zowel numeriek als grafisch geïllustreerd. Lokale verdelingsvrije technieken paren fraaie theoretische eigenschappen aan conceptuele eenvoud en flexibiliteit om structuur aan te brengen in vele gegevensbestanden. Geruime aandacht wordt besteed aan de invloed van de grenzen op de keuze van de bij de smoothing gebruikte parameters. Deze beschouwingen illustreren de noodzaak van flexibeler benaderingen. Adaptieve lokale kernel-gewogen log-likelihood methoden worden besproken. In hoeverre er gladgemaakt wordt verschilt van plaats tot plaats, en de methoden staan aanpassingen toe gebaseerd op de betrouwbaarheid van de data. Deze methoden passen zich netjes aan aan de complexiteit van het sterfte-oppervlak, door geschikte, op de data stoelende, keuze van de adaptieve gladstrijkparameters.

Ten slotte behandelt dit proefschrift een aantal onderwerpen die van belang zijn voor de praktijk, en wel het construeren van portefeuille-specifieke prospectieve overlevingstafels, het bepalen van het modelrisico en, zij het in mindere mate, het risico van het oordeel van experts bij de keuze van de externe data.

Summary

Life tables are used to describe the one-year probability of death within a well defined population as a function of attained age and calendar year. These probabilities play an important role in the determination of premium rates and reserves in life insurance. The crude estimates on which life tables are based might be considered as a sample from a larger population and are, as a result, subject to random fluctuations. Most of the time, however, the actuary wishes to smooth these quantities to enlighten the characteristics of the mortality of the group considered which he thinks to be relatively regular.

This dissertation aims at providing a comprehensive and detailed description of non-parametric graduation methods of experience data originating from life insurance. The term non-parametric refers to the flexible functional form of the regression curve. Like parametric methods, they too are liable to give biased estimates, but in such a way that it is possible to balance an increase in bias with a decrease in sampling variation.
In the actuarial literature, the process of smoothing a mortality table is known as graduating the data. The little hills and valleys of the rough data are to be graded into smoothness, just as in building a road over rough terrain. Smoothing alone, however, is not graduation. Graduated rates must be representative of the underlying data and graduation will often turn out to be a compromise between optimal fit and optimal smoothness.

Local polynomials regression and local kernel-weighted log-likelihood are discussed extensively. Important issues concerning the choice of the smoothing parameters, statistical properties of the estimators, criteria used for models selection, construction of confidence intervals and comparisons between the models are covered with numerical and graphical illustrations.
Local non-parametric techniques combine excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. Considerable attention is devoted to the influence of the boundaries on the choice of the smoothing parameters. These considerations illustrate the need for more flexible approaches. Adaptive local kernel-weighted log-likelihood methods are introduced. The amount of smoothing varies in a location dependent manner and the methods allow adjustments based on

the reliability of the data. These methodologies adapt neatly to the complexity of mortality surface, clearly because of the appropriate data-driven choice of the adaptive smoothing parameters.

Finally, this manuscript deals with some important topics for practitioners. Those concern the construction and validation of portfolio specific prospective mortality tables, assessment of the model risk and, to a lesser extent, the risk of expert judgment related to the choice of the external data used.

Résumé (Summary in French)

Les tables de mortalité sont utilisées pour décrire la probabilité annuelle de décès d'une population en fonction de l'âge atteint et de l'année calendaire. Ces probabilités jouent un rôle important dans la détermination des primes et réserves en assurance vie. Les estimations brutes, sur lesquelles se basent les tables de mortalité, peuvent être considérées comme un échantillon provenant d'une population plus importante et sont, par conséquent, soumises à des fluctuations aléatoires. Toutefois, l'actuaire souhaite la plupart du temps lisser ces quantités afin de faire ressortir les caractéristiques de la mortalité du groupe considéré qu'il pense être relativement régulières.

Cette dissertation fournit une description détaillée des méthodes de graduation non-paramétrique de données d'expérience issues de l'assurance vie. Le terme non-paramétrique renvoie à une forme fonctionnelle de la courbe de régression. Comme les méthodes paramétriques, elles sont toutes aussi susceptibles de donner des estimations biaisées, mais de telle sorte qu'il est possible de compenser une augmentation du biais avec une diminution de la variation de l'échantillonnage.
Dans la littérature actuarielle, le processus de lisser une table de mortalité est appelé graduation. Les collines et vallées des données brutes sont lissées de façon similaire á la construction d'une route sur un terrain accidenté. Le lissage seul, cependant, n'est pas la graduation. Les taux gradués doivent être représentatifs des données sous-jacentes et la graduation se révélera souvent comme un compromis entre ajustement et lissage optimal.

Les régressions polynomiales locales et méthodes de vraisemblance locale sont examinées en détail. Les questions importantes concernant le choix des paramètres de lissage, les propriétés statistiques des estimateurs, les critères utilisés pour la sélection des modèles, la construction des intervalles de confiance ainsi que les comparaisons entre les modèles sont couvertes avec des illustrations numériques et graphiques. Les techniques non-paramétriques locales combinent d'excellentes propriétés théoriques avec une simplicité et une flexibilité conceptuelle pour trouver une structure dans de nombreuses bases de données. Une attention particulère est consacrée à l'influence des bordures sur le choix des paramètres de lissage. Ces considérations illustrent le besoin d'avoir à disposition des approches plus flexibles. Des méthodes

adaptatives de vraisemblance locale sont alors introduites. Le montant de lissage varie en fonction de l'emplacement et ces approches permettent des ajustements de la fenêtre d'observation en fonction de la fiabilité des données. Ces méthodes s'adaptent parfaitement à la complexité de la surface de mortalité en raison du choix adaptatif approprié des paramètres de lissage. Enfin, ce manuscrit traite de sujets importants pour les praticiens. Ceux-ci concernent la construction et la validation de tables de mortalité prospectives pour des portefeuilles d'assurance, l'évaluation du risque de modèle, et dans une moindre mesure, du risque d'opinion d'experts lié au choix de la table de référence externe utilisée.