UNIVARIATE GRADUATION OF MORTALITY BY LOCAL POLYNOMIAL REGRESSION

Julien TOMAS[1]

Amsterdam School of Economics Research Institute[2]
ISFA - Laboratoire SAF[3]

Abstract:

Life tables are used to describe the one-year probability of death within a well defined population as a function of attained age. 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 larger population and are, as a result, subject to random fluctuations. However, the actuary wishes most of the time to smooth these quantities to enlighten the characteristics of the mortality of the group considered which he thinks to be relatively regular.

The present article discusses a non-parametric graduation method of experience data originating from life insurance. We introduce local univariate polynomial regression. We discuss the choices of the smoothing parameters and criteria used for models selection. We graduate the mortality 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 by local polynomial regression with those obtained by the Whittaker-Henderson smoothing.

Résumé :

Les tables de mortalité sont utilisées pour décrire la probabilité annuelle de décès d'une population en fonction de l'âge atteint. 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

[1] Corresponding author: j.tomas@uva.nl. All data and computer programs are available on request.
[2] Roetersstraat 11, 1018 WB, Amsterdam
[3] Université de Lyon - Université Claude Bernard Lyon 1 - 50 avenue Tony Garnier - 69366 Lyon Cedex 07 - France

BULLETIN FRANÇAIS D'ACTUARIAT, Vol. 12, n° 23, janvier – juin 2012, pp. 5-58

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.

Cet article discute d'une méthode de graduation non-paramétrique de données d'expérience issues de l'assurance vie. Nous introduisons la régression polynomiale locale univariée. Nous discutons du choix des paramètres de lissage et des critères utilisés pour la sélection des modèles. Nous graduons les taux de mortalité à travers le choix des paramètres de lissage. Le lissage et les intervalles de confiance correspondants sont obtenus pour l'ensemble des tranches d'âge. Des tests sont utilisés pour comparer les taux lisses obtenus par la régression polynomiale locale avec ceux obtenus selon le modèle de Whittaker-Henderson.

Keywords: Local Polynomials, Life insurance, Graduation, Whittaker-Henderson.

JEL - Code: C14, G22.

1. INTRODUCTION

1.1 Life tables and graduation: The stochastic formalization of life time

The age at which a person will decease 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.

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 as rate of mortality. The denominator of the rate (of which the numerator is the relevant number of deaths) is commonly referred to as *population at risk* or the *exposed to risk*. To be specific, let assume that we are given the number of deaths recorded, d_i, and the number of individuals initially exposed to the risk of death, l_i, all aged x_i last birthday, and that our experience is limited to this single age x_i where $i = 1, 2, \ldots, n$. The crude estimate of the observed mortality rate, q_i, is denoted by $\overset{\circ}{q}_i$,

$$\overset{\circ}{q}_i = \frac{d_i}{l_i}.$$

Then $\overset{\circ}{q}_i$ represents the one-year observed probability of death for a particular population at age x_i which lies above or below the true underlying value. 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 less reliable will be the resulting $\overset{\circ}{q}_i$. 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_i of persons observed. If the number of individuals in the group had been considerably larger, the set of observed probabilities $\overset{\circ}{q}_i$, would have displayed a much more regular progression with x_i. In the limit, it would have exhibited a smooth progression explain Copas and Haberman (1983, p.136).

The idea of a group of persons attaining age x_i and being gradually reduced in numbers, until they are all dead, by the operation of mortality 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 recalls Alistair (1989), from 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 resume Bloomfield and Haberman (1987).

Figure 1 displays the one-year transformed crude probabilities of death (year 2008), logit scale (see Section 2.2), for ages $x_i = 0,1,\ldots,98$ and each gender for the dutch population provided by the Human Mortality Database (2011). The Human Mortality Database (HMD) has been 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 which can be accessed online for research purposes.

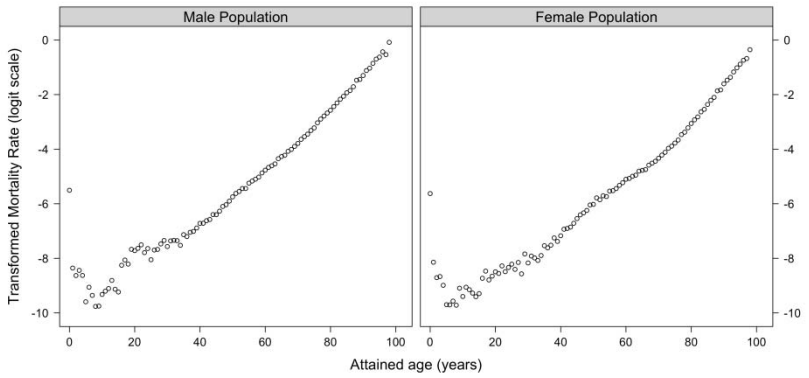

Figure 1: Transformed crude one-year probabilities of death $\overset{\circ}{q}_i^t$*, logit scale, for Dutch Male (left panel) and Dutch females (right panel) in 2008. Source: HMD.*

From Figure 1, 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. 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 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 by Benjamin and Pollard (1980) in medical terms as follows:

- 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 lung) 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 and provides an example of the need for cause analysis.

Thus, graduation aims 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:

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

1.2 Getting out of a Procrustean bed

Assume n data points $\{(x_i, \overset{\circ}{q}_i)\}_{i=1}^{n}$ have been collected, then the regression relationship can be modeled as

$$\overset{\circ}{q}_i = f(x_i) + u_i, \quad i = 1, 2, \ldots, n;$$

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 to 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 prespecified 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 assumed that the functional form (i.e. Thiele law, Perks laws, Gompertz-Makeham class of models, etc...) 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 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 such the approaches taken by Heligman and Pollard (1980) applied initially for graduating post-war Australian national mortality. The Heligman-Pollard model is an eight parameters model containing three terms, each representing a distinct component of mortality. The first, a rapidely declining exponential, reflects the fall in mortality during the early childhood. The third term in the formula is the well known Gompertz exponential. It reflects the near geometric rise in mortality at the adult ages. It is generally considered to represent the ageing or deterioration of the body, i.e, the senescent mortality. The remaining term, is a function similar to the log-normal. It reflects accident mortality for males and accident plus maternal mortality for the female population. Although the law may not always give a fit close enough for actuarial purposes, it does reproduce the three distinct features of mortality. The model is applicable over the entire age range. It has relatively few parameters, all of which have demographic interpretation and together fully describe the age pattern of mortality. See Keyfitz (1981) for an extensive review of the choice of functions for mortality analysis.

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 recalls Hardle (1990). Fisher pointed out that the non-parametric approach gave generally poor efficiency whereas Pearson was more concerned about the specification question. Both point 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 model which may result in a more variable estimates, especially for small sample size.

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

We have previously seen that the crude rates, $\overset{\circ}{q}_i$, on which the model is based, 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, q_i, 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 $\overset{\circ}{q}_j$ near $\overset{\circ}{q}_i$ 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 towards 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, i.e., the little hills and valleys of the rough 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. It is a very difficult concept to define mathematically and we

deliberately avoid here a detailed presentation. 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 those points which are such that 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 require that third-order differences are small, which is consistent with the widely held view that low-order polynomials are smooth.

1.4 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, a graduation 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 oversmooth 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 oversmoothing and undersmoothing is called the smoothing parameters selection problem. To give insight into the smoothing parameters selection problem consider figure 2 below.

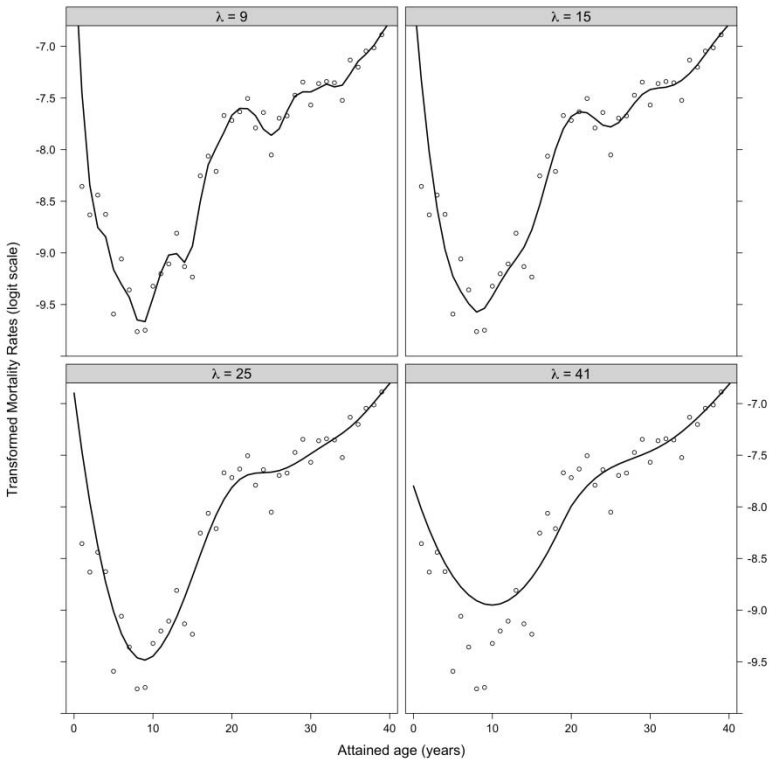

Figure 2: Estimated curve and transformed crude mortality rates (dots), logit scale, for Dutch Male 2008. Source: HMD.

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 4.

1.5 Content of the paper

This article begins by presenting, in Section 2, a general theory of local univariate polynomial regression, showing this method falls into the class of linear smoothers. Then Section 3 develops important properties, including bias and variance, which allow us in Section 4 to develop methods for statistical inference, model diagnostics and choices of the smoothing parameters. We emphasize on results that have immediate practical

consequences. To illustrate the discussion, we present two examples in Section 5. Section 6 provides comparisons with the Whittaker-Henderson framework. Finally, Section 7 summarizes the conclusions drawn in the paper.

2. LOCAL POLYNOMIAL REGRESSION

2.1 Premises

The underlying model for local regression is

$$\overset{\circ}{q}_i = f(x_i) + u_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

The distribution of the $\overset{\circ}{q}_i$, including the means, $f(x_i)$, are unknown. However, the u_i are assumed to be independently, identically distributed normal random variables, with zero mean and a constant, 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 $\overset{\circ}{q}_i$. For example, one common distributional assumption is that the $\overset{\circ}{q}_i$ have a constant variance, we need to ensure that these assumptions are reflected in the data and, if not, to make appropriate adjustments, see the following Section 2.2.

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.

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 greater weight to the x_j in the neighborhood that are close to x_i and lesser 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 $\overset{\circ}{q}_i$. In other words, as with parametric fitting, we are modeling the data.

2.2 Transforming mortality data

Before model (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,

$$\overset{\circ t}{q}_i = \overset{\circ}{q}_i + r_i \quad , \quad \text{for } i = 1, 2, \ldots, n;$$

where t denotes the transformation and the residuals r_i are assumed to be independently, 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 (1) becomes

$$\overset{\circ t}{q}_i = f(x_i) + \varepsilon_i \quad , \quad \text{for } i = 1, 2, \ldots, n; \tag{2}$$

where ε_i are independently, 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, t, in binary analysis is the logit transformation. For our application,

$$\overset{\circ t}{q}_i = \log\left(\frac{\overset{\circ}{q}_i}{1-\overset{\circ}{q}_i}\right),$$

with back-transform

$$\hat{q}_i = \frac{\exp\left(\sum_{j\in N_{j|\lambda}} s_j(x_i)\overset{\circ t}{q}_j\right)}{1+\exp\left(\sum_{j\in N_{j|\lambda}} s_j(x_i)\overset{\circ t}{q}_j\right)}, \quad \text{for } i = 1,2,\ldots,n.$$

By smoothing on a logistic scale and then back-transforming, we are guaranteed that the predicted values stay in an appropriate scale, $0 \le \hat{q}_i \le 1$. 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 paper. Overall, the choice of transformation remains subjective, and the relative success of a particular transformation seems to depend on the data set.

However transformations do not always achieve normality, neither lead to skewness zero or homoscedasticity. 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 note the dependent variable $\overset{\circ t}{q}_i$ by y_i to lighten the notation.

2.3 Theory

We assume a model of the form of 2,

$$y_i = f(x_i) + \varepsilon_i \quad , \quad \text{for } i = 1,\ldots,n;$$

where $f(x_i)$ 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, $\mathbb{E}[\varepsilon_i] = 0$, and have finite variance, $\mathbb{E}[\varepsilon_i^2] = \sigma_i^2 < \infty$.

We now turn to non-parametric estimation of f. Globally, no strong assumptions are made about f. Locally around a point x_i, we assume that f can be well approximated by a member of a simple class of parametric functions.

Assume that the function f has $(p+1)$ th continuous derivative at the point x_i.

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

$$\begin{aligned} f(x_j) &\approx \sum_{p=0}^{P} (f^{(p)}(x_i)/p!)(x_j - x_i)^p \qquad (3) \\ &= f(x_i) + f'(x_i)(x_j - x_i) + f''(x_i)(x_j - x_i)^2 \frac{1}{2} + \ldots + f^{(p)}(x_i)(x_j - x_i)^p \frac{1}{p!} \\ &= \sum_{p=0}^{P} \beta_p(x_i)(x_j - x_i)^p. \end{aligned}$$

We then carry through a weighted polynomial regression:

$$\sum_{j=1}^{n} \left(y_j - \sum_{p=0}^{P} \beta_{i,p}(x_j - x_i)^p \right)^2 W\left(\frac{x_j - x_i}{h} \right), \qquad (4)$$

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. A weight function $W(u)$ should require:

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

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

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

iv. $W(u) = 0$ for $|u| \geq 1$.

$W(u)$ is some weight function like those given in Table 1, below. 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.

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

Table 1: Example of weight functions with $u = | x_j - x_i | / h$.

Figure 3 displays some of the weight functions presented above. For a weight function $W(u)$, the weights decrease with increasing distance $\left| x_j - x_i \right|$. 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. Such a weight function produces smoothed points that have a smooth appearance and it is widely appreciated in the literature that a smooth weight function results in a smoother estimate, see Cleveland and Loader (1996, p. 10-11). 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.

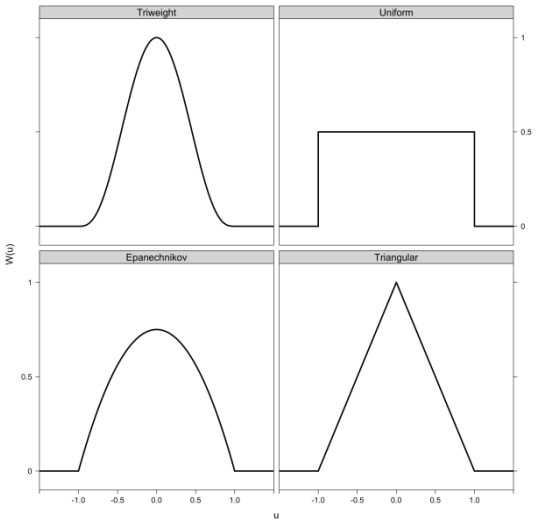

Figure 3: Weighting system shape of some weight functions.

If $\{\widehat{\beta}_p(x_i)\}$ denotes the solution to the above weighted least squares problem (4), then it is clear from approximation (3) that $p!\widehat{\beta}_p(x_i)$ estimates $f^{(p)}(x_i)$, $p = 0,1,\ldots,P$. The weighted sum of square can be written in matrix form as

$$(\mathbf{y} - \mathbf{Xb})^T \mathbf{W}(\mathbf{y} - \mathbf{Xb}),$$

with

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

and $\mathbf{W}$ is 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 $\mathbf{WX}$ has full column rank, least squares theory gives the explicit expression for the minimizer

$$\hat{\mathbf{b}}(x_i) = (\mathbf{X}^T \mathbf{WX})^{-1} \mathbf{X}^T \mathbf{W}\, \mathbf{y}, \tag{5}$$

and $\mathbf{b} = (\beta_0, \beta_1, \ldots, \beta_P)$. Hence,

$$\widehat{\beta_0}(x_i) = \widehat{f}(x_i) = \mathbf{e}_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}\, \mathbf{y}. \quad (6)$$

Here and throughout, we let $\mathbf{e}_v$ denote a column vector of length $P+1$ having 1 as its v th 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 $f(x_i)$, one has to minimize (4) for a grid of target values x_i. For each target value one gets specific parameter estimates $\mathbf{b}(x_i)$.

Also, the form of the estimate is simple in that it is linear in y_i. Because local polynomial regressions, solve a least squares problem, $\widehat{f}(x_i)$ 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

$$\widehat{f}(x_i) = \sum_{j=1}^{n} s_j(x_i) y_j, \quad (7)$$

where the smoothing weights on the observed responses are given by

$$s_j(x_i) = w_j \sum_{p=0}^{P} \beta_p (x_j - x_i)^p. \quad (8)$$

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 $\mathbf{X}^T \mathbf{W} \mathbf{X}$ is non-singular.

As we can see in (8) the smoother weights $s_j(x_i)$ depends on λ and $\mathbf{X}$ in a highly non-linear way. The only linearity we have in equation (7) is linearity in $\mathbf{y}$. This linear representation (7) provides a basis for the theoretical development of local regression estimation. Likewise in a matrix form,

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

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

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

with rows

$$\mathbf{s}(x_i)^T = \left(s_1(x_i), s_2(x_i), \dots, s_n(x_i)\right) = \mathbf{e}_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}. \tag{9}$$

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.

3. STATISTICAL PROPERTIES

3.1 Bias, variance, influence and degree of freedom

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 $k(\hat{\mathbf{b}})$. On the other hand, we can approximate the non linear function using a first-order Taylor series expansion about $\mathbf{b}$. Assuming the first order differentiability of $k(.)$, we have

$$k(\hat{\mathbf{b}}) = k(\mathbf{b}) + \frac{\partial k(\mathbf{b})}{\partial \mathbf{b}^T}\left(\hat{\mathbf{b}} - \mathbf{b}\right) + o\left\|\hat{\mathbf{b}} - \mathbf{b}\right\|.$$

Then for $\hat{f}(x_i) = \hat{\beta}_0(x_i)$, we obtain

$$\hat{f}(x_i) = f(x_i) + \frac{\partial k(\mathbf{b})}{\partial \beta_0^T}\left(\hat{\mathbf{b}} - \mathbf{b}\right) + o\left\|\hat{\mathbf{b}} - \mathbf{b}\right\|,$$

and,

$$\mathbb{E}\left[\hat{f}(x_i)\right] = f(x_i) + \frac{\partial k(\mathbf{b})}{\partial \beta_0^T}\mathbb{E}\left[\hat{\mathbf{b}} - \mathbf{b}\right].$$

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

$$\begin{aligned} \mathbb{V}ar\left[\hat{f}(x_i)\right] &\approx \mathbb{E}\left[\left(\hat{f}(x_i) - f(x_i)\right)^2\right] \\ &\approx \mathbb{E}\left[\left(\frac{\partial k(\mathbf{b})}{\partial \beta_0^T}\left(\hat{\mathbf{b}} - \mathbf{b}\right)\right)^2\right] \end{aligned} \tag{10}$$

$$= \frac{\partial k(\mathbf{b})}{\partial \beta_0^T} \mathbb{E}\left[\left(\hat{\mathbf{b}}(x_i) - \mathbf{b}(x_i) \right)^2 \right] \frac{\partial k(\mathbf{b})}{\partial \beta_0}. \quad (11)$$

We still need to estimate $\mathbb{V}ar\left[\hat{\mathbf{b}}(x_i)\right] = \mathbb{E}\left[\left(\hat{\mathbf{b}}(x_i) - \mathbf{b}(x_i) \right)^2 \right]$. However, standard weighted least squares theory provides explicit mean and variance expressions of the solution (5),

$$\mathbb{E}\left[\hat{\mathbf{b}}(x_i)\right] = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{f} = \mathbf{b} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{e}, \quad (12)$$

where $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))^T$ and $\mathbf{e} = \{\varepsilon_j\}_{j=1}^n = \mathbf{f} - \mathbf{X}\mathbf{b}$; and,

$$\begin{aligned} \mathbb{V}ar\left[\hat{\mathbf{b}}(x_i)\right] &= \mathbb{E}\left[\left(\hat{\mathbf{b}}(x_i) - \mathbf{b}(x_i) \right)^2 \right] = \mathbb{E}\left[\left(\hat{\mathbf{b}}(x_i) - \mathbf{b}(x_i) \right)\left(\hat{\mathbf{b}}(x_i) - \mathbf{b}(x_i) \right)^T \right] \\ &= \mathbb{E}\left[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{e}\mathbf{e}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \right] \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbb{E}\left[\mathbf{e}\mathbf{e}^T\right]\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}. \end{aligned} \quad (13)$$

From (2), $\mathbb{E}[\mathbf{e}\mathbf{e}^T] = \sigma^2(x_j)\mathbf{I}_n$. Using local homoscedasticity, namely that $\sigma(x_j) \approx \sigma(x_i)$ for x_j in a neighborhood of x_i, then equation (13) can be approximated by

$$\mathbb{V}ar\left[\hat{\mathbf{b}}(x_i)\right] = \sigma^2(x_i)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}. \quad (14)$$

Therefore,

$$\begin{aligned} \mathbb{V}ar\left[\hat{f}(x_i)\right] &= \sigma^2(x_i)\mathbf{e}_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{e}_1 \\ &= \sigma^2(x_i)\mathbf{S}\mathbf{S}^T, \end{aligned} \quad (15)$$

since $\partial k(\mathbf{b}) / \partial \beta_0^T = \mathbf{e}_1^T$.

Then by (9) we obtain compact forms for the mean and variance of the local regression estimate, similar to Loader (1999b, p. 288)

$$\begin{aligned} \mathbb{E}\left[\hat{f}(x_i)\right] &= \sum_{j=1}^n s_j(x_i) f(x_j) \\ \mathbb{V}ar\left[\hat{f}(x_i)\right] &= \sigma^2(x_i)\sum_{j=1}^n s_j^2(x_i) = \sigma^2(x_i) \| \mathbf{s}(x_i) \|^2. \end{aligned} \quad (16)$$

The variance reducing factor $\| \mathbf{s}(x_i) \|^2$ measures the reduction in variance due to the local regression. It usually decreases with the bandwidth. The above distributional results are the same as those for parametric least-squares except that for least-squares $\mathbf{S}$ is replaced by the so called *hat matrix*, the projection operator onto the space spanned by the fitting variables. $\mathbf{S}$ shares with the hat matrix the property that if $\mathbf{z}$ is a vector in this space then $\mathbf{S}\mathbf{z} = \mathbf{z}$. In other words, the smooth weight diagram is constant preserving, the rows of $\mathbf{S}$ 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 argue 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.

As in linear models, a quantity of interest is the influence function, closely related to the variance. The influence values, $\text{infl}(x_i)$, are the diagonal elements $s_i(x_i)$ of the smooth weight diagram,

$$\text{infl}(x_i) = \mathbf{e}_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{e}_1.$$

These measure the sensitivity of the fitted curve $\hat{f}(x_i)$ to the individual data points.

Although the notion of degrees of freedom does not really apply to smoothers, the usefulness of the degrees of freedom is in providing a measure of the amount of smoothing that is comparable between different estimates applied to the same dataset. Among the several possible definitions, we denote

$$\begin{aligned} \upsilon_1 &= \sum_{i=1}^{n} \text{infl}(x_i) = \text{tr}(\mathbf{S}) \\ \upsilon_2 &= \sum_{i=1}^{n} \left\| \mathbf{s}(x_i) \right\|^2 = \text{tr}(\mathbf{S}\mathbf{S}^T). \end{aligned} \tag{17}$$

υ_2 is the equivalent degrees of freedom of $\hat{f}(x_i)$. 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. More extensive discussion of the degrees of freedom of a smoother can be found in Cleveland and Devlin (1988).

3.2 Assessment of bias and variance and construction of pointwise confidence intervals

The bias and variance in equations (12) and (13) are not directly accessible, as they depend on unknown quantities, the residual $\mathbf{e}$ 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 Fan and Gijbels (1995a, p. 218-219) and Fan and Gijbels (1995b, p. 376-378).

The bias of the estimator $\hat{\mathbf{b}}$ comes from the approximation error in the Taylor

expansion. Recall the bias vector given in (12) and let

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

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

$$\varepsilon(x_j) \approx \beta_{p+1}(x_j - x_i)^{p+1} + \ldots + \beta_{p+a}(x_j - x_i)^{p+a} \equiv \tau_j, \tag{18}$$

where $\beta_k = f^{(k)}(x_i)/k!$ 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 $\hat{\beta}^*_{p+1}, \ldots, \hat{\beta}^*_{p+a}$ be the resulting estimated regression coefficients and denote the weighted residual sum of squares by

$$\hat{\sigma}^{*2}(x_i) = \frac{1}{\mathrm{tr}(\mathbf{W}^*) - \mathrm{tr}((\mathbf{X}^{*T}\mathbf{W}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{W}^{*2}\mathbf{X}^*)} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 W\left(\frac{x_j - x_i}{h^*}\right), \tag{19}$$

where $\hat{y}_j$ are the fitted values from the $(p+a)$ th order local polynomial fit. Moreover, $\mathbf{X}^*$ and $\mathbf{W}^*$, similar to $\mathbf{X}$ and $\mathbf{W}$, 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 $\hat{\tau}$, leading to an estimated bias vector

$$\begin{aligned} \widehat{\mathrm{bias}}_p(x_i) &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\tau} \\ &= (\mathbf{T})^{-1} \begin{pmatrix} \hat{\beta}^*_{p+1} t_{p+1} & \cdots & \hat{\beta}^*_{p+a} t_{p+a} \\ & \vdots & \\ \hat{\beta}^*_{p+1} t_{2p+1} & \cdots & \hat{\beta}^*_{p+a} t_{2p+a} \end{pmatrix}, \end{aligned} \tag{21}$$

where $\mathbf{T} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ is a $(p+1)\times(p+1)$ matrix whose $(j,k)th$ element is t_{j+k-2} with

$$t_k = \sum_{j=1}^{n} (x_j - x_i)^k W\left(\frac{x_j - x_i}{h}\right). \tag{22}$$

The variance matrix of the estimator (13) can be estimated by substituting $\hat{\sigma}^{*2}(x_i)$, defined in (19), into (14). This provides an estimated variance matrix

$$\widehat{\text{var}}_p(x_i) = \hat{\sigma}^{*2}(x_i)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}. \tag{23}$$

Expressions (21) and (23) give the estimated bias and variance not only for $\hat{f}(x_i)$ but also for $\hat{f}^{(v)}(x_i) = v!\beta_v(x_i), v = 0,1,\ldots,P$.

The estimated bias for $\hat{f}^{(v)}(x_i)$ is the $(v+1)$th element of (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 (23), denoted by $\widehat{\text{var}}_{p,v}(x_i)$, times $(v!)^2$. For instance,

$$\mathbb{E}\left[\hat{f}(x_i)\right] - f(x_i) = \mathbf{e}_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\tau}, \tag{24}$$

and,

$$\begin{aligned}\mathbb{V}ar\left[\hat{f}(x_i)\right] &= \hat{\sigma}^{*2}(x_i)\mathbf{e}_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{e}_1 \\ &= \hat{\sigma}^{*2}(x_i)\|\mathbf{s}(x_i)\|^2 .\end{aligned} \tag{25}$$

Recall that the approximated bias (12) and variance (13) depends respectively on the quantities $\varepsilon_1,\ldots,\varepsilon_n$ and $\sigma^2(x_i)$, which are unknown. These quantities will be estimated by fitting a local polynomial of degree $p+a$ locally via equation (4), using a pilot bandwidth h^*. This gives estimates $\hat{\beta}_0^*,\hat{\beta}_1^*,\ldots,\hat{\beta}_{p+a}^*$ and $\hat{\sigma}^{*2}(x_i)$, which are then substituted respectively into expressions (18), yielding estimates $\hat{\tau}_1,\hat{\tau}_2,\ldots,\hat{\tau}_n$ of $\tau_1,\tau_2,\ldots,\tau_n$, and (23) leading to the estimated variance. Finally, the estimated bias is computed by substituting the estimates $\hat{\tau}_1,\hat{\tau}_2,\ldots,\hat{\tau}_n$ into (20).

As recommended by Fan and Gijbels (1995b, p. 377) we modify the bias estimate in expression (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 $t_{p+a+1},t_{p+a+2},\ldots,t_{2p+a}$ in (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 asymptotics properties, since it only concerns the higher order terms and no leading terms.

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

By (25) a local polynomial estimate $\hat{f}(x_i)$ has the distribution

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

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

$$\hat{I}(x_i) = \left(\widehat{f}(x_i) - \hat{b}(x_i) - c\,\widehat{\sigma}^*(x_i)\,\|\mathbf{s}(x_i)\|,\ \widehat{f}(x_i) - \hat{b}(x_i) + c\,\widehat{\sigma}^*(x_i)\,\|\mathbf{s}(x_i)\|\right), \qquad (26)$$

where c is the appropriate quantile of the standard normal distribution ($c = 1.96$ for 95% confidence) and $\hat{b}(x_i)$ is a bias estimate as defined in (24).

Remark that 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 $\hat{I}(x_i)$ is just a construction of an undersmoothed interval centered around the local quadratic estimate $\widehat{f}(x_i) - \hat{b}(x_i)$.

Other authors have expressed the bias and the variance in other fashions, see Section 4.2 or Cleveland et al. (1988, p. 100) however we do not provide here any comparisons between the approaches.

3.3 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 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 (24) and variance (25) espressions are suitable for our applications. However, they only give 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 Muller (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.

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

$$f(x_j) = f(x_i) + (x_j - x_i) f'(x_i) + \ldots + (x_j - x_i)^p \frac{f^{(p)}(x_i)}{p!}$$
$$+ (x_j - x_i)^{p+1} \frac{f^{(p+1)}(x_i)}{(p+1)!} + (x_j - x_i)^{p+2} \frac{f^{(p+2)}(x_i)}{(p+2)!} + \ldots$$

As an application of Henderson's theorem, we know that the 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,

$$\mathbb{E}\left[\widehat{f}(x_i) \right] - f(x_i) = \frac{f^{(p+1)}(x_i)}{(p+1)!} \sum_{j=1}^{n} s_j (x_i) (x_j - x_i)^{p+1}$$
$$+ \frac{f^{(p+2)}(x_i)}{(p+2)!} \sum_{j=1}^{n} s_j (x_i) (x_j - x_i)^{p+2} + \ldots \qquad (27)$$

The bias has a leading term involving the $(p+1)$ st derivative $\widehat{f}^{(p+1)}(x_i)$. We keep the $\widehat{f}^{(p+2)}(x_i)$ term in (27) because in our 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, similarily to Muller (1987, p. 234 Corollary 3) for kernel regression. Thus the first term in the bias expansion disappears. In this case the second term is dominant.

From expression (22), the matrix $\mathbf{X}^T \mathbf{W} \mathbf{X}$ has components t_k of the form $\sum_{i=1}^{n} w_j (x_j - x_i)^k$. Under mild conditions, in particular $nh \to \infty$,

$$\frac{1}{nh} \sum_{j=1}^{n} w_j^l \frac{(x_j - x_i)^k}{h^k} = \int W(v)^l v^k f(x_i + hv) dv + o(1), \qquad (28)$$

This result is valid for fixed h. Under the additional assumption $h \to 0$, (28) simplifies to

$$\frac{1}{nh} \sum_{j=1}^{n} w_j^l \frac{(x_j - x_i)^k}{h^k} = f(x_i) \int W(v)^l v^k dv + o(1). \qquad (29)$$

For regular design, the limit (29) follows from the theory of Riemann sums, see Loader (1999b, p. 38-39). Applying (28) and (29) to the matrix $\mathbf{X}^T \mathbf{W}^l \mathbf{X}$ gives

$$\frac{1}{nh} \mathbf{H}^{-1} \mathbf{X}^T \mathbf{W}^l \mathbf{X} \mathbf{H}^{-1}$$
$$= \begin{cases} \int W(v)^l \mathbf{c}(v) \mathbf{c}(v)^T f(x_i + hv) dv + o(1) & \text{hfixed} \\ f(x_i) \int W(v)^l \mathbf{c}(v) \mathbf{c}(v)^T dv + o(1) & \text{h} \to 0, \end{cases} \qquad (30)$$

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

Asymptotic approximations to quantities such as the bias and variance are now easily derived. Under the small bandwidth limits, the variance (25) has the following asymptotic approximation

$$\mathbb{V}ar\left[\widehat{f}(x_i)\right] = \frac{\sigma^{*2}(x_i)}{nh\, f(x_i)} \mathbf{e}_1^T \Lambda_1^{-1} \Lambda_2 \Lambda_1^{-1} \mathbf{e}_1 + o((nh)^{-1}), \tag{31}$$

where $\Lambda_l^{-1} = \int W(v)^l \mathbf{c}(v)\mathbf{c}(v)^T dv$.

Substituting (30) into expression (6) for the local regression estimate leads

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

where

$$W^{\circ}(v) = \mathbf{e}_1^T \Lambda_1^{-1} \mathbf{c}(v) W(v). \tag{32}$$

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 (31) becomes

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

The first term of the bias expansion (27) is approximated by

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

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 (27), which has size $o(h^{p+2})$. For p even, we obtain

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

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 4 shows the squared bias, variance and *MSE* into 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.

The intuition 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 to small, the variance term tends to be larger. So the *MSE* quantity protects against both extreme choices.

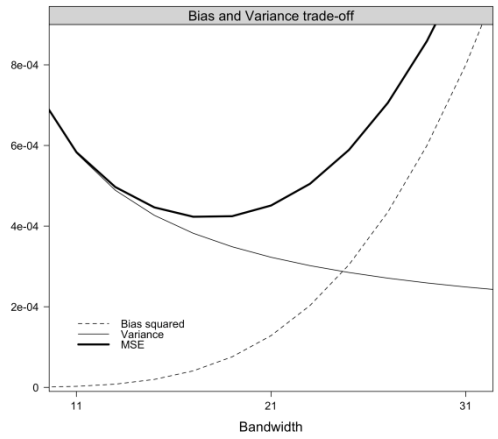

Figure 4: Squared bias (thin dashed), variance (thin solid) and mse (thick solid) of a local polynomial fit for the Dutch male population, 2008. Source: HMD.

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, for p odd the bias does not depend on the density $g(x_i)$; in this sense the estimate is *design adaptive* to use the terminology of Fahrmeir and Tutz (2001). For p even, the term contains the density $g(x_i)$ in the denominator, meaning that bias is lower if the density $g(x_i)$ is high.

To give an illustration on how the trade-off between bias and variance works in practice, consider Figures 5 and 6.

Figure 5 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 6 shows the residuals for each of the 20 fits in Figure 4, but for the all age range, from 0 to 98. Superposed on each plot is a loess smooth.

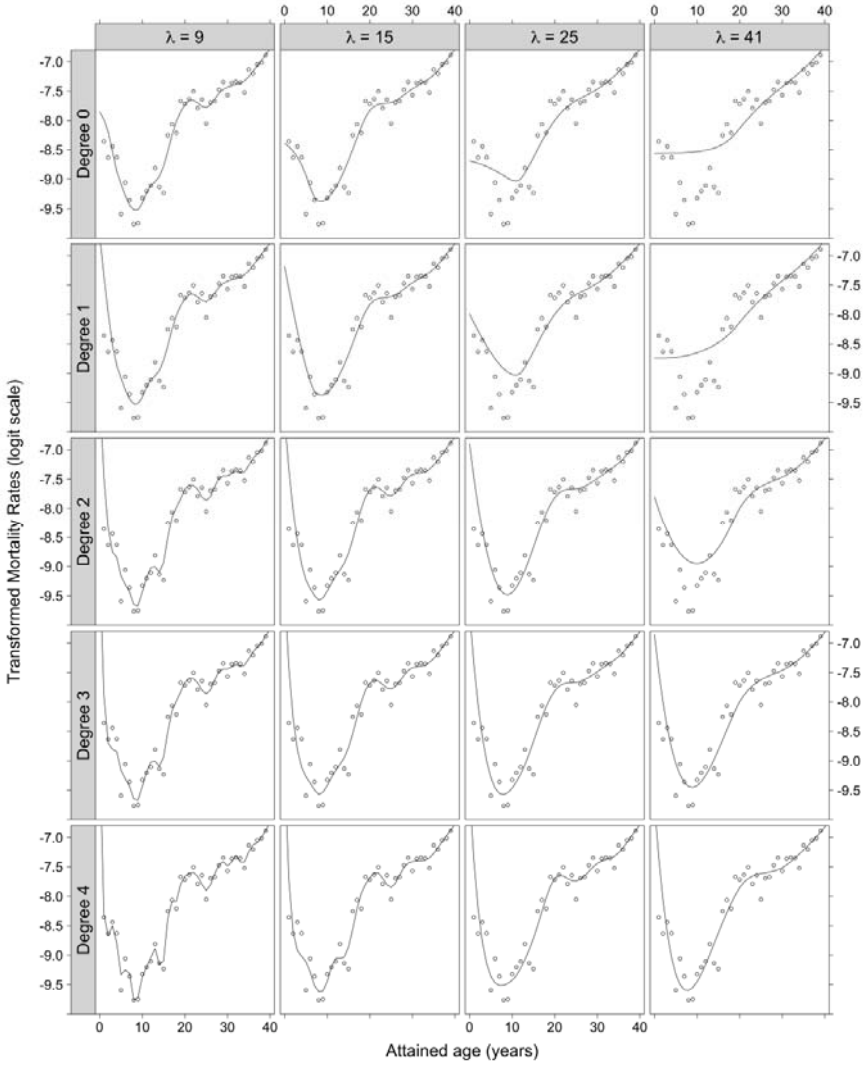

Figure 5: Fits for four bandwidths and five local fitting methods for the Dutch male population, 2008. Source: HMD.

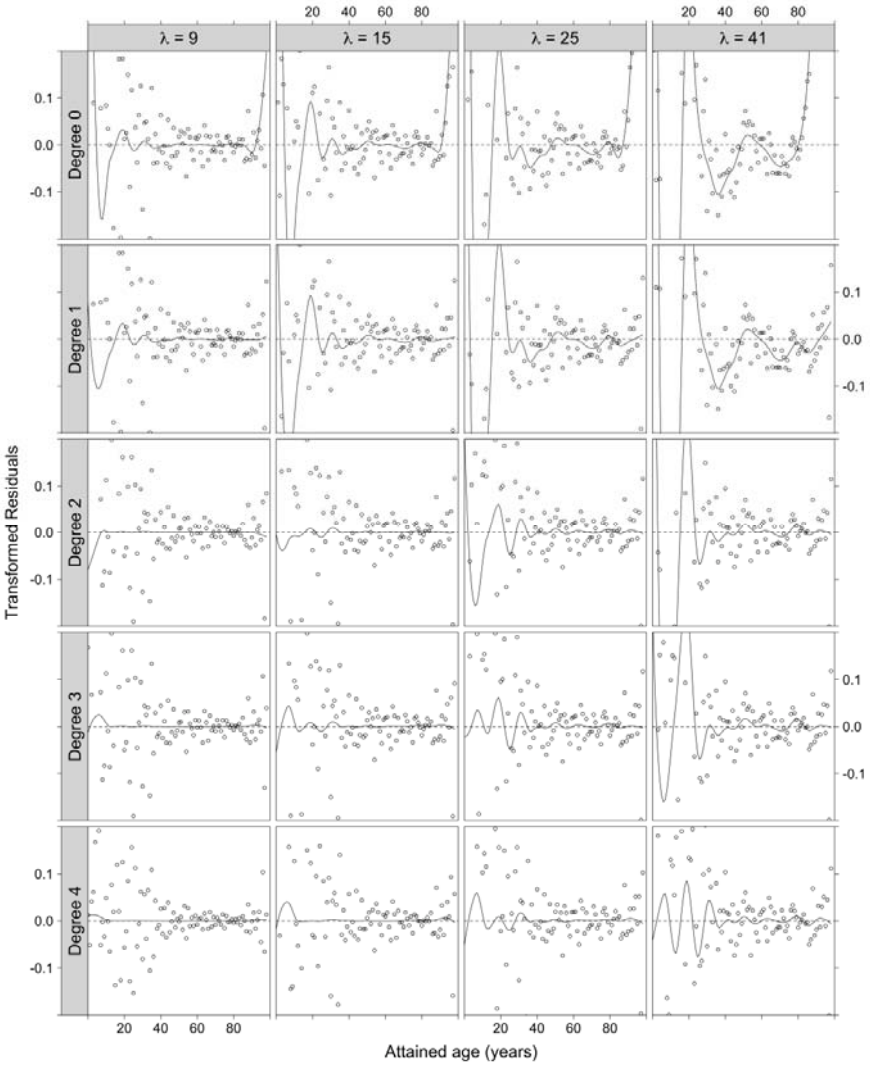

Figure 6: Transformed Residuals plots for the fits in the left panel for the Dutch male population, 2008. Source: HMD.

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 done 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 misses completely 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 6.

To some extents, 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, 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 do we 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.

4. 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 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 argue Cleveland and Loader (1996).

From parametric regression, we shall think about two families of criteria respectively based on prediction error and on estimation error.

4.1 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 $\hat{y}_0 = \hat{f}(x_0)$, what is the prediction error?

One measure is

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

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

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{-i}(x_i)\right)^2. \tag{33}$$

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

The leave-one-out cross validation criteria was introduced for parametric models as the PRESS procedure (prediction error sum of squares). Formally computing each of the leave-one-out regression estimates $\hat{f}^{-i}(.)$ would be highly computational, and so at a first sight computation of the CV as in (33) looks prohibitively expensive. But there is a remarkable simplification, valid for all common linear smoother, by 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 $\mathbf{S}$.

Actually, it is not clear what 'leave-one-out' means in the context of smoothing. In general there is no necessary relationship between a smoother for n data pairs and a smoother for $n-1$ data pairs. One method of finding such 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 to be an $(n-1)\times(n-1)$ smooth weight diagram, we must renormalize the rows to sum to one.

Let recall that $s_i(x_i)$ denote the diagonal elements of the original $n \times n$ smooth weight diagram $\mathbf{S}$. 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 $\hat{f}(x_i) = \sum_j s_j(x_i)\, y_j$, one may choose

$$\hat{f}^{-i}(x_i) = \frac{1}{1 - s_i(x_i)} \sum_{\substack{j=1 \\ j \neq i}}^{n} s_i(x_j) y_j, \tag{34}$$

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

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

Then the essential term $y_i - \hat{f}^{-i}(x_i)$ in (33) is given by

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

and may be computed from the regular fit $\hat{f}(x_i)$ based on n observations and weights $s_i(x_i)$. By using (34) one gets the criterion

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

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

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

is easier to compute as it is the single average squared error corrected by a factor. The generalized cross validation can be seen as a special case of minimizing

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

where $\varphi(.)$ is a penalty function that decreases with increasing smoothness of $\widehat{f}$ and $\widehat{\sigma}^2 = (1/n)\sum_i (y_i - \widehat{f}(x_i))^2$ is the average squared residuals, see Hurvich et al. (1998, p. 273).

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

$$\log(\widehat{\sigma}^2) + 2\,\mathrm{tr}(\mathbf{S})/n. \tag{35}$$

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 response $y_i \sim N(\mu_i, \sigma^2)$, one obtains apart from additive constants $AIC = n\left(\log(\widehat{\sigma}^2) + \frac{2}{n}p\right)$. In (35) the trace $\mathrm{tr}(\mathbf{S})$ plays the role of the effective numbers of parameters used in the smoothing fit, see Loader (1999b). Thus, replacing p by $\mathrm{tr}(\mathbf{S})$ yields to (35). If $\varphi(\mathbf{S}) = -\log\{1 - 2\,\mathrm{tr}(\mathbf{S})/n\}$ is chosen, one obtains the criterion suggested by Rice (1984).

A last alternative can be mentioned. Hurvich et al. (1998, p. 88) proposed to use the criterion *AICC*, a corrected version of the *AIC*,

$$AICC = \log(\hat{\sigma}^2) + \frac{1 + tr(\mathbf{S})/n}{1 - (tr(\mathbf{S}) + 2)/n} = \log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(\mathbf{S}) + 1)}{n - \mathrm{tr}(\mathbf{S}) - 2}. \tag{36}$$

The first term in (36) measure the quality of the adjustment while the second term evaluate the model complexity.

It follows from Hardle 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 $\mathrm{tr}(\mathbf{S}) \to 0$, a situation that is not consistent with a small smoothing parameter λ.

Figure 7 makes this distinction clear. It gives the penalty functions $\varphi(\mathbf{S})$ as a function of $\mathrm{tr}(\mathbf{S})$ 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 $n = 100$).

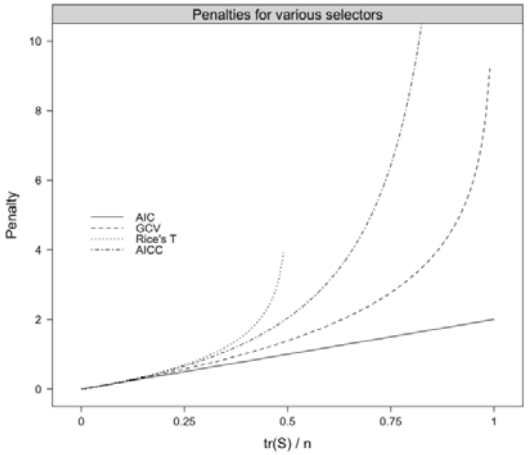

Figure 7: $\varphi(.)$ *penalties for various selectors as a function of* $\mathrm{tr}(\mathbf{S})/n$.

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

In consequence, we would use *AIC* and *GCV* selector when the data present a relatively smooth pattern as we are more likely to look for an undersmooth fit. While The Rice's T statistic and *AICC* would be used reciprocally, as they lead to an oversmooth fit which is satisfactory when the data are relatively volatile.

4.2 Criteria based on estimation error

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

$$R(f,\widehat{f}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\widehat{f}(x_i) - f(x_i)\right)^2\right]. \tag{37}$$

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

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

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

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

Likewise in matrix notation, knowing that

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

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

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

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

$$\mathbb{E}\left[\left\|\mathbf{y}-\widehat{\mathbf{f}}\right\|^2\right] - \sigma^2(n-2\upsilon_1+\upsilon_2). \tag{38}$$

With (25) and (38), and making the propers arrangements, an unbiased estimate of (37) is

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

This statistic is known as the *Cp* criterion, and has been introduced by Mallows (1973) for parametric regressions. It provides an unbiased estimate of $R(f,\widehat{f})$. This statistic was extended to local regression by Cleveland and Devlin (1988). To implement the *Cp* criterion (or unbiased risk estimate) one needs to known an estimate σ^2. However in practice, we do not know 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, $\widehat{\sigma}^2_{\lambda}$, where λ is small by

$$\hat{\sigma}^2 = \frac{1}{n - 2\upsilon_1 + \upsilon_2} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2 .$$

4.3 Plug-in methods 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 4.1 and 4.2. 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 $\hat{f}$ is written as a function of the unknown f, and is approximated through Taylor series expansions. A pilot estimate of f 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.

$$\widehat{MSE}_{p,v}(x_i, h) = (v!)^2 \left(\widehat{\text{bias}}^2_{p,v}(x_i) + \widehat{\text{var}}_{p,v}(x_i) \right). \tag{39}$$

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,

$$IRSC(h) = \int RSC(t, h) dt,$$

with the RSC defined as

$$RSC(x_i, h) = \hat{\sigma}^{*2}(x_i) \left(1 + (p+1) / N \right), \tag{40}$$

where N^{-1} is the first diagonal element of the matrix $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^2 \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ and $\hat{\sigma}^{*2}(x_i)$ is the normalized weighted residual sum of squares after fitting locally a $(p+a)$ th order polynomial defined as expression (19). Note that N reflects the effective number of local data points since $\mathbb{V}ar\left[\hat{\mathbf{b}}(x_i) \right] = \sigma^2(x_i) / N$ by equation (14).

The intuition behind statistic (40) 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 $\hat{\sigma}^{*2}(x_i)$. When the bandwidth is to 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 $\hat{\beta}^*_{p+1}(x_i)$, $\hat{\beta}^*_{p+2}(x_i)$ and $\hat{\sigma}^{*2}(x_i)$. 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 $\hat{\beta}_v$, which are respectively the $(v+1)$ th element of vector (20) and the $(v+1)$ th diagonal element of the estimated expression (23). Combining these estimates yield to the estimated *MSE* (39). This leads to the bandwidth selector

$$\hat{h}_{p,v} = \arg\min_h \left\{ \int \widehat{MSE}_{p,v}(t,h)dt \right\}.$$

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 (22), are not further replace by their asymptotics counterparts. The quantities t_k are already known, and Fan and Gijbels (1995b) 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 $f_X(x_i)$.

However, for higher order fit ($p \geq 2$) 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 indicate 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.

4.4 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 4.1 and 4.2 for different fits and select the fit with the lowest score.

However, as argue 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 that the statistics provided by the whole profile of the selectors curves. Cleveland and Devlin introduced then graphical tools for displaying these statistics.

As an illustration, Figure 8, below, displays the *AIC* scores against the fitted degrees of freedom $\text{tr}(\mathbf{SS}^T)$. 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.

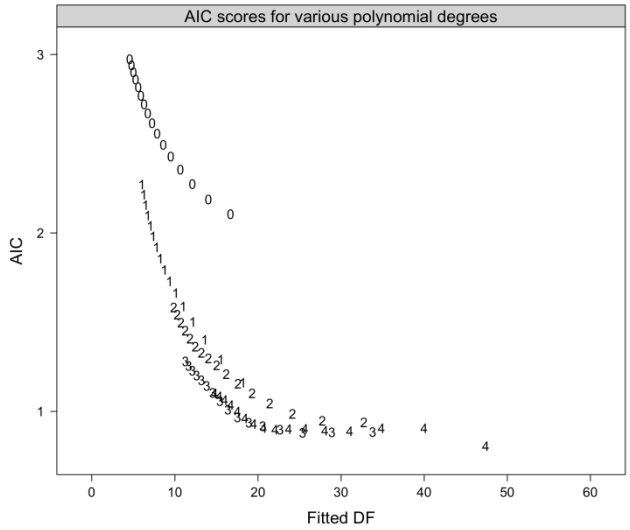

Figure 8: AIC scores for various polynomial degrees and triweight weight function for Dutch Male population, 2008. Source: HMD.

From Figure 8, 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 resultant 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 9 shows the fits and corresponding residuals plot for the constellation picked by the lowest *AIC* score and the one elected by aid of 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. Superposed on the residual plot is a loess smooth with local quadratic fitting and

$\lambda = 19$. The smooths help for search for clusters of residuals that may indicate lack of fit. Such residual plots provide an exceedingly 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 dataset, the fit is very nearly interpolating the data which results in unacceptably high variance.

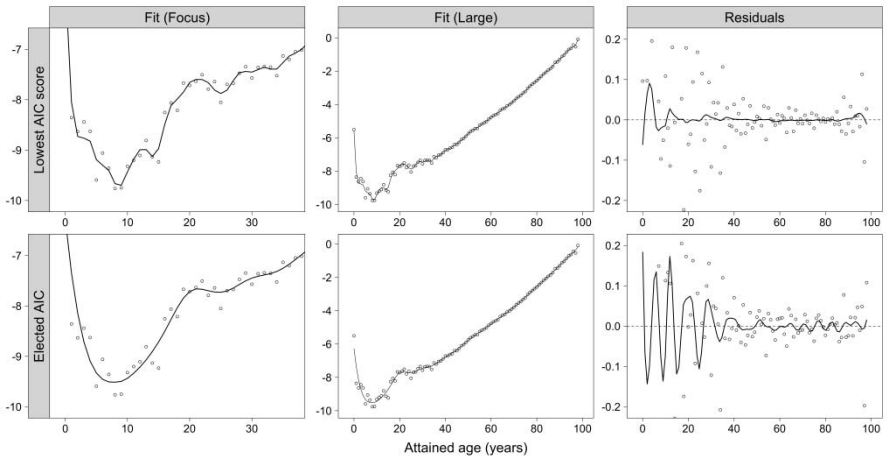

Figure 9: Fits and residuals plots elected by the AIC score with a triweight weight function for Dutch Male population, 2008. Source HMD.

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, 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 tunning 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 exclusive reliance in practice 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 the design space.

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.

5. LOCAL POLYNOMIALS METHOD FOR GRADUATION

5.1 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 (2011). The scripts are available on request. Figure 10 displays the observed statistics of the two datasets.

- The data for the first application are reported by the Human Mortality Database (2011). The dependent variable is the measurements in a logit scale of the one-year probability of death for the Dutch Male population for the year 2008 at age x_i ; x_i being the independent variable.
- The data for the second application are the Female counterpart.

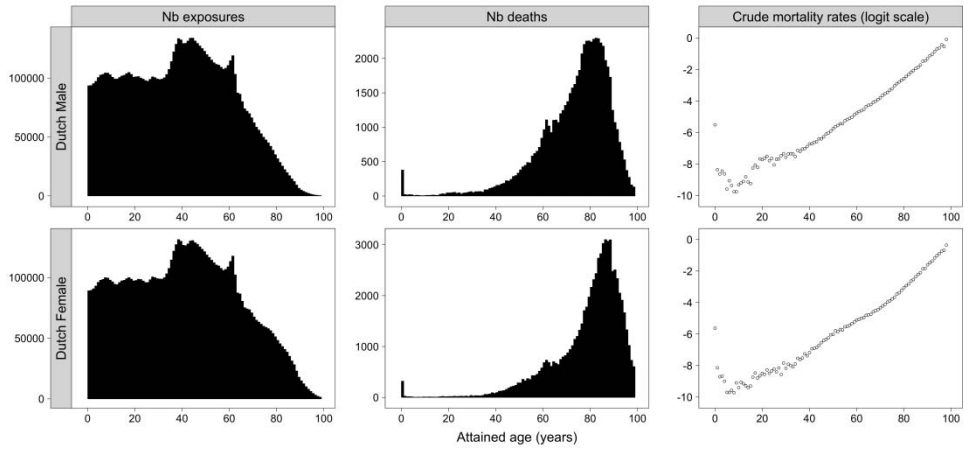

Figure 10: Observed statistics for Dutch Male and Female population, 2008. Source: HMD.

5.2 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 constellation, we use the criteria presented in Section 4 and graphical diagnostics to guide our modeling.

Both datasets present a relatively wiggly pattern. For these applications we picked the optimal constellation selected by the Rice's T statistic and $AICC$ as the final fit. Due to strong penalties on $tr(S)/n$, these criteria have tendencies to lead to oversmoothing, 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 displays the elected optimal constellation of smoothing parameters for the local polynomials method together with the fitted degrees of freedom. Recall $\lambda = (2*h)+1$.

	λ	Degree	$W(.)$	Fitted DF
Dutch Male	19	3	Triweight	18,46
Dutch Female	21	3	Triweight	16,76

Table 2: Elected optimal constellation of smoothing parameters and fitted degrees of freedom

A local cubic fit is needed to capture the mortality patterns. The choice differ by the elected bandwidth. The weight function has much less effect on the bias-variance trade-off than the two others 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. An 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 3 presents the theoretical optimal bandwidth provided by the plug-in method developed in Section 4.3. We fit a polynomial of degree 3 and use the corresponding optimal weight functions elected in Table 2. The values of λ are reported below.

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

Table 3: Pilot and optimal bandwidths selected by the plug-in method

The optimal bandwidths confirmed our choices presented in Table 11, being relatively close and agreeing with our ranking.

5.3 Plots of the fits on the transformed scale

Figure 11 presents the mortality rates (logit scale) graduated by our local polynomials method with the optimal constellation of smoothing parameters displayed in Table 2.

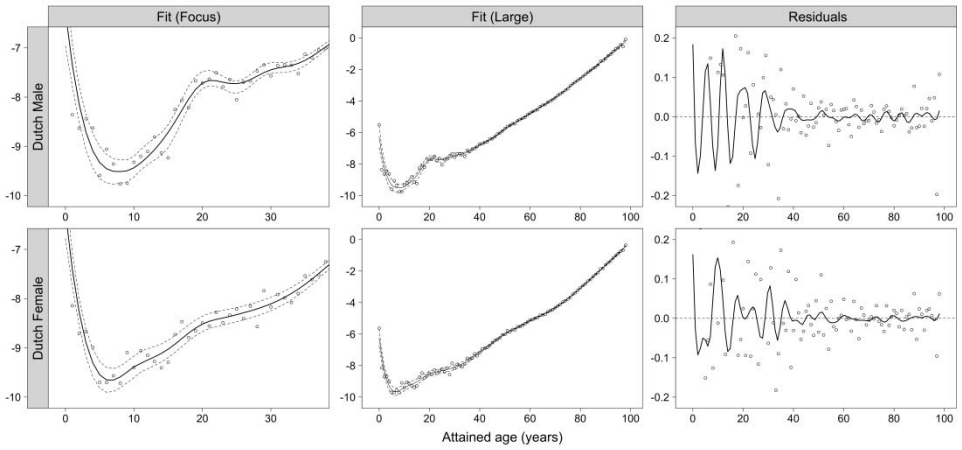

Figure 11: Graduated mortality rates by local polynomial (logit scale) with 95% pointwise confidence intervals and corresponding transformed residuals plots for Dutch Male and Female population, 2008. Source: HMD.

In conjunction of the plots of the fits, we display the residuals plots. Superposed 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 relatively linear trend after 30 years old; it is normal to get higher residuals at the left boundary than to the rest of the curve. It shows us where the observed mortality rates differ from what we think relatively regular.

A last feature is shown by examining the confidence intervals in Figure 11. 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 death 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.

5.4 Plots of the smoothers

The weight function associated with the i-th point is used to compute the weights in the i-th row, $\mathbf{s}(x_i)$, of the 99 x 99 smoother $\mathbf{S}$ and is shown in Figures 12 and 13, below, with the influence values.

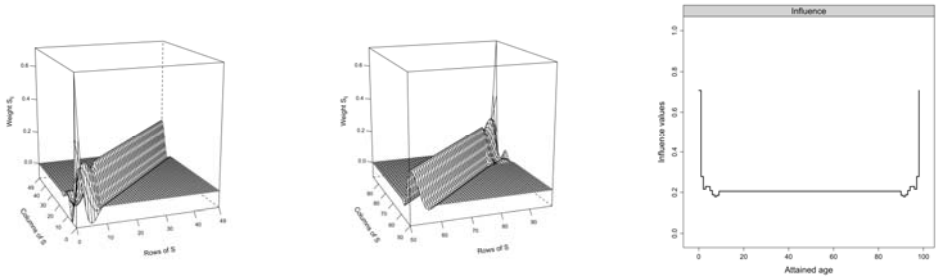

Figure 12: Smoother S_{ij} : left panel: $i, j = 0,\ldots,49$, center panel: $i, j = 50,\ldots,98$ and influence values for the Dutch Male population, 2008. Source: HMD.

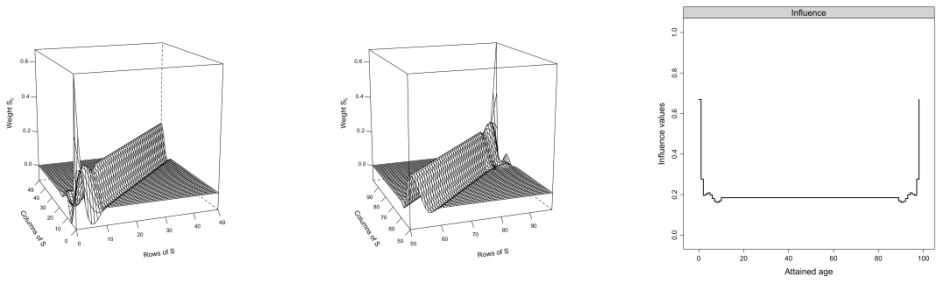

Figure 13: Smoother S_{ij} : left panel: $i, j = 0,\ldots,49$, right panel: $i, j = 50,\ldots,98$ and influence values for the Dutch Female population, 2008. Source: HMD.

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, becomes asymmetric and some weights are negative. Moreover, the height of the kernel increases because fewer observations are available.

We deliberately avoid here a presentation on the boundary correcting kernel. The interested reader is invited to look at Tomas (2011b) for a detailed presentation. However, 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 (2011). 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 argue Fan et

al. (1998). At the boundaries, the residuals sum of squares, *RSC* criterion and estimated derivatives can be too large because of numerical instabilities and scarcity of the data, see Tomas (2011b).

The influence values measure the sensitivity of the fitted curves $\widehat{f}(x_i)$ to the individual data points. It show us the amount of smoothing applied locally. For instance, in Figure 12 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 ($\approx 0,21$) 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 to the rest of the curve.

5.5 Plots of the graduated series and diagnostic checks

Having produced estimates on the transformed scale, we now back-transform the graduated rates. Figure 14 presents the mortality rates graduated on the original scale by our local polynomials method.

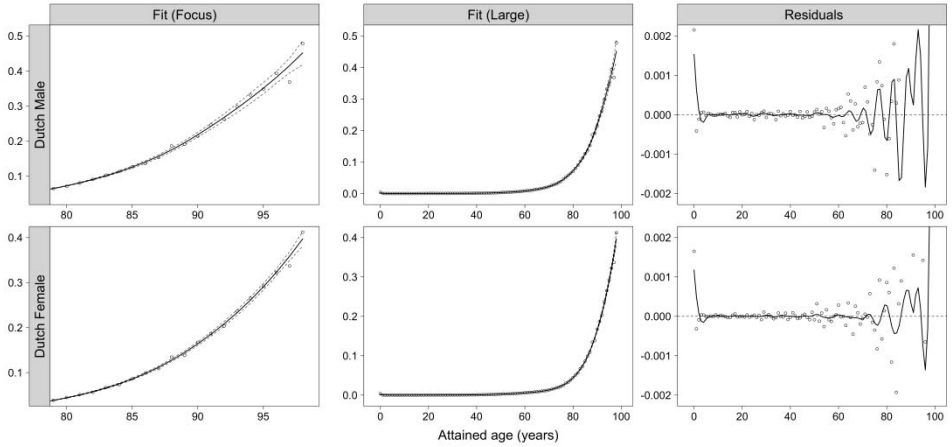

Figure 14: Graduated mortality rates by local polynomials (original scale) with 95% pointwise confidence intervals and corresponding residuals plots for Dutch Male and Female population, 2008. Source: HMD.

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

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

Figure 15 displays the expected number of death with the statistic described above. We notice that most values are less than two and the statistic has a mean close to zero for both population, indicating that the assumptions made by the model are valid. Several other diagnostic plots and non-parametric test could be considered, see Gavin et al. (1995, p.183) and Cleveland et al. (1988).

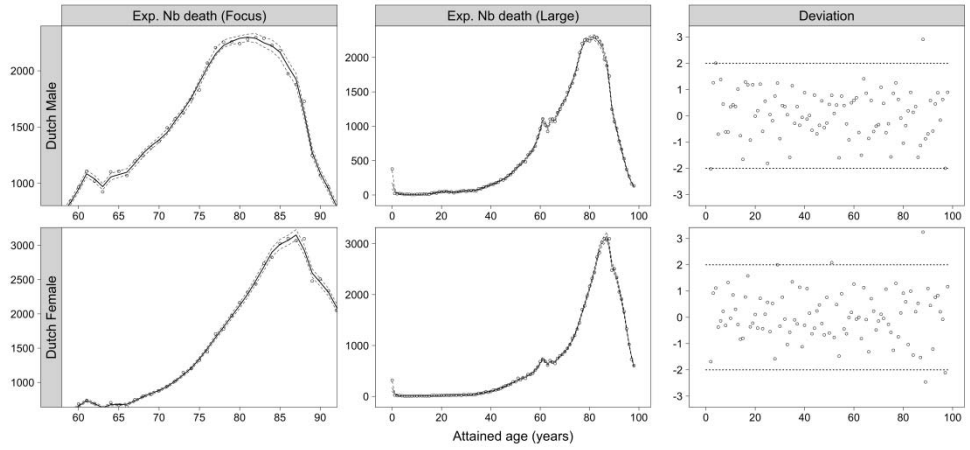

Figure 15: Expected number of death with 95% pointwise confidence intervals and deviation between actual and expected death for Dutch Male and Female population, 2008. Source: HMD.

6. WHITTAKER-HENDERSON SMOOTHING

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 approach 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, and concluding that the general spline function is preferable to Whittaker-Henderson graduation due to his greater flexibility.

In this section we choose to use the Whittaker-Henderson model which is simpler to implement.

6.1 The Whittaker-Henderson model as a linear smoother

We show that the Whittaker-Henderson model falls into the class of linear smoothers. it will allow us to use the methodology develops in Section 4 for model diagnostic and choice of parameters.

The Whittaker-Henderson model is non-parametric and forms 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,

$$M = F + h * S,$$

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

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

where v_i represents the weight for observation i, taken generally as the ratio $l_i / \max(l_i)$; and z being an other 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 such that,

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

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

$$F = (\mathbf{y} - \hat{\mathbf{y}})^T \mathbf{V} (\mathbf{y} - \hat{\mathbf{y}}).$$

For the smoothness criterion, writing $\Delta^z \mathbf{y} = (\Delta^z y_i)_{1 \le i \le n-z}$, yields to $S = (\Delta^z \mathbf{y})^T \Delta^z \mathbf{y}$.

To find $\Delta^z \mathbf{y}$, 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 sign of the coefficients alternates and starts positively for z even, $\Delta^z \mathbf{y} = K_z * \mathbf{y}$.

The M criterion can finally be written as

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

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

$$\hat{\mathbf{y}} = (\mathbf{V} + hK_z^T K_z)^{-1}\mathbf{V}\mathbf{y}. \tag{41}$$

We see that the form of the estimate is simple in that it is linear in the y_i. In consequence, similarly to the local polynomials method, we can apply the criteria presented in Section 4.1 to find the optimal value of parameters h and z.

6.2 Comparisons

We picked the constellation, $h = 5$ and $z = 3$ for the male, and $h = 20$ and $z = 3$ for the female population, given by the Rice's T criterion, leading to $20,99$ and $17,06$ fitted degrees of freedom respectively.

Figures 16 and 17 present graphical comparisons of the local polynomials approach and the Whittaker-Henderson model.

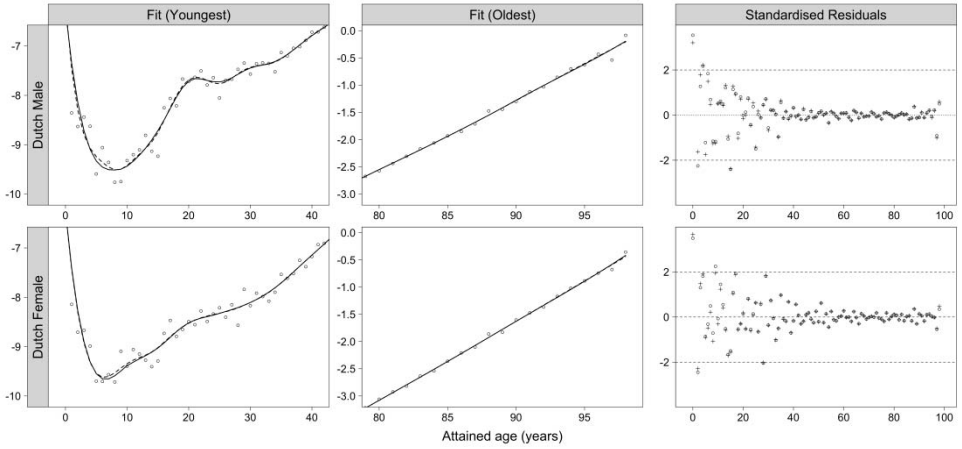

Figure 16: Graphical comparisons between the local polynomials approach (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008: Graduated series and standardised residuals. Source: HMD.

In Figure 16, the top left panel presents the graduated mortality rates (logit scale) for the Dutch Male population. The graduated series 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 relatively identical. The fitted degrees of freedom are very close, illustrating that the models show the same amount of features.

The right panels display the standardised residuals. The circles represent the residuals from the local polynomials approach and the crosses the ones from the Whittaker-Henderson smoothing. The standardised residuals are mainly in the interval $[-2;2]$ which indicates that the models adequately model the variability of these datasets.

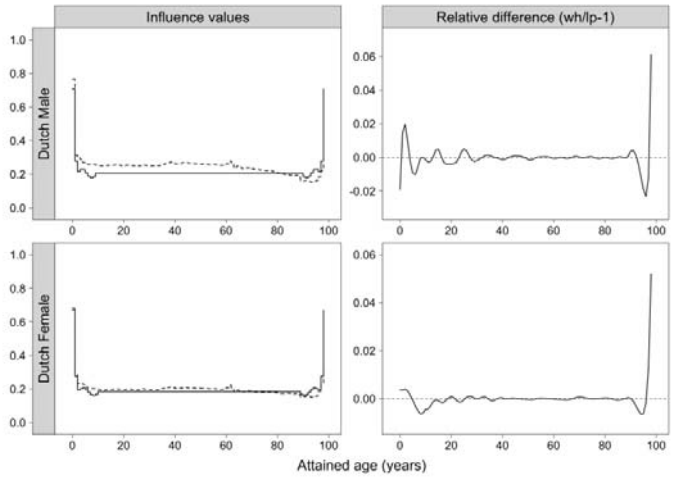

Figure 17: Graphical comparisons between the local polynomials approach (full line) and the Whittaker-Henderson smoothing (dotted line) for the Dutch Male and Female population, 2008: Influence values and relative difference between the graduated series. Source: HMD.

In Figure 17, the influence values, obtained by the local polynomials for the male population are, up to age 80, below the ones computed with Whittaker-Henderson model, $\text{infl}_{WH}(x_i) = diag((\mathbf{V} + hK_z^T K_z)^{-1}\mathbf{V})$, top 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 constitute 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 constitute 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 Debon 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 4 presents the results.

		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
Standardised	> 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
χ^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 4: Comparisons between the local polynomials approach and the Whittaker-Henderson smoothing for the Dutch Male and Female population, 2008. Source: HMD.

The two approaches display favorable results making it difficult to choose one of them. As an advantage for the Whittaker-Henderson, we observe that is not necessary to give a special treatment for the observations in the boundary and the computation time is 100 times smaller. However we have used a prototype implementation in R to perform the local polynomials approach. This can be improved on by at least a factor of 10, if a lower level language such as C is used.

7. DISCUSSION AND RELATED WORK

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. In consequence, the selection of the smoothing parameters may be mainly driven by minimizing the residuals sum of squares in the boundaries rather than for the whole data points. It may force the criteria to select a smaller bandwidth at the boundary to reduce the bias, but this may lead to undersmoothing in the middle of the table.

One 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 fittings 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 decided to include the young age groups to show the applicability and relevancy of the approach to find structure in presence of irregular profile. Moreover, as pointed by a referee, it is worth remembering that the double exponential which appears in Heligman and Pollard (1980) and 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 Tomas (2011a), the concepts of Sections 2 and 3 are incorporated and extended within the framework of local likelihood and localized Generalized Linear Models.

ACKNOWLEDGMENT

We wish to thank Professor F. Planchet for helpful and constructive suggestions which he provided in relation to earlier draft of this article. We are also grateful to a referee for a careful reading of the manuscript and comments which lead to an improved version of the paper.

REFERENCES

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.

BLOOMFIELD, D. S. F. and HABERMAN, S. (1987). Graduation: some experiments with kernel methods. *Journal of the Institute of Actuaries*, **114**, 339-369.

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. Hardle 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.

COPAS, J. B. and HABERMAN, S. (1983). Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries*, **110**, 135-156.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377-403.

DEBON, 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.

DIEWERT, W. E. and WALES, T. J. (2006). A "new" approach to the smoothing problem. In *Money measurement and computation*, page 104-144. Palgrave Macmillan.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425-455.

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., FARMEN, M., and GIJBELS, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society*, **60**(3), 591-608.

FELIPE, A., GUILLEN, M., and PEREZ-MARIN, 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., KNEIP, A., and KOHLER, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the american statistical association*, **86**(415), 643-652.

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*, **XLVII**, 173-209.

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.

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.

Human Mortality Database (2011). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 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.

KEYFITZ, N. (1981). The limits of population forecasting. *Population and development review*, **7**(4), 579-593.

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.

MÜLLER, H. G. (1987). Weighted local regression and kernel method for nonparametric curve fitting. *Journal of the American Statistical Association*, **82**, 231--238.

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**(409), 66-72.

PLANCHET, F. and WINTER, P. (2007). L'utilisation des splines bidimentionnels pour l'estimation de lois de maintien en arrêt de travail. *Bulletin Français d'Actuariat*, **13**(7), 83-106.

R Development Core Team (2011). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

RICE, J. A. (1984). Bandwidth choice for non-parametric regression. *Annals of Statistics*, **12**(4), 1215-1230.

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**(3), 1346--1370.

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.

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.

TAYLOR, G. (1992). A bayesian interpretation of whittaker-henderson graduation. *Insurance: Mathematics and Economics*, **11**(1), 7-16.

TOMAS, J. (2011a).A local likelihood approach to univariate graduation of mortality. *Bulletin Français d'Actuariat*, **11**(22), 105--153.

TOMAS, J. (2011b). On boundaries effects and practical considerations for univariate graduation of mortality by local likelihood models. pages 1-32. Unpublished.

TSYBAKOV, A. B. (1986). Robust reconstruction of functions by the local approximation method. *Problems of Information Transmission*, **22**(2), 133--146.