Kernel Smoothing

M.P. Wand
Department of Biostatistics
Harvard School of Public Health
Boston, MA, US

M.C. Jones
Department of Statistics
The Open University
Milton Keynes, UK

CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

Contents

Preface

Kernel smoothing refers to a general class of techniques for nonparametric estimation of functions. Suppose that you have a univariate set of data which you want to display graphically. Then kernel smoothing provides an attractive procedure for achieving this goal, known as kernel density estimation. Another fundamental example is the simple nonparametric regression or scatterplot smoothing problem where kernel smoothing offers a way of estimating the regression function without the specification of a parametric model. The same principles can be extended to more complicated problems, leading to many applications in fields as diverse as medicine, engineering and economics. The simplicity of kernel estimators entails mathematical tractability, so one can delve deeply into the properties of these estimators without highly sophisticated mathematics. In summary, kernel smoothing provides simple, reliable and useful answers to a wide range of important problems.

The main goals of this book are to develop the reader's intuition and mathematical skills required for a comprehensive understanding of kernel smoothing, and hence smoothing problems in general. Exercises designed for achieving this goal have been included at the end of each chapter. We have aimed this book at newcomers to the field. These may include students and researchers from both the statistical sciences and interface disciplines. Our feeling is that this book would be appropriate for most first or second year statistics graduate students in the North American system, honours level students in the Commonwealth system and students at similar stages in other systems. In its role as an introductory text this book does make some sacrifices. It does not completely cover the vast amount of research in the field of kernel smoothing, and virtually ignores important work on non-kernel approaches to smoothing problems. It is hoped that the bibliographical notes near the end of each chapter will provide sufficient access to the wider field.

In completing this book we would like to extend our most sincere gratitude to Peter Hall and Steve Marron. Peter is responsible for a substantial portion of the deep theoretical understanding of kernel smoothing that has been established in recent years, and his generosity as both supervisor and colleague has been overwhelming. Steve Marron, through many conversations and countless e-mail messages, has been an invaluable source of support and information during our research in kernel smoothing. His probing research, philosophies, insight and high standards of presentation have had a strong influence on this book. We also give special thanks to David Ruppert and Simon Sheather for their advice, ideas and support. It was largely Bernard Silverman's influence that first led to M.C.J.'s interest in this topic, for which he is most grateful. Our ideas and opinions on smoothing have also been moulded by our contact, collaboration and collegiality with many other prominent researchers in the field, including Adrian Bowman, Ray Carroll, Chris Carter, John Copas, Dennis Cox, Luc Devroye, Geoff Eagleson, Joachim Engel, Randy Eubank, Jianqing Fan, Theo Gasser, Wenceslao González-Manteiga, Wolfgang Härdle, Jeff Hart, Nancy Heckman, Nils Hjort, Iain Johnstone, Robert Kohn, Oliver Linton, Hans-Georg Müller, Jens Nielsen, Doug Nychka, Byeong Park, M. Samiuddin, Bill Schucany, David Scott, Joan Staniswalis, Jim Thompson and Tom Wehrly.

Drafts of this text have been read by Glen Barnett, Angus Chow, Inge Koch, Alun Pope, Simon Sheather, Mike Smith, Frederic Udina, Sally Wood and an anonymous reader, as well as several students who participated in a course given from this book in the Department of Statistics at the University of New South Wales. Their feedback, comments and corrections are very gratefully acknowledged. Parts of this book were written while one of us (M.P.W.) was visiting Rice University, National University of Singapore and University of British Columbia. The support of these institutions is also gratefully acknowledged. We would also like to thank Yusuf Mansuri of the Australian Graduate School of Management for the excellent computing and word processing support that he has provided throughout this project.

Finally, we must express our deepest thanks to our partners Handan and Ping and our families for their constant support and encouragement.

Kensington and Milton Keynes Matt Wand
April 1994 Chris Jones

CHAPTER 1

Introduction

1.1 Introduction

Kernel smoothing provides a simple way of finding structure in data sets without the imposition of a parametric model. One of the most fundamental settings where kernel smoothing ideas can be applied is the simple regression problem, where paired observations for each of two variables are available and one is interested in determining an appropriate functional relationship between the two variables. One of the variables, usually denoted by X, is thought of as being a *predictor* for the other variable Y, usually called the *response* variable.

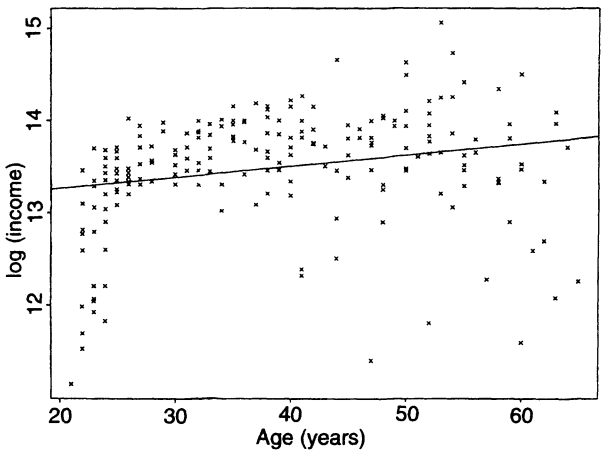

Figure 1.1. *Scatterplot of age/log(income) data. The ordinary least squares line is also shown.*

Figure 1.1 is a scatterplot with each cross representing the age

(the X variable) and log(income) (the Y variable) of 205 Canadian workers (source: Ullah, 1985). There is interest in modelling log(income) as a function of age. A first attempt might be to fit a straight line to the data. The line shown in Figure 1.1 is the ordinary least squares straight line fit to the observations. What are we assuming when modelling Y as a linear function of X? The usual assumption is that the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of age/log(income) pairs satisfies the relationship

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where the *errors* ε_i are symmetric random variables having zero mean. However, this assumption, which entails that the observations are randomly scattered about a straight line, appears to be far from valid for this scatterplot.

The linear model (1.1) is an example of a *parametric regression* model. Let us clarify this terminology. A well known result from elementary statistics is that the function m for which $E\{Y - m(X)\}^2$ is minimised is the conditional mean of Y given X, that is

$$m(X) = E(Y|X).$$

This function, the best mean squared predictor of Y given X, is often called the *regression* of Y on X. It follows from the definition of m that

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where $E(\varepsilon_i) = 0$ for each i. In model (1.1) we are therefore making the assumption that the functional form of the regression function m is known except for the values of the two *parameters* β_0 and β_1. This is the reason for the term *parametric* since the family of functions in the model can be specified by a finite number of parameters.

There are several other parametric regression models which one could use. Some examples are

$$\begin{aligned} & Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \\ & Y_i = \beta_1 \sin(\beta_2 X_i) + \varepsilon_i \\ \text{and} \quad & Y_i = \beta_1 / (\beta_2 + X_i) + \varepsilon_i. \end{aligned}$$

The choice of parametric model depends very much on the situation. Sometimes there are scientific reasons for modelling Y as a particular function of X, while at other times the model is

based on experience gained through analysis of previous data sets of the same type. There is, however, a drawback to parametric modelling that needs to be considered. The restriction of m belonging to a parametric family means that m can sometimes be too rigid. For example, the models above respectively require that m be parabolic, periodic or monotone, each of which might be too restrictive for adequate estimation of the true regression function. If one chooses a parametric family that is not of appropriate form, at least approximately, then there is a danger of reaching incorrect conclusions in the regression analysis.

The rigidity of parametric regression can be overcome by removing the restriction that m belong to a parametric family. This approach leads to what is commonly referred to as *nonparametric regression*. The philosophical motivation for a nonparametric approach to regression is straightforward: when confronted with a scatterplot showing no discernible simple functional form then one would want to let the data decide which function fits them best without the restrictions imposed by a parametric model (this is sometimes referred to as "letting the data speak for themselves"). However, nonparametric and parametric regression should not be viewed as mutually exclusive competitors. In many cases a nonparametric regression estimate will suggest a simple parametric model, while in other cases it will be clear that the underlying regression function is sufficiently complicated that no reasonable parametric model would be adequate.

There now exist many methods for obtaining a nonparametric regression estimate of m. Some of these are based on fairly simple ideas while others are mathematically more sophisticated. For reasons given in Section 1.3 we will study the kernel approach to nonparametric regression.

Figure 1.2 shows an estimate of m for the age/log(income) data, using what is often called a *local linear kernel estimator*. The function shown at the bottom of the plot is a *kernel* function which is usually taken to be a symmetric probability density such as a normal density. The value of the estimate at the first point u is obtained by fitting a straight line to the data using *weighted* least squares, where the weights are chosen according to the height of the kernel function. This means that the data points closer to u have more influence on the linear fit than those far from u. This local straight line fit is shown by the dotted curve and the regression estimate at u is the height of the line at u. The estimate at a different point v is found the same way, but with the weights chosen according to the heights of the kernel when centred around v. This

estimator fits into the class of *local polynomial* regression estimates (Cleveland, 1979). Nonparametric regression estimators are often called *regression smoothers* or *scatterplot smoothers*, while those based on kernel functions are often called *kernel smoothers.*

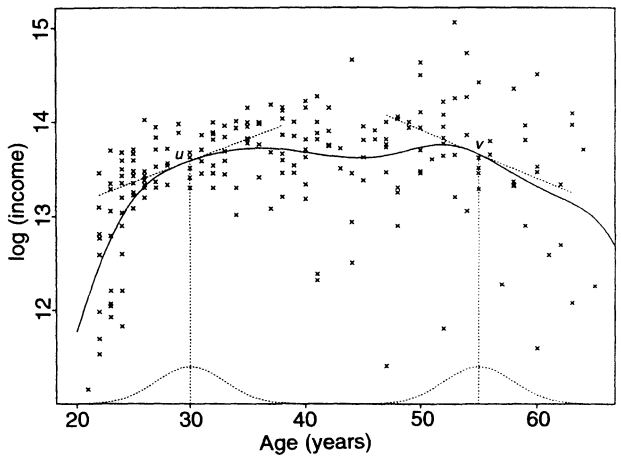

Figure 1.2. *Local linear kernel regression estimate based on the age/log(income) data. The solid curve is the estimate. The dotted curves are the kernel weights and straight line fits at points u and v.*

While the kernel-based nonparametric regression estimator described here means that we have a much more flexible family of curves to choose from, this increased flexibility has its costs and leads to several new questions. Some examples are

- What are the statistical properties of kernel regression estimators?
- What influence does the shape of the kernel function have on the estimator?
- What influence does the scaling of the kernel function have on the estimator?
- How can this scaling be chosen in practice?
- How can kernel smoothing ideas be used to make confidence statements rather than just giving point estimates?
- How do dependencies in the data affect the kernel regression estimator?
- How does one best deal with multiple predictor variables?

While some of these questions have reasonably straightforward solutions, others are the subject of ongoing research and may not

be completely resolved for many years. Some can be answered by relatively simple mathematical arguments, while others require deeper analyses that are beyond the scope of this book.

We chose to introduce the motivation and ideas of kernel smoothing by nonparametric regression because of the familiarity of the regression problem to most readers. However, as we will see, kernel smoothing can be applied to many other important curve estimation problems such as estimating probability density functions, spectral densities and hazard rate functions. The first of these is discussed in the next section.

1.2 Density estimation and histograms

Perhaps an even more fundamental problem than the regression problem is the estimation of the common probability density function, or *density* for short, of a univariate random sample. Suppose that $X_1, \ldots, X_n$ is a set of continuous random variables having common density f. The parametric approach to estimation of f involves assuming that f belongs to a parametric family of distributions, such as the normal or gamma family, and then estimating the unknown parameters using, for example, maximum likelihood estimation. On the other hand, a *nonparametric density estimator* assumes no pre-specified functional form for f.

The oldest and most widely used nonparametric density estimator is the *histogram.* This is usually formed by dividing the real line into equally sized intervals, often called *bins.* The histogram is then a step function with heights being the proportion of the sample contained in that bin divided by the width of the bin. If we let b denote the width of the bins, usually called the *binwidth,* then the histogram estimate at a point x is given by

$$\hat{f}_H(x;b) = \frac{\text{number of observations in bin containing } x}{nb}.$$

Two choices have to be made when constructing a histogram: the binwidth and the positioning of the bin edges. Each of these choices can have a significant effect on the resulting histogram. Figure 1.3 show four histograms based on the same set of data. These data represent 50 birthweights of children having severe idiopathic respiratory syndrome (source: van Vliet and Gupta, 1973). The first two histograms are based on a small and a large binwidth ($b = 0.2$ and $b = 0.8$) respectively. The bottom two are

based on the same medium sized binwidth ($b = 0.4$), but with bin edges shifted by half a binwidth.

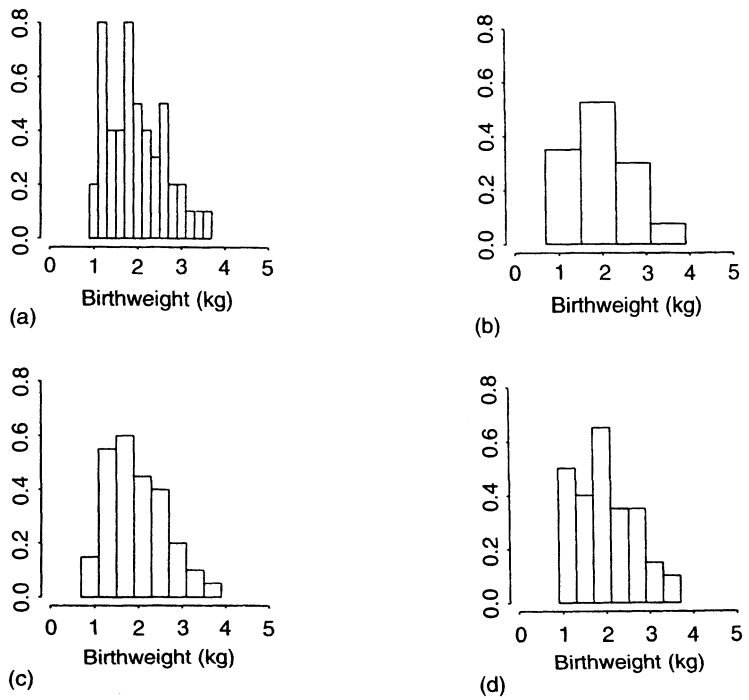

Figure 1.3. *Histograms of birthweight data. Figures (a) and (b) are based on binwidths of 0.2 and 0.8 respectively. Figures (c) and (d) are each based on a binwidth of 0.4 but with left bin edge at 0.7 and 0.9 respectively.*

Notice that each histogram gives a different impression of the shape of the density of the data. A smaller binwidth leads to a relatively jagged histogram, while a larger binwidth results in a smoother looking histogram as shown in Figures 1.3 (a) and (b). Figures 1.3 (c) and (d) show that the placement of the bin edges also has an effect since the density shapes suggested by these histograms are quite different from each other, despite the equal binwidths.

The binwidth b is usually called a *smoothing parameter* since it controls the amount of "smoothing" being applied to the data.

All nonparametric curve estimates have an associated smoothing parameter. We will see in the following chapters that, for kernel estimators, the scale of the kernel plays a role analogous to that of the binwidth. In parametric polynomial regression the degree of

the polynomial can be thought of as being a smoothing parameter.

The sensitivity of the histogram to the placement of the bin edges is a problem not shared by other density estimators such as the kernel density estimator introduced in Chapter 2. The bin edge problem is one of the histogram's main disadvantages. The logical remedy to this problem is the *average shifted histogram* (Scott, 1985), which averages several histograms based on shifts of the bin edges, but this can be shown to approximate a kernel density estimator, and thus provide an appealing motivation for kernel methods. The histogram has several other problems not shared by kernel density estimators. Most densities are not step functions, yet the histogram has the unattractive feature of estimating all densities by a step function. A further problem is the extension of the histogram to the multivariate setting, especially the graphical display of a multivariate histogram. Finally, the histogram can be shown not to use the data as efficiently as the kernel estimator. This deficiency is discussed at the end of Section 2.5. Despite these drawbacks, the simplicity of histograms ensures their continuing popularity.

1.3 About this book

As the title and the previous two sections suggest, this book is about kernel smoothing as a means of obtaining nonparametric curve estimators. Kernel estimators have the advantage of being very intuitive and relatively simple to analyse mathematically. Even if one prefers to use other nonparametric smoothing methods, such as those based on spline functions, an understanding of the main issues involved can best be gained through studying kernel estimators.

Kernel estimators have been around since the seminal papers of Rosenblatt (1956) and Parzen (1962), although the basic principles were independently introduced by Fix and Hodges (1951) and Akaike (1954). Since then articles written about kernel estimators number in the thousands, and there will be many more written in the future. It follows that there are still many unresolved and controversial issues. This book makes no endeavour to survey the field of kernel estimation, nor does it try to provide an answer for every question concerning the practical implementation of kernel estimators. Instead our goal is to present the reader with the aspects of kernel smoothing which we see as being most fundamental and practically most relevant at the time of writing.

The choice of topics in the pursuit of this goal is necessarily a personal one. Moreover, because of ongoing research into the practical implementation of kernel estimators we will, in the main, avoid the more unsettled issues in the field. The main purpose of this book is to enhance the reader's intuition and mathematical skills required for understanding kernel smoothing, and hence smoothing problems in general.

We believe that the readability of the book is improved by postponing detailed referencing to bibliographical notes, which are provided near the end of each chapter. These are followed by a set of exercises which aim to familiarise the reader with the above-mentioned mathematical skills.

We begin our study of kernel smoothing with the univariate kernel density estimator in Chapter 2. This is because kernel density estimation provides a very simple and convenient way of developing an understanding of the main ideas and issues involved in kernel smoothing in general.

As we show in Chapter 2, one of the central issues in kernel smoothing is the choice of the smoothing parameter, often called the *bandwidth* for kernel estimators. The choice of the bandwidth from the data has become an important topic in its own right in recent years and is still a burgeoning area of research. In Chapter 3 we discuss some of the more popular approaches to bandwidth selection, again in the simple density estimation context. While our coverage of this topic is far from complete, our aim is to give the reader a flavour for the types of approaches and problems faced when selecting a bandwidth from the data.

Chapter 4 is devoted to the extension of the kernel density estimator to multivariate data. While this extension is fairly obvious in principle, the mathematical analyses and practical implementation are non-trivial, as Chapter 4 shows.

In Chapter 5 we return to the important problem of nonparametric regression. The notions of kernel smoothing learnt from studying the kernel density estimator prove to be useful for understanding the more complicated kernel regression problem.

Chapter 6 is a collection of extra topics from the kernel smoothing literature which portray various extensions of the material from the previous chapters, and indicate the wide applicability of the general ideas.

There are four appendices. Appendix A lists notation used throughout this book. Tables in Appendix B contain useful results for several common densities. Facts about the normal density,

relevant to kernel smoothing, are given in Appendix C. Appendix D describes computation of kernel estimates.

1.4 Options for reading this book

This book may be read in several different ways. To gain a basic understanding of the main ideas of univariate kernel estimation for important settings one should consult Chapters 2 and 5. Chapter 4 could be added for an understanding of multivariate kernel smoothing. Chapter 3 stands out as the only chapter concerned with full data-driven implementation of kernel estimators, but is also the least settled topic in this book, and could be omitted without loss of continuity. Chapter 6 is a selection of extra topics which could be covered depending on time and interest. We have ordered the chapters in what we see as being the most natural sequence if this book is to be completely covered.

1.5 Bibliographical notes

For a detailed study of the histogram we refer the reader to Scott (1992).

Important references for kernel smoothing will be given at the end of the appropriate chapter. At this point we will just mention some recent books on the subject. A very readable introduction to kernel density estimation is given by Silverman (1986). Kernel density estimation is also treated in books by Devroye and Györfi (1985), Härdle (1990a) and Scott (1992). Recent books that treat nonparametric kernel regression to varying extents are Eubank (1988), Müller (1988), Härdle (1990b) and Hastie and Tibshirani (1990). Wahba (1990) and Green and Silverman (1994) are recent monographs on spline approaches to nonparametric regression, and Tarter and Lock (1993) treats orthogonal series density estimation.