# Graphics and Inference in Nonparametric Modelling

Stuart Gordon Young

Thesis submitted for the degree of Ph.D.

University of Glasgow

Department of Statistics, Faculty of Science

*October 1996*

# Summary

This thesis is concerned with statistical modelling techniques which involve nonparametric smoothing. Its principal aim is to introduce a number of graphical methods of assessing nonparametric models, and also examine the important area of formal inference in the nonparametric setting.

Chapter 1 gives a broad introduction to the ideas and methods contained in this thesis, and includes a description of kernel smoothing, and some key results.

Chapter 2 leads off with a general treatment of three-dimensional data, and a method of displaying such data, by including a surface, relating to a model of interest. This surface can bring out the three-dimensional structure by providing a helpful visual reference. This basic idea is applied to a number of areas, such as regression, survival data, principal component analysis, and linear discrimination. Nonparametric modelling is introduced towards the end of this chapter, when three-dimensional density estimation is considered, and a contouring procedure examined. In the area of nonparametric smoothing, whether density estimation or regression, much of the literature deals with the selection of a smoothing parameter. A common feature of many of the nonparametric techniques introduced in this thesis is that the influence of the smoothing parameter is reduced. With this often contentious issue assuming less importance, more emphasis can be placed on meaningful interpretations of the data. The three-dimensional contouring procedure discussed here is the first example of this approach.

Exploring the nonparametric theme more fully, Chapter 3 deals with nonparametric regression, and in particular, graphical methods of comparing several curves. This comparison is achieved by means of a "reference band", which, while not removing the need for a formal test, can provide a useful visual interpretation of a model. Reference bands for equality are derived and explored in a variety of settings. Reference bands for parallelism are also derived for nonparametric regression models. The subject of bias, so often a concern in nonparametric modelling, is eliminated in this context by appropriate choice of smoother.

The important topic of inference is addressed in Chapter 4, which introduces tests for a nonparametric analysis of covariance model, which provide formal means of analysis to accompany some of the graphical methods introduced in Chapter 3. Tests for equality and parallelism are derived, and their power assessed via a simulation study.

Chapter 5 extends the ideas of Chapter 4 by applying them to binary response data, and a test for equality is introduced for this case. The approach here is necessarily different from that in Chapter 4, and involves deriving the mean and variance of the test statistic exactly, in order to permit a null distribution to be found. An example from Chapter 3 is re-visited here, and the formal test applied, to accompany the reference band produced previously. As in Chapter 4, the size and power of the new test is investigated via a simulation study. Competing null distributions are also assessed for their suitability.

Finally, Chapter 6 widens the scope of the thesis, by considering more general models in several covariates. Graphical methods and formal tests, which involve looking for patterns in residuals, are considered for semiparametric models, containing both parametric and nonparametric components. Inference in bivariate smoothing is also addressed, with a particular application in image analysis. This is extended to a general test for comparing smooth surfaces, using the principles of Chapter 4. The thesis closes with a discussion of the results presented, and what further areas could be investigated.

# Acknowledgements

I would like to thank the SERC for providing a Research Studentship Award, under which most of the work contained in this thesis was carried out.

I would also like to give my utmost thanks to Prof. Adrian Bowman, whose supervision and patience has been invaluable over the years.

Some of the work presented here has been published in journals, and I am grateful for the comments of referees, which improved earlier versions of the text.

My heartfelt gratitude goes to my family and friends for all their support and encouragement. This thesis has been a long time coming, and my parents deserve praise for their tolerance in the bleakest moments. I am grateful to my colleagues in the Statistics Department, for all they have done for me, and I feel I must mention Mary Nisbet and Myra Smith in particular. My thanks also to Dougy Watt, who in sharing an office with me, also shared my frustration, as I shared his.

And thanks to Lynne, for being there, for always showing me the bright side, and for giving me the final push when it was needed.

*To Mum and Dad, Lynne, and absent friends*

# Contents

**1 Introduction** **1**

1.1 Scope of this thesis . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Nonparametric regression . . . . . . . . . . . . . . . . . . . . . . . 4

1.2.1 Smoothing parameter . . . . . . . . . . . . . . . . . . . . . . 9

1.2.2 Choosing a smoothing parameter by cross-validation . . . . . 11

1.2.3 Plug-in methods of smoothing parameter selection . . . . . . . 12

1.2.4 Degrees of freedom of a smoother . . . . . . . . . . . . . . . . 13

1.2.5 Bias . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

1.2.6 Boundary effects . . . . . . . . . . . . . . . . . . . . . . . . . 18

**2 Three-dimensional plots of data and models** **19**

2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2.2 The geometry of three-dimensional projections . . . . . . . . . . . . 21

2.3 Regression data . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

2.3.1 Displays with two covariates . . . . . . . . . . . . . . . . . . . 24

2.3.2 Adding information on a third explanatory variable . . . . . . 30

2.3.3 Displaying survival data . . . . . . . . . . . . . . . . . . . . . . . . 34

2.4 More general three-dimensional data . . . . . . . . . . . . . . . . . . . 36

2.4.1 Discrimination reference planes . . . . . . . . . . . . . . . . . . . 36

2.4.2 Principal component reference planes . . . . . . . . . . . . . . . 39

2.4.3 Density contours . . . . . . . . . . . . . . . . . . . . . . . . . . 41

2.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46

**3 Graphical comparison of nonparametric curves 49**

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.2 Nonparametric regression curves . . . . . . . . . . . . . . . . . . . . . 50

3.2.1 A reference band for equality . . . . . . . . . . . . . . . . . . . . 53

3.2.2 Regression with binary responses . . . . . . . . . . . . . . . . . . 58

3.3 Survival curves . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

3.3.1 Reference bands for parallelism . . . . . . . . . . . . . . . . . . . 64

3.4 Comparing nonparametric and parametric curves . . . . . . . . . . . 67

3.4.1 A reference band for linearity . . . . . . . . . . . . . . . . . . . . 67

3.4.2 A reference band for residual plots . . . . . . . . . . . . . . . . . 70

3.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

**4 Nonparametric analysis of covariance 76**

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76

4.2 Tests of equality and parallelism . . . . . . . . . . . . . . . . . . . . . 80

4.2.1 A test statistic for equality . . . . . . . . . . . . . . . . . . . . . . 80

4.2.2 Distribution of the test statistic . . . . . . . . . . . . . . . . . 82

4.2.3 Estimation and inference in the parallel model . . . . . . . . . 84

4.3 A practical example . . . . . . . . . . . . . . . . . . . . . . . . . 87

4.3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . 87

4.3.2 White Spanish Onions . . . . . . . . . . . . . . . . . . . . 87

4.3.3 Brown Spanish Onions . . . . . . . . . . . . . . . . . . . . 91

4.4 A simulation study . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.4.2 Test of equality . . . . . . . . . . . . . . . . . . . . . . . 95

4.4.3 Test of parallelism . . . . . . . . . . . . . . . . . . . . . . 97

4.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101

**5 Analysis of covariance with binary response data 102**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102

5.2 A nonparametric test statistic . . . . . . . . . . . . . . . . . . . . 104

5.2.1 Obtaining a null distribution . . . . . . . . . . . . . . . . . 106

5.2.2 Expectation of the test statistic . . . . . . . . . . . . . . . . 107

5.2.3 Variance of the test statistic . . . . . . . . . . . . . . . . . 108

5.2.4 Using the two-moment approximation . . . . . . . . . . . . . 110

5.3 A practical example . . . . . . . . . . . . . . . . . . . . . . . . . 111

5.4 Some simulation studies . . . . . . . . . . . . . . . . . . . . . . . 113

5.4.1 Power . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114

5.4.2 Size . . . 120

5.5 Discussion . . . 122

**6 Introducing more covariates 123**

6.1 Introduction . . . 123

6.2 Testing a model with a single nonparametric component . . . 125

6.2.1 Illustrating the method of residual smoothing by example . . . 127

6.3 Multivariate smoothing . . . 130

6.3.1 An application in image analysis . . . 132

6.3.2 Applying the test to the example data . . . 134

6.4 Multivariate analysis of covariance . . . 136

6.4.1 Bias in the multivariate case . . . 136

6.4.2 A test of bivariate equality . . . 137

6.4.3 The curse of dimensionality . . . 139

6.5 Discussion and final comments . . . 140

**Appendix 143**

**References 145**

# List of Figures

1.2.1 Illustration of the effect of the smoothing parameter . . . . . . . . . . 10

1.2.2 Illustration of the relationship between smoothing parameter and degrees of freedom . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

1.2.3 Example of smoothing with two choices of smoothing parameter . . . 16

2.2.1 Illustration of perspective geometry . . . . . . . . . . . . . . . . . . . . 22

2.3.2 Scatterplots of Church of England data . . . . . . . . . . . . . . . . . 26

2.3.3 Three dimensional plot of the Church of England data . . . . . . . . 28

2.3.4 A second three-dimensional plot of the Church of England data . . . 29

2.3.5 Regression with the stack loss data . . . . . . . . . . . . . . . . . . . 31

2.3.6 Stack loss data with the transitional state points removed . . . . . . . 33

2.3.7 Plot of the VA lung cancer data with fitted survival model . . . . . . 35

2.4.8 Discrimination with the Conn's Syndrome data . . . . . . . . . . . . 37

2.4.9 Discrimination with a new observation added for diagnosis . . . . . . 38

2.4.10 Geyser data with the plane of the first two principal components . . 40

2.4.11 The 75% contour of a three-dimensional density estimate of the geyser data . . . . . . . . . . 44

2.4.12 The 50% contour of a three-dimensional density estimate of the geyser data . . . . . . . . . . 47

3.2.1 Brown Spanish Onions data . . . . . . . . . . 54

3.2.2 Reference band for equality of nonparametric regression curves . . . . 57

3.2.3 Reference band for equality, with a different choice of smoothing parameter . . . . . . . . . . 59

3.2.4 Reference band for equality of nonparametric regression with binary data . . . . . . . . . . 60

3.3.5 Reference band for equality of survivor functions . . . . . . . . . . 63

3.3.6 Reference bands for parallelism . . . . . . . . . . 65

3.4.7 A reference band for linearity on the geyser data . . . . . . . . . . 69

3.4.8 A reference band for trend in a residual plot . . . . . . . . . . 71

3.4.9 A reference band for the stack loss data . . . . . . . . . . 73

4.1.1 White Spanish Onions data . . . . . . . . . . 77

4.3.2 Brown Spanish Onions data . . . . . . . . . . 88

4.3.3 Significance trace for the test of parallelism on White Onions data . . 90

4.3.4 A second significance trace for the White Onions data . . . . . . . . . 91

4.3.5 Significance trace for the test of equality on Brown Onions data . . . 92

4.4.6 A set of sample simulation plots . . . . . . . . . . 94

5.3.1 Significance trace for equality of nonparametric regression with binary data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 112

5.4.2 Plot of logistic models for the simulation study . . . . . . . . . . . . . 115

5.4.3 A set of sample simulation plots . . . . . . . . . . . . . . . . . . . . . . . 116

6.2.1 Significance trace for test of a linear term in air flow . . . . . . . . . . 128

6.2.2 Significance trace for test of a quadratic term in air flow . . . . . . . 129

6.3.3 Plots of the two sample images . . . . . . . . . . . . . . . . . . . . . . 132

6.3.4 Plots of the observed and smoothed differences . . . . . . . . . . . . . 133

6.3.5 Significance trace for the test of equality of images . . . . . . . . . . . 135

# List of Tables

2.3.1 Data on giving in the Church of England . . . . . . . . . . . . . . . . 25

4.4.1 Powers for the test of equality for the underlying linear function . . . 95

4.4.2 Powers for the test of equality for the underlying sine function . . . . 96

4.4.3 Powers for the test of equality for the underlying quadratic function . 97

4.4.4 Powers for the test of parallelism for the underlying linear function . 98

4.4.5 Powers for the test of parallelism for the underlying sine function . . 99

4.4.6 Powers for the test of parallelism for the underlying quadratic function 100

5.4.1 Powers for the test of equality based on 1000 simulations with $n_i = 25$ 117

5.4.2 Powers for the test of equality based on 1000 simulations with $n_i = 50$ 118

5.4.3 A second simulation study for the test of equality based on 1000 simulations with $n_i = 25$ . . . . . . . . . . . . . . . . . . . . . . . . . 119

5.4.4 A second simulation study for the test of equality based on 1000 simulations with $n_i = 50$ . . . . . . . . . . . . . . . . . . . . . . . . . 119

5.4.5 Assessing the size of the test of equality based on 1000 simulations with $n_i = 25$ . . . . . . . . . . . . . . . . . . . . . . . . . . 121

5.4.6 Assessing the size of the test of equality based on 1000 simulations with $n_i = 50$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

# Chapter 1

# Introduction

## 1.1 Scope of this thesis

A helpful plot can often express a great deal about a set of data. While the formal side of an analysis should never be ignored, a plot, when appropriate, can provide information in an immediate and accessible manner.

This thesis opens with graphical methods, but later chapters also cover the important subject of inference, and in particular, methods of inference to complement many of the plots introduced earlier.

The main subject matter is nonparametric (and semiparametric) modelling, but Chapter 2 leads off with a general treatment of ways of displaying three-dimensional data. Representations of such data in the form of spinning point clouds are quite common, however this chapter investigates static displays, incorporating a parametric model included for reference and to bring out the three-dimensional structure.

This method is applied in a number of settings, such as regression, survival data, discrimination and principal component analysis. The chapter ends by introducing the nonparametric theme, by way of three-dimensional density estimation, and a contour of such an estimate. The contouring procedure extends proposals made by Bowman & Foster (1993), and a method of displaying the results is presented.

Later chapters concentrate on nonparametric regression, beginning with Chapter 3, which introduces the idea of a reference band for the comparison of two nonparametric regression curves. This is an analysis of covariance model, where parametric assumptions are relaxed, and the covariate effect is assumed only to be smooth. The reference band is derived from the standard error of the difference between the two curves at each point, has a simple hypothesis testing interpretation, and uses properties of a nonparametric smoother to alleviate problems of bias. An introduction to nonparametric smoothing is given in Section 1.2. In addition to the standard nonparametric regression context, the method is also applied to examples involving binary response and survival data. The chapter closes with a look at comparisons between parametric and nonparametric models, which can be used to assess goodness-of-fit.

The use of a reference band does not remove the need for a global test of the effects of interest. Rather, the benefit lies in providing a very useful graphical tool to explore where any identified differences may lie, or in explaining why apparent differences do not actually contribute strong evidence of statistical significance to the global comparison of curves. This important area of inference is addressed first in Chapter 4, which introduces global tests of equality and additivity, for the analysis of covariance model. This chapter provides inferential procedures to accompany some of the visual methods introduced in Chapter 3. Highly-accurate moment-based approximations to the null distributions of the suggested test statistics are used to construct the tests, and power is assessed through simulation studies.

In nonparametric modelling, the choice of smoothing parameter, which controls the degree of smoothing of the data, can often be a very important consideration. It is demonstrated that, in these applications, the choice of smoothing parameter can take on less significance, with respect to the question of interest. This is illustrated by means of a "significance trace", first introduced by Azzalini & Bowman (1991). Reducing the importance of the smoothing parameter is a theme common to many of the nonparametric techniques presented in this thesis.

Chapter 5 extends the ideas of Chapter 4 by applying them to the case where the response is binary, and a probability, or proportion, is modelled. This again ties in with Chapter 3, by introducing a global test to accompany the idea of a reference band. The nature of the data in the binary response case dictates that the methods of Chapter 4, involving quadratic forms in Normal variates, cannot be used to determine the null distribution here. In this situation, the mean and variance of the test statistic are derived exactly, and these moments are used to provide an approximate null distribution. The test is then applied to a set of data already discussed in Chapter 3, to complement the pictorial method of the reference band. Finally, simulation studies are again used to investigate the size and power of the new test.

The test of additivity proposed in Chapter 4 introduces the topic of semiparametric models. A semiparametric model is one which contains both parametric and nonparametric components, and a "parallel curves" model, such as the one described in Chapter 4, is the simplest example. More complicated models arise when additional covariates, in particular continuous covariates, are present, and these are addressed in Chapter 6. This chapter begins by considering regression models with several covariates, one of which is modelled as a smooth function. Once again, this provides a formal method of inference to back up graphical techniques introduced earlier. This is followed by an extension of the analysis of covariance model to two continuous covariates, leading to a comparison of surfaces. A test of equality is derived, with particular application in image analysis, and an example in that area is examined. The principle is then extended to the general case of two or more surfaces, and a further test introduced. The chapter closes with a discussion on Generalized Additive Models, as developed by Hastie & Tibshirani (1990), and how the scope of the methods of inference introduced in this thesis might be extended.

## 1.2 Nonparametric regression

Suppose data are collected in the form $(x_i, y_i)$, $i = 1, \ldots, n$, and it is of interest to model the relationship between the covariate, $x$, and the response, $y$. In other words, it is required to model the conditional mean of $y$ given $x$. Let $g(x) = E(y|x)$. The usual Normal linear model is of the form

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1.2.1}$$

where the $\varepsilon_i$ are independent, identically-distributed Normal random variables with mean 0 and variance $\sigma^2$. This is a *parametric* model, and assumes that the regression function $g$ is linear, and can be expressed in terms of the unknown parameters $\alpha$ and $\beta$. Many other parametric models are possible, for example higher-order polynomials, and specific situations or prior knowledge may prompt functions other than the simple straight-line one in formula (1.2.1). The drawback of parametric regression is that the functional form chosen for $g$ can sometimes be too rigid, and if it is not appropriate (at least approximately), then incorrect conclusions could result from the subsequent analysis.

A *nonparametric* approach is much more flexible, and specifies only that $g$ is a *smooth* function. The motivation is quite straightforward, and it is that when a scatterplot is produced which shows no obvious parametric form, it is simplest to allow the data to determine which shape $g$ should follow. In that sense, nonparametric regression is a *data-driven* technique.

Nonparametric smoothing methods have, in the main, been developed relatively recently. This development has been motivated principally because of advances in computational capabilities, which make these methods practical, but also from the realisation that a parametric approach does not always allow the flexibility required.

A number of nonparametric regression techniques have been developed, such as spline smoothing and wavelets. The method used throughout this thesis is *kernel smoothing*, which benefits from mathematical and intuitive simplicity, and basically entails weighted averaging of the observed responses.

To illustrate, for a particular covariate value, $x$, an estimate of $g(x)$ is obtained by taking a weighted average of the responses $y_1, \ldots, y_n$. Responses corresponding to those observed covariate values closest to $x$ receive more weight than those further away: logically, points closest to the point of interest should carry more information about the behaviour of the regression function at $x$ than those further away. The principal concern with a nonparametric approach is how these weights are chosen.

Perhaps the simplest kernel regression smoother is one known as the Nadaraya-Watson technique, introduced independently by Nadaraya (1964) and Watson (1964). This can be expressed as

$$\hat{g}_{nw}(x) = \frac{\sum_{j=1}^{n} k\left(\frac{x - x_j}{h}\right) y_j}{\sum_{j=1}^{n} k\left(\frac{x - x_j}{h}\right)}. \tag{1.2.2}$$

Here the weights are chosen by evaluating a kernel function $k$ at the design points, and then dividing by the sum of the weights, so that they add up to 1. The kernel function often takes the simple form of a symmetric probability density function with single mode at 0, and throughout this thesis, standard Normal kernels are used. The amount of local averaging, and therefore the smoothness of the estimator, is controlled by the value of $h$, which is referred to as a *smoothing parameter* (or *bandwidth*). It is clear from equation (1.2.2) that in this case this smoothing parameter is the standard deviation of a Normal density. A large value of $h$ gives a wide kernel which encompasses a large number of points in the weighting process, and produces a very smooth estimator. This will possess small variance but large bias. As $h$ decreases, smoothness also decreases, as the kernel covers fewer points, producing an estimator with smaller bias but larger variance. The subject of bias is very important throughout this thesis, and will be dealt with in greater detail in Section 1.2.5.

An alternative to the Nadaraya-Watson smoother, with bias properties which will later be exploited, was proposed by Gasser & Müller (1979) as

$$\hat{g}_{gm}(x) = \sum_j y_j \int_{t_{j-1}}^{t_j} \frac{1}{h} k\left(\frac{x-v}{h}\right) dv$$

where $t_0 = -\infty, t_n = \infty$, and $t_j = \frac{x_j + x_{j+1}}{2}$. This notation assumes that the data have been ordered in increasing value of the $x_j$'s. The values chosen for $t_0$ and $t_n$ ensure that the weights sum to 1. The drawback with these values is that a strong *boundary effect* is created, since near either end the tail point will receive a relatively large weight. Modifications are possible, however see Section 1.2.6 for a discussion on how boundary effects have been treated throughout this thesis.

A third estimator, which has been receiving much attention recently, is the class called *local polynomial* kernel estimators, and specifically, *local linear* smoothers. An early reference for this method, in the field of time series, is Macauley (1931). It was introduced as a regression estimator by Stone (1977), and is a member of the family of robust local regression estimators of Cleveland (1979). Two other significant references are Cleveland & Devlin (1988) and Fan (1993).

As the name suggests, a local polynomial kernel estimator works by estimating the regression function at a particular point by "locally" fitting an $n$-th degree polynomial. This local fitting is achieved by weighted least squares. In fact, the Nadaraya-Watson method, described above, is a special case of this class of estimators, as it corresponds to fitting a local polynomial of degree zero (i.e. a *local constant*). In this thesis, local line fitting is often focussed upon, as this method displays favourable asymptotic and boundary properties.

In the local linear case, the estimator is defined as $\hat{g}_{llf}(x) = \hat{\alpha}$ where $\hat{\alpha}$ and $\hat{\beta}$ minimise

$$\sum_{i=1}^{n} \{y_i - \alpha - \beta(x - x_i)\}^2 \, k\left(\frac{x - x_i}{h}\right). \tag{1.2.3}$$

Wand & Jones (1995) provide a good introduction to kernel smoothing techniques, and in particular the method of local line fitting. They demonstrate that, for local line fitting, the weighted least squares solution for formula (1.2.3) can be expressed as

$$\hat{g}_{llf}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\hat{s}_2(x;h) - \hat{s}_1(x;h)(x - x_i)\} \, k\left(\frac{x-x_i}{h}\right) y_i}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}$$

where

$$\hat{s}_r(x;h) = n^{-1} \sum_{i=1}^{n} (x_i - x)^r k\left(\frac{x - x_i}{h}\right).$$

Splines, although not considered in this thesis, offer an alternative to kernel methods, and a suitable reference for this approach is Green & Silverman (1994).

### 1.2.1 Smoothing parameter

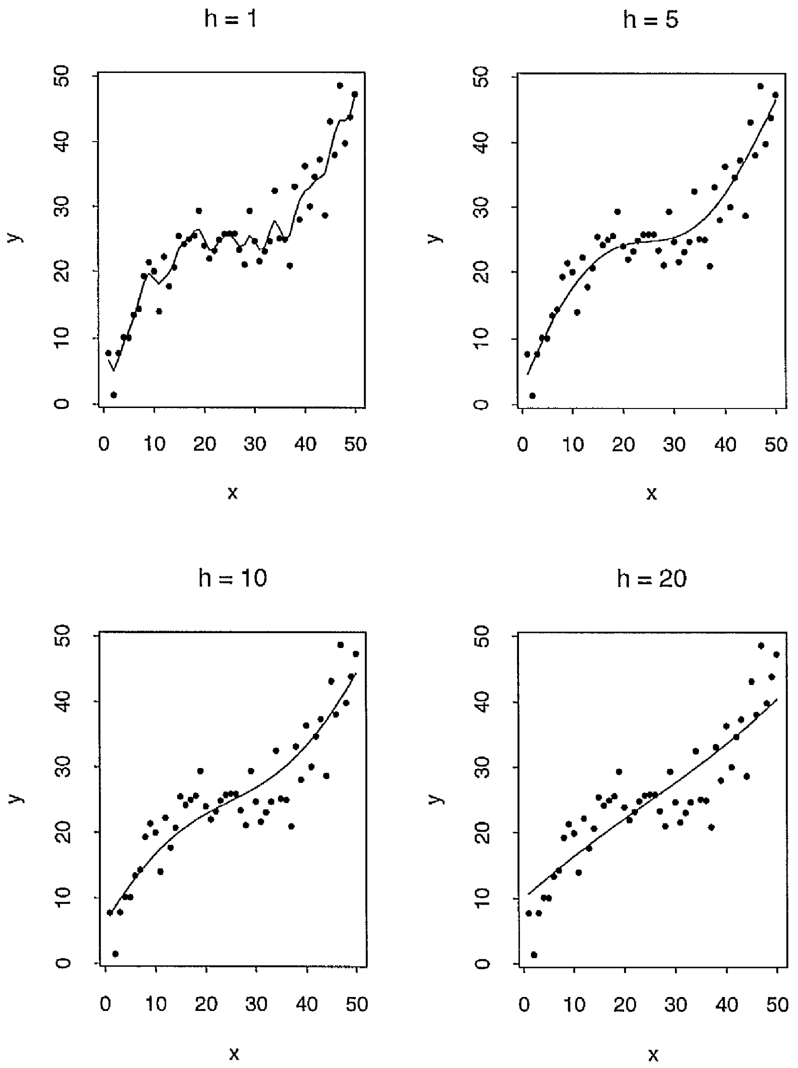
As outlined in the previous section, the choice of smoothing parameter, $h$, influences the fit of the nonparametric curve. This is illustrated in Figure 1.2.1, which shows four different local line smooths, corresponding to four different values of $h$, for some simulated data.

The plots clearly demonstrate the effect of the smoothing parameter. When $h = 1$, the fit is too rough, and follows the data too closely. In the second plot, for $h = 5$, the fit is improved considerably, and a smoother curve is formed. When $h$ is increased to 10, this seems to be too large, causing the data to be oversmoothed, and is only emphasized further when $h = 20$, which illustrates how the local line fitting method tends to a parametric least squares line as the smoothing parameter increases.

Much of the literature on nonparametric regression is concerned with choice of smoothing parameter. One of the aims of this thesis is to lessen the importance of choosing a smoothing parameter, at least with respect to inference. While choice of $h$ has a clear effect on how a smooth will appear visually, it can be demonstrated that in some circumstances it may not have a significant effect on inferential conclusions drawn from the data. This will be explored further in later chapters.

Since the area of smoothing parameter selection is very much a secondary topic for this thesis, it is not dwelt upon at length. Nevertheless, it is a key issue in general, and so a brief introduction is presented here to some of the methods which are commonly-used.

Figure 1.2.1: Illustration of the effect of the smoothing parameter



Härdle (1990) reports that the idea behind all smoothing parameter selection algorithms is to estimate the *Average Squared Error* (ASE) or an equivalent measure, in the hope that the value of $h$ minimizing this estimate, performs similarly for the ASE itself. In choosing a smoothing parameter, the aim is to balance the bias effects against the variance. As explained earlier, smaller values of $h$ lessen the bias but increase the variance, and vice versa. Section 1.2.5 discusses bias in more detail. The selection methods described in the following sections aim to facilitate the discovery of a value for $h$ which provides an appropriate amount of smoothing.

### 1.2.2 Choosing a smoothing parameter by cross-validation

The method of cross-validation is based on nonparametric regressions where, for each regression, a single observation is left out. Suppose the $j$-th observation is omitted, giving:

$$\hat{g}_j(x_j) = \sum_{i \neq j} w_i y_i$$

where the $w_i$'s are the weights, chosen by an appropriate kernel method and determined by a smoothing parameter, $h$. In other words, the curve is estimated at point $x_j$, without including in the calculation the corresponding observed value $y_j$. With this collection of modified estimates, the *cross-validation function* is formed:

$$CV(h) = \frac{1}{n} \sum_{j=1}^{n} \left[y_j - \hat{g}_j(x_j)\right]^2 w(x_j)$$

where $w$ is a weight function. The method acquired its name because it validates the ability to predict each $y_j$ $(j = 1, \ldots, n)$ across the sub-samples which exclude $(x_j, y_j)$.

The smoothing parameter is taken to be the value of $h$ which minimises $CV(h)$.

Cross-validation was first proposed in the context of kernel smoothing by Clark (1975), and Härdle (1990) provides greater detail and explanation of the method.

King, Hart & Wehrly (1991) observe that cross-validation is designed for a different purpose than maximising the power of a test. Values of $h$ that maximise power can differ from those which produce appealing visual estimates. This general principle, regarding the difference between inference and visual appeal, is also noted by Hall & Hart (1990), and is relevant, when considering the methods of inference which will be introduced in later chapters.

### 1.2.3 Plug-in methods of smoothing parameter selection

Recent research on the problem of smoothing parameter selection has found that the traditional methods, such as the cross-validation procedure described in Section 1.2.2, exhibit inferior asymptotic and practical performance. One reference for this is Härdle, Hall & Marron (1988).

Attention has switched to "plug-in" selection rules, which involve estimation of unknown functionals appearing in formulae for the asymptotically optimum value of $h$. These estimates are "plugged-in" to the asymptotic formula, and hence the name. Gasser, Kneip & Köhler (1991) report that these methods have much lower variability than cross-validation estimators for a variety of situations, and that, despite the asymptotic motivation, plug-in methods perform well even for small sample sizes.

Ruppert, Sheather & Wand (1995) propose several plug-in rules for local line fitting, under the assumption of homoscedastic errors. Their simplest proposal is a variant on the methods of Härdle & Marron (1995), and the most sophisticated is an adaptation from Sheather & Jones (1991). All plug-in methods require some initial estimates from the data, and Ruppert et al (1995) employ a variation of the "blocking method" of Härdle & Marron (1995), with the number of blocks determined by Mallows' $C_p$ (1973). Fast binning algorithms are used to provide simple implementation of the procedures.

Fan & Gijbels (1995) propose a selection method for local polynomial estimation which combines both plug-in and cross-validation ideas.

### 1.2.4 Degrees of freedom of a smoother

This is the method used by Hastie & Tibshirani (1990). Switching to vector-matrix notation, if $\mathbf{y}$ is the vector of responses, then the fitted values $\hat{\mathbf{y}}$ can be expressed as $\hat{\mathbf{y}} = \mathbf{S}_h\mathbf{y}$ for a particular smoothing parameter, $h$, and corresponding smoother matrix, $\mathbf{S}_h$. Hastie & Tibshirani (1990) define the degrees of freedom ($df$) of a smoother to be $df = tr(\mathbf{S})$.

This is interpreted as "the effective number of parameters" of a smoother, and can be used to make comparisons between different smoothers with respect to the amount of fitting done. Equally, a smoothing parameter can be selected simply by specifying the degrees of freedom required.

Hastie & Tibshirani (1990) report that two other definitions of degrees of freedom exist, namely $n - tr(2\mathbf{S} - \mathbf{S}\mathbf{S}^T)$ and $tr(\mathbf{S}\mathbf{S}^T)$. These are labelled *degrees of freedom for error* ($df_{err}$) and *degrees of freedom for variance* ($df_{var}$) respectively. All three definitions can be motivated by analogy with the Normal linear model, and are used for different purposes. This will be seen in Chapter 6, when $df_{err}$ is used in an estimator.

Figure 1.2.2 illustrates the relationship between smoothing parameter and degrees of freedom ($df$) for a simple example, and using local-line fitting. Here the explanatory values are simply 20 observations equally-spaced in the range $(0, 1)$. Note that the degree of freedom does not depend on the response variable, but only on the smoothing parameter and the explanatory. The smoothing parameter is generally the major factor in determining $df$, with the explanatory variable having little influence.

Figure 1.2.2 shows how $df$ varies smoothly with $h$, and it can be seen that $df$ decreases as $h$ increases, as expected. As seen in Section 1.2.1, the smaller the value of smoothing parameter, the rougher the fit, corresponding to a higher number of degrees of freedom (or parameters). Similarly, with a larger value of $h$, the fit tends towards a linear regression, and so $df$ approaches 2.

As an indication of how this works in practice, a simple example is shown in Figure 1.2.3, with some simulated data.

Here, it can be seen that 3 degrees of freedom ($h = 0.257$) are not quite sufficient to model the data, and 4 degrees of freedom ($h = 0.158$) is a better choice.

Similarly, the plots seen previously in Figure 1.2.1 represent 21, 5, 3 and 2.4

Figure 1.2.2: Illustration of the relationship between smoothing parameter and degrees of freedom

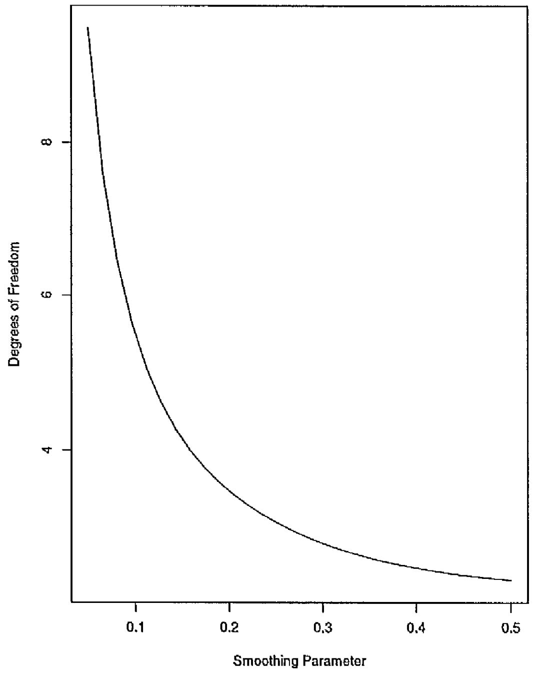
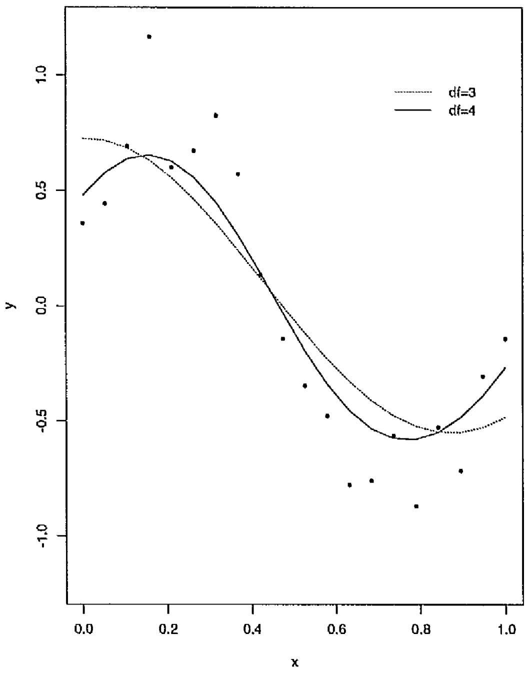


Figure 1.2.3: Example of smoothing with two choices of smoothing parameter



degrees of freedom respectively.

### 1.2.5 Bias

One drawback with nonparametric smoothing is that it naturally incurs bias, i.e. $E\{\hat{g}(x) - g(x)\} \neq 0$. However, the nature of the bias depends on the type of estimator chosen, and judicious choice of smoother can alleviate the problem in certain contexts, as will be seen throughout this thesis.

The Nadaraya-Watson (1964) estimator, described in Section 1.2, has asymptotic bias

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\}\{g''f + 2g'f'\}/f, \tag{1.2.4}$$

where $f$ is the density of the design points. Notice that when a standard Normal kernel is adopted, $\int t^2 k(t)dt = 1$. In the Gasser-Müller (1979) and local linear cases the asymptotic bias of the estimator is

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\} g''. \tag{1.2.5}$$

This has the considerable advantage of being independent of the design density, and Fan (1992) refers to estimators of this sort as *design-adaptive*. It is this property which is exploited in later chapters, for example when curves for different groups are contrasted. With design-independence, the bias terms cancel under a null hypothesis of equality of the curves, since the bias is dependent only upon the second derivative of the true function. It is clear that this is not the case if the Nadaraya-Watson estimator is used, unless the underlying densities of the design points are the same.

The local linear smoother has the additional appealing property that it is *exactly* unbiased when the true regression function is linear.

Formulae (1.2.4) and (1.2.5) above are given by Härdle (1990), Chu & Marron (1991), and Wand & Jones (1995).

### 1.2.6 Boundary effects

Special consideration has to be given to the performance of kernel smoothers near the boundary of the covariate space. This is because, as the boundary is approached, part of the span of the kernel is devoid of data. This has implications for the bias of a kernel smoother. Wand & Jones (1995) give a good explanation of behaviour near the boundary, and show that the boundary bias for the local linear estimator is of the same form as for the interior of the covariate space, except for the kernel dependent constant. This means that the property of design-independent bias also holds near the boundary. In all the examples considered throughout this thesis, no adjustment has been made for boundary effects.

# Chapter 2

# Three-dimensional plots of data and models

## 2.1 Introduction

Graphical methods play a crucial role in the initial exploration and presentation of data, and in the common case where interest is in the relationship between two variables the scatterplot is an invaluable tool. Additional information about a third variable can be added to scatterplots in a limited way through the use of coding, by colour or symbol. Some software packages allow the creation of scatterplots of three-dimensional data in the form of spinning point clouds. MacSpin, reviewed by Hinde (1988), is one such example, but many other packages now provide similar functions. Such plots can be very instructive, but they are dependent on animation to create the three-dimensional effect, making paper representation impossible. Huber (1987) stresses the importance of dynamic interaction with such plots. An alternative approach is to use stereographic projections, involving two two-dimensional projections with slightly shifted perspectives which, with training or with a set of special spectacles, can recreate the visual perception of the third dimension.

The aim of this chapter is to discuss a simpler approach to the representation of three-dimensional data, producing static displays which require only a graphics device with three colours or shadings. Simple geometric projection has long been used to produce two-dimensional images of three-dimensional objects. This is the basis of much engineering and "visualization" software, whose representation of objects has become very sophisticated and visually stunning. AVS is one such system, described by Upson et al (1989). However, much simpler displays of surfaces and objects in a plot, corresponding to an underlying model, rather than simply the points which correspond to data, create a helpful visual frame of reference. This frame of reference is meaningful because it refers to an underlying model.

In Section 2.2 the simple geometry governing projections is outlined. The visual representation of a variety of three-dimensional structures is then explored. The initial examples all consist of parametric models, in the form of a plane, and are applied in such contexts as regression, principal component analysis, survival data and discrimination.

The model need not be parametric, however. Nonparametric methods can also be used to identify smooth model structures, and this is addressed in Section 2.4.3, where contours of a three-dimensional density estimate are considered. This section introduces the topic of nonparametric modelling, which is the main theme of this thesis, and which is explored more fully in later chapters.

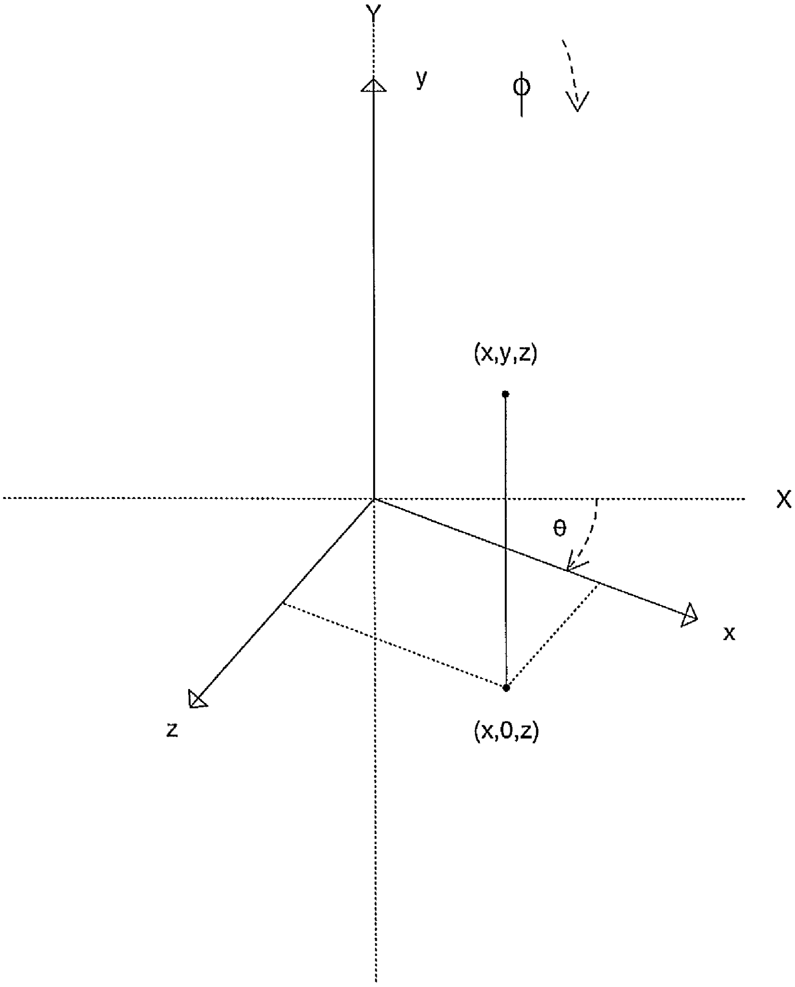
## 2.2 The geometry of three-dimensional projections

The formulae which govern the representation of a three-dimensional object by projection onto a two-dimensional surface can be derived by simple trigonometry. Texts such as Angell (1981) give a thorough discussion of this general topic. Figure 2.2.1 illustrates the basic idea. Here $X$ and $Y$ refer to co-ordinates on the projection surface, such as paper or a computer screen. The axes $x$, $y$ and $z$ refer to the co-ordinates of the data or object to be plotted. Computing environments such as XLISPSTAT (Tierney (1990)) allow three-dimensional point clouds, or "wire-frame" objects, to be rotated about three axes, namely the screen axes $X$ and $Y$, and an axis perpendicular to the $X$-$Y$ plane.

In this chapter a more restricted approach will be adopted, where rotation can take place only about the screen co-ordinate $X$ or the data co-ordinate $y$. This does not restrict the projections which can be viewed but it does have the consequence that lines which are parallel to the $y$-axis will always be represented by vertical lines in the projection. This is particularly helpful in displaying regression data, where the response variable can be associated with the $y$-axis, ensuring that errors and residuals appear as vertical lines, in a manner which is consistent with two-dimensional scatterplots.

Figure 2.2.1: Illustration of perspective geometry

x,y,z, are three-dimensional co-ordinates; X,Y are screen co-ordinates
Phi and theta are the angles of rotation about the X-axis and y-axis respectively



If the rotation about $X$ is denoted by $\phi$ and rotation about $y$ by $\theta$ (where $\phi$ and $\theta$ are chosen by the user), then simple trigonometry shows that a point $(x, y, z)$ in three-dimensional space has screen co-ordinates

$$\begin{aligned} X &= xcos(\theta) - zsin(\theta), \\ Y &= ycos(\phi) - (xsin(\theta) + zcos(\theta))sin(\phi). \end{aligned}$$

This is illustrated in Figure 2.2.1, which also displays the fact that the line $(x, y, z) - (x, 0, z)$, which is parallel to the $y$-axis, has a vertical representation. These basic formulae control the representation of all points, lines and objects. Problems caused by different units on the $x$, $y$ and $z$ axes can be dealt with simply by a preliminary scaling of each to the range $(-1, 1)$.

There is a variety of methods for enhancing the visual perception of projections of this kind. The formulae may be amended to reduce the size of objects appearing at the rear of the three-dimensional plot. Other methods of "depth cueing" include giving points at the front of the plot greater intensity (of brightness or of ink) than points at the rear. Realism can also be increased by mimicking the effects of light shining from different positions, as illustrated by Cleveland (1993, p.269). In this chapter, it is argued that the representation of a model provides a reference which itself greatly assists three-dimensional perception. No other methods of enhancement have been used. The simplicity of the formulae mean that figures can be produced in a wide variety of statistical computing environments, with simple point plotting, line drawing and colour filling facilities, although in one case grey-level shading is required. All of the pictures described in this chapter were produced in the S-Plus environment (Becker et al, 1988).

## 2.3 Regression data

### 2.3.1 Displays with two covariates

Regression is one of the most commonly used of all statistical procedures, and is the main focus of attention in this thesis. Three-dimensional representation allows the simultaneous effects of two continuous covariates on a response variable to be explored. Table 2.3.1 displays data on the amount of giving, in Pounds sterling per annum per member, in the dioceses of the Church of England from reports published in 1983. Potential explanatory variables are the employment rate of the diocese (*Employment*), the percentage of the population which appears on the electoral roll of the Church (*Elect*), and the percentage of the population which usually attends on Sundays (*Attendance*). An economic analysis of these data, with a wider range of covariate values, was carried out by Pickering (1985). Bowman & Robinson (1990) examine a subset of the data and describe graphical software which uses three-dimensional representation to illustrate the meaning of a regression model with two covariates.

Figure 2.3.2 displays a scatterplot matrix for these data. Fitted regression lines of *Giving* on each of the explanatory variables *Employment*, *Elect* and *Attend* are superimposed on the appropriate panels of the plot. *Giving* shows a gentle, but non-significant, increase with *Employment*. More surprisingly, *Giving* decreases markedly,

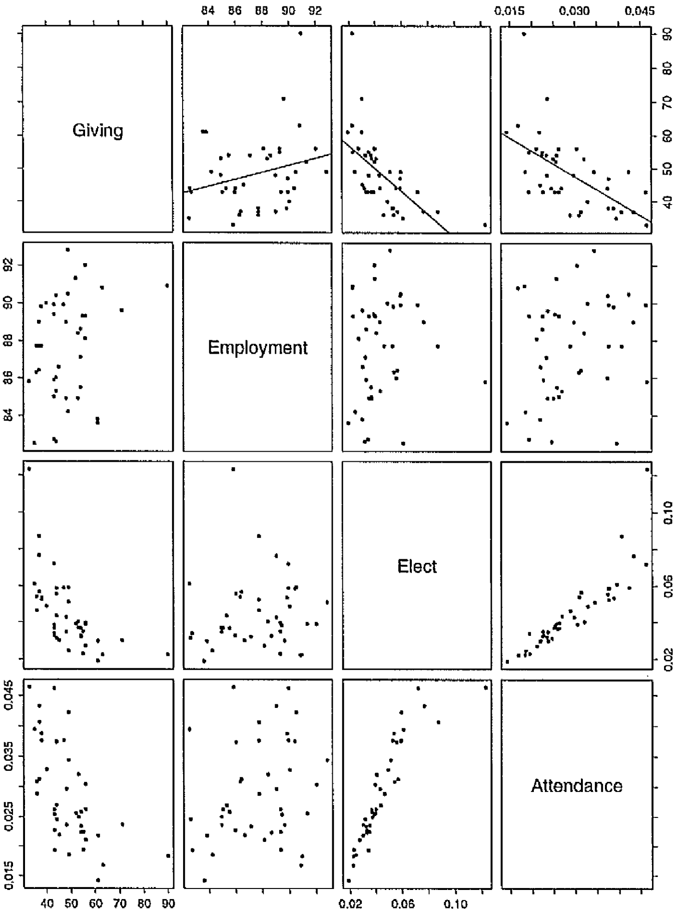
Table 2.3.1: Data on giving in the Church of England, reported in 1983. Each row refers to a different diocese.

| Giving | Employment (%) | Electoral roll (%) | Attendance (%) |
|---|---|---|---|
| 43 | 89.9 | 7.2 | 4.6 |
| 61 | 83.6 | 1.9 | 1.4 |
| 37 | 86.4 | 5.7 | 3.1 |
| 54 | 87.1 | 3.2 | 2.3 |
| 71 | 89.6 | 3.0 | 2.4 |
| 48 | 89.0 | 4.3 | 3.0 |
| 37 | 87.7 | 8.7 | 4.1 |
| 55 | 89.3 | 2.3 | 1.9 |
| 44 | 85.3 | 4.4 | 2.7 |
| 44 | 90.4 | 5.9 | 3.7 |
| 43 | 85.0 | 3.7 | 2.6 |
| 54 | 88.6 | 3.2 | 2.2 |
| 43 | 82.7 | 3.4 | 1.9 |
| 38 | 89.8 | 5.3 | 3.9 |
| 44 | 86.0 | 5.6 | 3.7 |
| 49 | 90.5 | 5.9 | 4.2 |
| 49 | 92.8 | 5.1 | 3.4 |
| 33 | 85.8 | 12.3 | 4.6 |
| 55 | 89.3 | 3.5 | 2.2 |
| 48 | 84.9 | 3.5 | 2.4 |
| 36 | 86.3 | 5.4 | 3.1 |
| 44 | 82.6 | 3.1 | 2.5 |
| 63 | 90.8 | 2.2 | 1.7 |
| 43 | 85.9 | 3.3 | 2.3 |
| 61 | 83.8 | 3.0 | 2.2 |
| 38 | 87.7 | 5.3 | 3.8 |
| 56 | 92.0 | 3.9 | 3.0 |
| 53 | 88.4 | 4.0 | 3.2 |
| 43 | 89.4 | 3.8 | 2.5 |
| 36 | 87.7 | 4.6 | 2.9 |
| 56 | 89.3 | 4.0 | 2.6 |
| 52 | 91.3 | 3.9 | 2.6 |
| 47 | 89.9 | 5.9 | 3.8 |
| 37 | 89.0 | 7.6 | 4.3 |
| 49 | 84.2 | 2.4 | 1.8 |
| 90 | 90.9 | 2.2 | 1.8 |
| 56 | 88.1 | 2.7 | 2.1 |
| 35 | 82.5 | 6.1 | 3.9 |
| 45 | 86.6 | 3.0 | 2.2 |
| 40 | 90.0 | 4.9 | 3.3 |
| 54 | 85.5 | 3.7 | 2.6 |
| 53 | 84.9 | 3.7 | 2.5 |

and significantly, with *Elect* and *Attendance.* One possible explanation is that in large congregations there can be less perceived need to give, since a larger pool of potential contributors is apparent. Alternatively, larger congregations may tend to include higher proportions of less committed members, with a consequent drop in the giving averaged over the membership.

Since *Elect* and *Attendance* are highly correlated, a reasonable initial model to fit to the data would include *Employment* and *Attendance* as explanatory variables. If

Figure 2.3.2: Scatterplots of Church of England data



this is done then *both* of these variables have significant effects on *Giving*. This is one of those cases where a variable, in this case *Employment*, becomes significant only when another variable is included in the model.

In order to understand what is happening, a three-dimensional representation of the data and model is helpful. Rather than display the data simply as points, and depend on dynamic rotation for the perception of the three-dimensional structure, the fitted regression plane has been added in Figure 2.3.3 as a shaded region. This gives a helpful reference for the location of each data point, whose residual can now also be displayed as a line segment. Observations which lie below the plane are marked with a $-$ rather than a $+$, and dotted rather than full line segments have been used. To enhance the three-dimensional nature of the plot a cube has been added to denote the plotting space. The end result is a static figure which gives a clear representation of the three-dimensional structure of the data and the fitted model.

From Figure 2.3.3, the positive effect of *Employment* and the negative effect of *Attendance* can be seen, represented by the edges where the plane intersects the faces of the cube. It can also be seen that the opposite nature of these effects will cause each to dilute the other when the covariates are fitted separately. If the data are viewed through the face of the cube defined by *Giving* and *Employment* then any regression effect across that face is blurred by the regression effect in the opposite direction due to *Attendance*. This effect can be explored further by plots which display a plane sloping only in the direction of *Employment*, or only in the direction of *Attendance*.

Further insight can be gained by plotting *Giving* against *Elect* and *Attendance*, as seen in Figure 2.3.4. The presence of collinearity is indicated by the fact that the data all lie around a diagonal line across the fitted plane. The effect of collinearity is apparent through imagined rotation of the plane about this diagonal line. This corresponds to large changes in the regression coefficients associated with *Elect* and *Attendance*, while the residuals, and their sum-of-squares, change very little.

Figure 2.3.3: Three dimensional plot of the Church of England data

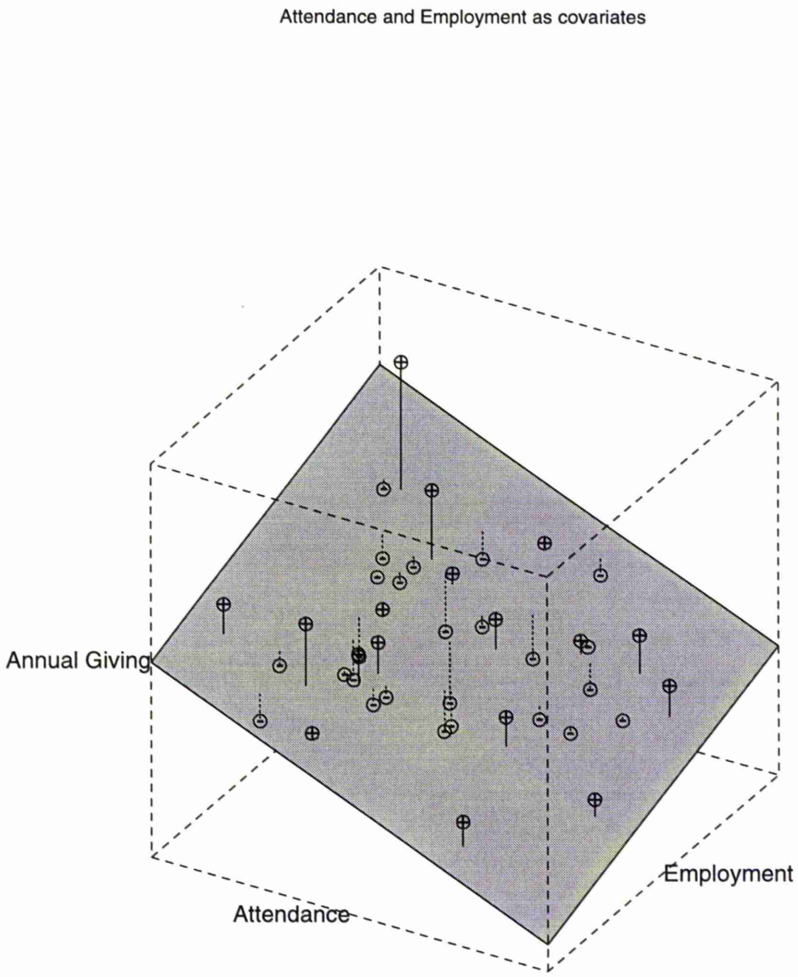
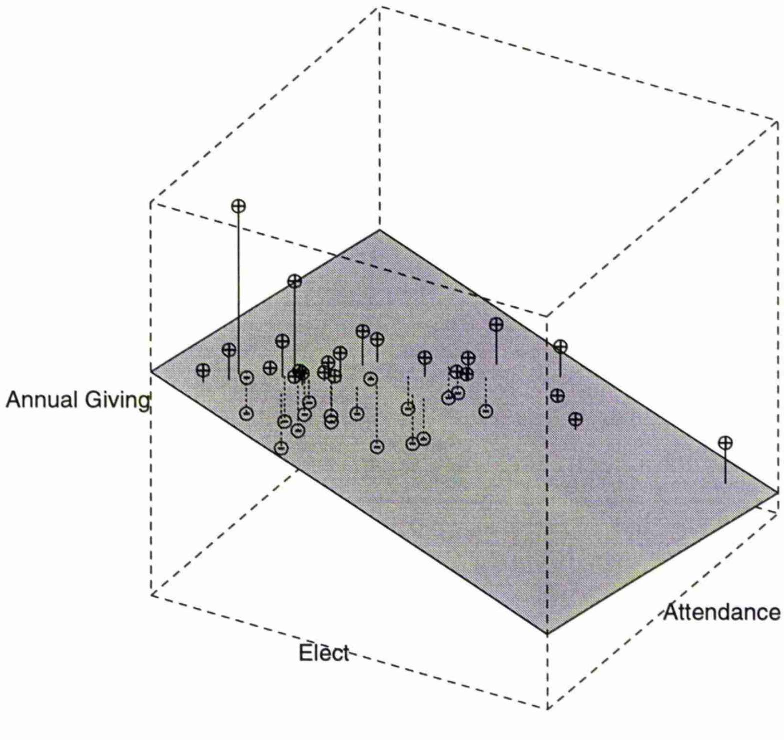


Figure 2.3.4: Church of England data with Elect and Attendance as covariates



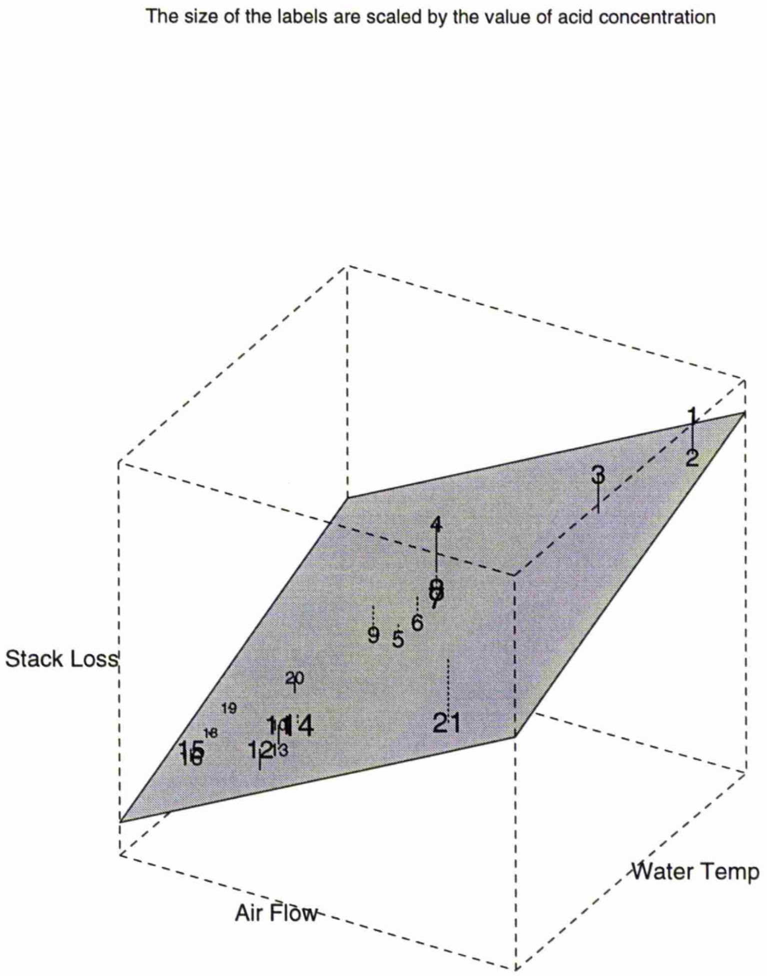
These kinds of illustration are invaluable in the context of teaching, where the nature of a multiple regression model can be communicated in a graphical manner, and an intuitive understanding of effects such as collinearity provided. Current computing environments allow them to be used in a more general fashion as routine exploratory tools, rather than as carefully and expensively produced book illustrations. Further examples of the potential insights available from this kind of plot are described in later sections.

### 2.3.2 Adding information on a third explanatory variable

A standard example on the use of diagnostics and exploratory techniques in multiple regression is the stack loss data of Brownlee (1965). The data refer to 21 successive days of operation of a plant oxidizing ammonia into nitric acid. The response is the percentage of ingoing ammonia which is lost as unabsorbed nitric acid. There are three explanatory variables, namely the flow of air to the plant, the temperature of the water entering the countercurrent nitric oxide absorption tower, and the concentration of nitric acid in the absorbing liquid. These data have been discussed by many authors. They offer an opportunity to explore further the effectiveness of three-dimensional representations.

Figure 2.3.5 displays the data with air flow and water temperature as covariates. As in two-dimensional scatterplots, information on an additional variable can be added through the use of colour or labelling. Here, the value of acid concentration is indicated for each observation by the size of its printed label (1,...,21). This

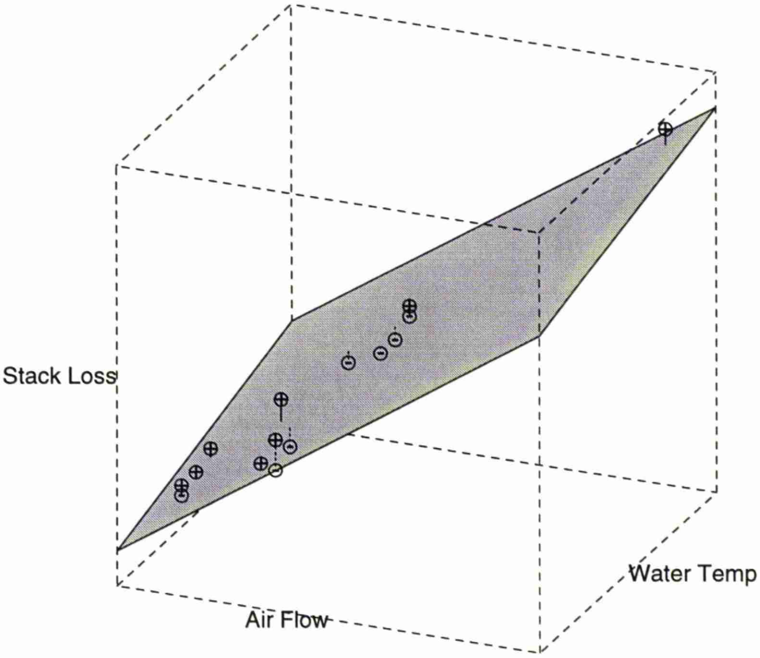
Figure 2.3.5: Regression with the stack loss data



variable has not been used in fitting the model plane. Several features, already identified by a number of authors, are apparent. First of all, large residuals can be seen for observations 1, 3, 4 and 21. Daniel & Wood (1980) concluded that these points represent transitional states and should be dropped. These authors also identified three highly clustered groups of points, (5,6,7,8), (10,11,12,13,14) and (15,16,17,18,19), which are apparent in the three-dimensional representation. There is no obvious effect of acid concentration since large and small values are dispersed throughout the plot. Finally, if points (1,3,4,21) are removed there is a suggestion of a systematic pattern in the residuals. This is illustrated in Figure 2.3.6, which shows the regression plane for the covariates air flow and water temperature (acid concentration has been omitted). Daniel & Wood concluded that a quadratic term in air flow is necessary, and this can be seen from the plot. This feature is investigated more fully in Chapter 3, when a graphical method is used to identify the curvature, and again in Chapter 6, where a formal nonparametric test is applied.

In this well known example all these effects were originally identified by a variety of methods. The three-dimensional representation has the merit of illustrating many of these effects in a single figure. It must always be emphasized that these graphical displays are not *replacements* for formal tests, but rather useful visualization tools to aid understanding and interpretation.

Figure 2.3.6: Stack loss data with the transitional state points removed



### 2.3.3 Displaying survival data

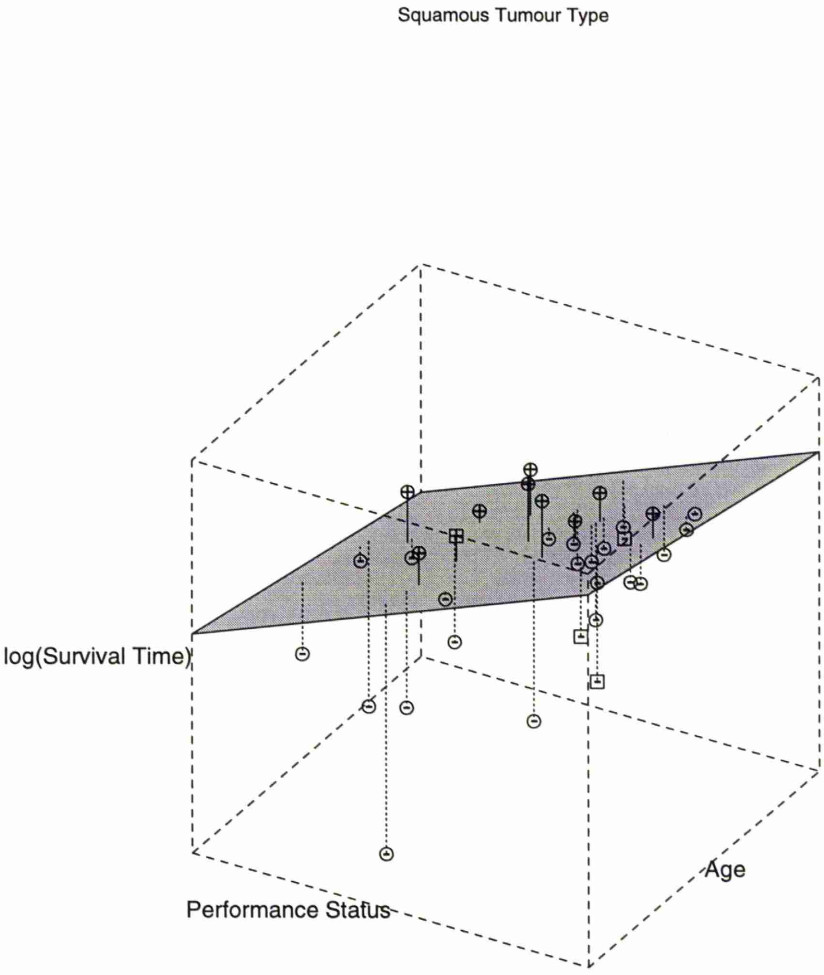
Another context in which it is natural to add information to a scatterplot is where the response variable is a survival time, with some observations subject to censoring. Figure 2.3.7 displays data corresponding to 35 patients who have a squamous tumour in the Veterans' Administration lung cancer trial, provided by Prentice (1973), with performance status and age as covariates. There are only a few censored observations and these are represented by squares instead of circles. With this type of data, the shape displayed by the raw observations can be difficult to interpret, especially if there is a high degree of censoring. The accelerated failure time model (Kalbfleisch & Prentice (1980)), where log survival is represented as a linear function of covariates, provides a convenient reference. The Cox (1972) proportional hazards model is more widely used in this context, but it cannot be represented graphically in such a simple fashion. The accelerated failure time model differs only in that the baseline hazard is assumed to be of a specific parametric form. It is likely to identify the principal covariate effects, unless the parametric assumption is markedly inappropriate. The model takes the following form

$$Y = \alpha + \boldsymbol{\beta}^T \mathbf{z} + \sigma W,$$

where $Y$ is the natural logarithm of the survival time, $\alpha$ is an intercept parameter, $\boldsymbol{\beta}$ is a vector of coefficients, $\mathbf{z}$ is a vector of covariates, $\sigma$ is a scale parameter, and $W$ is a random variable, taken to be Weibull in this example.

The positive association between performance status and survival, and the absence

Figure 2.3.7: Plot of the VA lung cancer data with fitted survival model



of a strong age effect, are apparent in the orientations of the edges of the fitted plane.

## 2.4 More general three-dimensional data

### 2.4.1 Discrimination reference planes

On some occasions it is useful to compare different groups of three-dimensional data. Discrimination problems are one example of this. Figure 2.4.8 displays data on Conn's syndrome from two groups of patients whose medical symptoms are similar but whose causal conditions are different. A variety of blood measurements, and the age of the patients, were recorded in an attempt to distinguish between the two groups. These data are given in Aitchison & Dunsmore (1975). A natural reference object in this case is the linear discriminant plane. The group labels on the observations in Figure 2.4.8 show that the two groups are well separated. The small number of incorrectly classified cases are marked by squares rather than by circles. New observations whose group membership is unknown could be added to the plot with a different label, for the purpose of diagnosis. This is shown in Figure 2.4.9, where a sample future case has been added to the plot, and labelled with the letter "U". It can be seen that the new patient lies with the Hyperplasia group, and would be diagnosed as such.

In these discrimination examples, the projections of the points onto the reference plane have again been represented by lines. It is natural for visual appearance to use a perpendicular projection, although this does not correspond to the Mahalanobis distance measure which underlies the discrimination procedure. However, it does

Figure 2.4.8: Discrimination with the Conn's Syndrome data

Group 1 is Adenoma, and group 2 is Hyperplasia

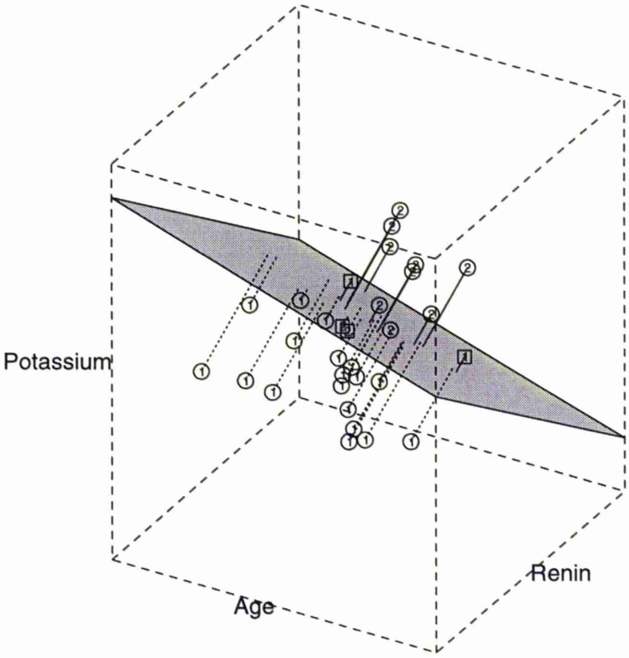
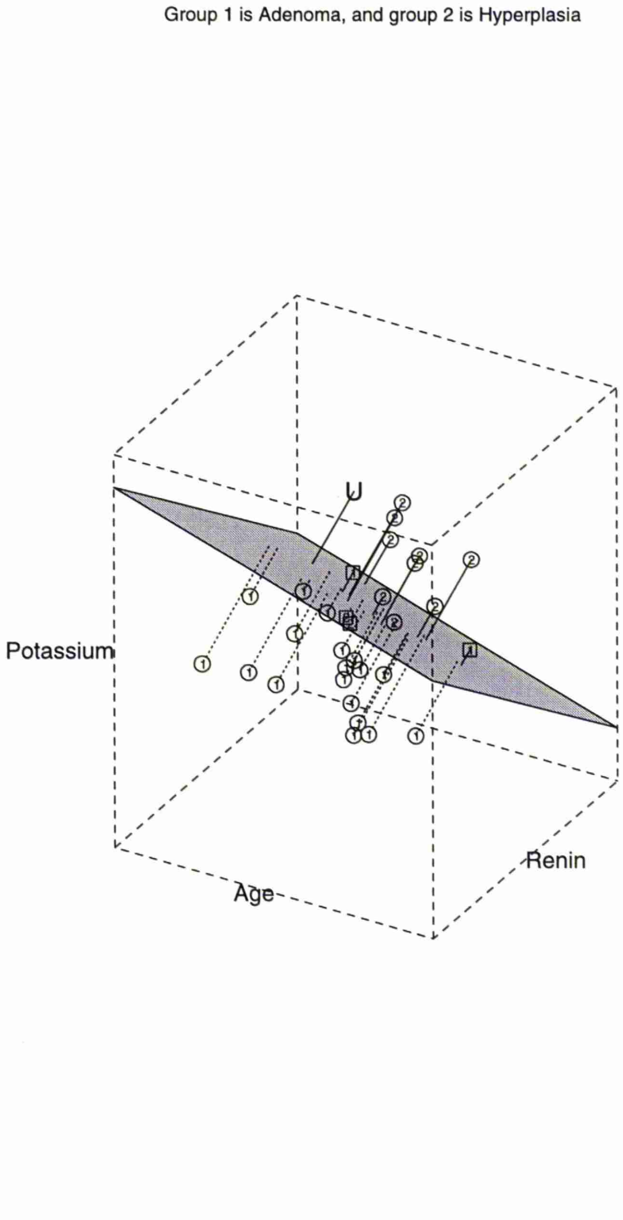


Figure 2.4.9: Discrimination with a new observation added for diagnosis



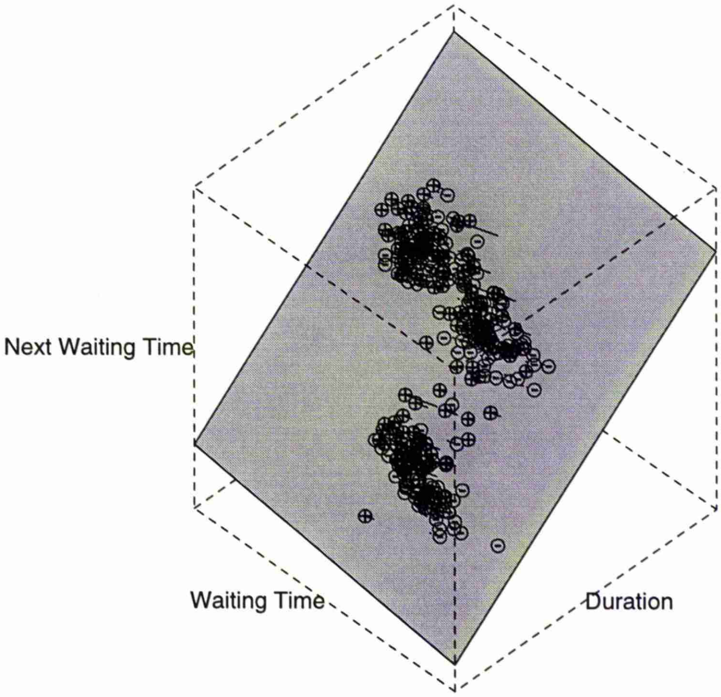
correctly indicate the side of the plane on which each observation lies. In order to produce lines which are visually perpendicular care must be taken to ensure that the same physical distance corresponds to the same unit distance along each screen axis. This has the consequence that the plotting region can now take the form of a cuboid rather than a cube.

### 2.4.2 Principal component reference planes

In order to represent general trivariate data without a regression structure, a natural reference is provided by the plane defined by the first two principal components of the three variables concerned. Figure 2.4.10 displays data from the Old Faithful Geyser, described by Azzalini & Bowman (1990). This consists of two interwoven time series describing the durations of the eruptions and the waiting times between eruptions of the geyser. Azzalini & Bowman (1990) identified the relationships between the three variables *waiting time*, *duration*, and *subsequent waiting time* to be important in determining the structure of the series. Clusters in the joint distribution of these variables are apparent by deduction from the marginal scatterplots. Figure 2.4.10 gives a direct representation of the relationships between the three variables, with the clusters clearly seen.

In the absence of a regression structure, the feature that "residuals" from the model should be represented as vertical lines also disappears. The deviations of the observations from the plane of the first two principal components are represented by lines which are perpendicular to the plane.

Figure 2.4.10: Geyser data with the plane of the first two principal components



It is also possible to produce a three-dimensional biplot representation of data matrices and Gower (1990) derives the form of this.

### 2.4.3 Density contours

The geyser data differs from previous examples in the large number of observations present in the dataset. Although the broad structure of the data is apparent from Figure 2.4.10, overplotting can sometimes obscure patterns. This is also true in two-dimensional scatterplots. Bowman & Foster (1993) addressed this problem by drawing contours of an estimate of the underlying density function, and other authors, such as Scott (1992), explain the need for bivariate and multivariate density estimation. Bowman & Foster (1993) carefully choose contours which contain specified proportions of the observations, in a manner reminiscent of a boxplot. This draws attention to features of the data which are otherwise difficult to see.

The same principle can be applied with three-dimensional data $(x_i, y_i, z_i)$, $i = 1, \ldots, n$. In a simplification of the nonparametric regression technique in Chapter 1, a density estimate is available through the kernel method as

$$\hat{f}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^3 s_x s_y s_z} k\left(\frac{x - x_i}{h s_x}\right) k\left(\frac{y - y_i}{h s_y}\right) k\left(\frac{z - z_i}{h s_z}\right) \qquad (2.4.1)$$

where $s_x, s_y, s_z$ denote the sample standard deviations of each variable and $k$ denotes the standard Normal density function. Silverman (1986), and Wand & Jones (1995) give excellent introductions to density estimation. Formula (2.4.1) is a particularly simple form, using an uncorrelated trivariate Normal density function as the kernel, and using a single smoothing parameter $h$, scaled by the sample standard deviation for each dimension. It is necessary to choose a value for the smoothing parameter. As discussed in Bowman & Foster (1993), a smoothing parameter optimal for a Normal distribution is a simple and often effective choice in this setting, especially when the sample size is large. For the estimator defined above this is

$$h = \left\{ \frac{4}{5n} \right\}^{\frac{1}{7}} .$$

For the geyser data this gives a value of $h$ equal to 0.4292. This choice, of Normal optimal smoothing parameter, is conservative in the sense that it is likely to oversmooth the data. It is therefore less likely to give prominence to spurious features in the data. In addition, it is *relative* as opposed to *absolute* density heights which are important in the construction of contours, and this lessens to some degree the influence of the smoothing parameter. This is a common theme in this thesis, as will be seen in later chapters, where nonparametric graphical methods and formal tests are presented, in which the choice of smoothing parameter is less important than usual for this subject. It is demonstrated that the ultimate conclusions which can be drawn remain unchanged for a wide range of smoothing parameter values.

Returning to the current example, in two dimensions, a contour of a density estimate is a closed curve, or a set of closed curves if multimodality is present. A contour of a function defined in terms of three arguments is a more unusual object. In fact, it is a closed surface, or set of closed surfaces if multimodality is present. Scott (1992) describes this approach and explores its use on a variety of datasets.

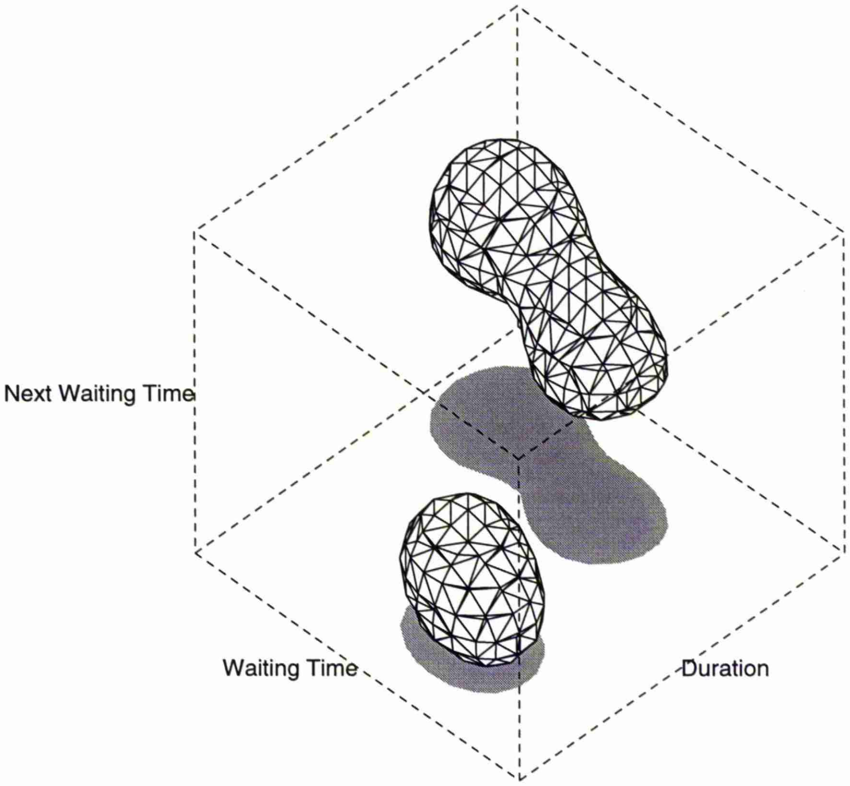
Figure 2.4.11 displays a contour of a three-dimensional density estimate of the geyser data. It is represented as a "wire-frame" object, constructed in a manner similar to that described by Scott (1992). This contour incorporates the feature that it contains exactly 75 percent of the observations.

There are three steps involved in constructing a contour of this type. First of all, the density estimate is evaluated at each observation, using formula 2.4.1. These estimates are then ranked, and the lower quartile, $Q_1$, found. To produce the plot, it is necessary to find values of $(x, y, z)$ which give

$$\hat{f}(x, y, z) = Q_1.$$

All such points form the surface of the density contour. These values are obtained by working with a cube of equally-spaced $(x, y, z)$ points, and calculating the density estimate at each point, again using formula 2.4.1. Then, for every pair of adjacent points, the corresponding density estimates are compared. If both density estimates are less than $Q_1$, then those points in the cube lie outside of the contour. Likewise, if both density estimates are greater than $Q_1$, then both points are enclosed by the contour surface. If either, or both, of the density estimates are equal to $Q_1$, then the corresponding point(s) lie(s) *on* the surface of the contour. Finally, if one density estimate is greater than $Q_1$, and the other less than $Q_1$, then the contour surface passes between these points, with the first point contained within the contour, and the second one excluded. In this case, linear interpolation is used to determine the actual $(x, y, z)$ values required to give $\hat{f}(x, y, z) = Q_1$.

Figure 2.4.11: The 75% contour of a three-dimensional density estimate of the geyser data



The number of points in the equally-spaced cube is flexible, and controls the appearance of the final plot. A small number of points produces a very coarse plot, since the surface is necessarily formed by relatively few points. The advantage of this approach however, is that it is very quick, and can be used to give a rapid initial impression of the structure of a data set. If a larger number of points is used, then a much finer plot can be produced, as the points are closer together, and more surface co-ordinates can be determined. This procedure does take slightly longer, but with the speed of computing packages now available, this is not a problem. It has been found that a grid size of around 25 is adequate for most applications.

The final stage involves taking all the points observed as lying on the contour surface, and converting them into a form suitable for plotting. This is achieved by finding sets of three $(x, y, z)$ points which are adjacent, and thus form a triangular region on the surface of the contour. When all of these polygons are plotted, the density contour is formed.

The space contained by the contour therefore corresponds to the upper reaches of the density estimate. This focusses attention on the principal features of the density and ignores the tail behaviour.

The density contour of the geyser data is disjoint, confirming the highly clustered nature of the data. In this example, a 75% contour was chosen, although of course any value can be selected. If a smaller percentage is specified for the fraction of data contained, three disjoint surfaces are displayed, corresponding to the three clusters in the data. This can be seen in Figure 2.4.12, where the 50% contour is shown. The 75% contour displayed in Figure 2.4.11 draws attention to the fact that the two upper clusters are not as well separated as the lower one is from the other two, and this feature is also clear in Figure 2.4.12.

When surfaces are disjoint the relative locations of the separate parts in three-dimensions may not always be apparent. This has been resolved in Figures 2.4.11 and 2.4.12 by displaying the shadow of the surfaces on the floor of the plot, as if a light were shining from the top face of the surrounding cube.

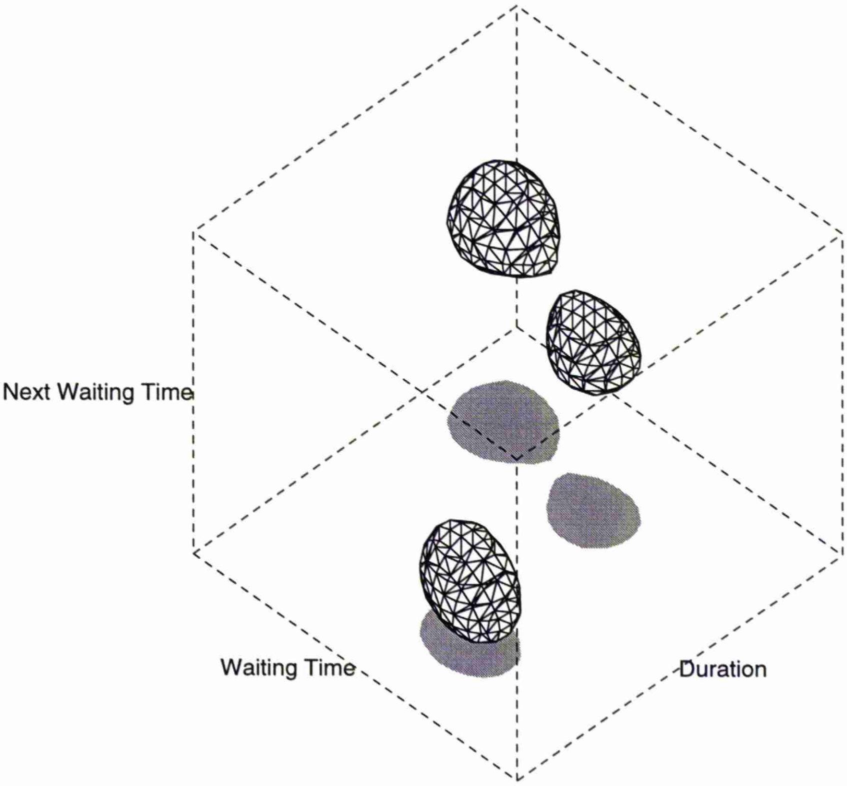
Scott (1992) describes a variety of more sophisticated techniques for constructing and displaying density contours, including the use of light sources to enhance visual perception, and removal of strips of the surface in order to display interior contours.

In this type of plot the data are not directly represented as points. It is the model alone, in this case a nonparametric one, which is being displayed in order to highlight the principal features of the data.

## 2.5 Discussion

A variety of plots involving three-dimensional representation have been described and illustrated on different datasets. It has been shown that static representations can be very effective for small sets of data when the display includes an appropriate model as a reference object. When there is a large amount of data, smoothing techniques can sometimes be used to focus attention on the main features of the data through

Figure 2.4.12: The 50% contour of a three-dimensional density estimate of the geyser data



a nonparametric model. The level of graphical facilities required to produce these plots is not high in view of the type of computing software and hardware now widely available.

This chapter serves a dual purpose, with regard to the rest of the thesis. First of all, it illustrates the importance of a helpful and informative plot, and later chapters present a number of graphical methods for displaying data in particular situations. Secondly, the idea of nonparametric modelling was introduced, and this is explored much more fully in the chapters to follow, in the regression context. The important area of inference is also addressed. The combination of these elements is the thrust of this thesis; namely the visual impact of an appropriate plot, and the back-up of a formal test. The influence of the smoothing parameter is another important consideration.

# Chapter 3

# Graphical comparison of nonparametric curves

## 3.1 Introduction

As a graphical method, the drawing of curves and lines to represent data or a model is ubiquitous. As explained in Chapter 1, in many cases it can be advantageous to model curves nonparametrically since this approach can extract from the raw data an indication of underlying structure without assuming any particular parametric form.

The aim of this chapter is to introduce graphical methods of focussing on the differences between nonparametric curves. The situation arises in an analysis of covariance context, where two or more groups are involved. The comparison of groups in the nonparametric context is complicated by the presence of bias, which has to be taken into consideration.

Here, the comparison is done by creating reference bands derived from the standard error of the difference between curves at each point of interest. It can be thought of as a graphical representation of a $t$-test of equality at each point. In the examples considered, nonparametric kernel smoothing, as defined in Section 1.2 of Chapter 1, is used to estimate the curves. In each case an appropriate reference band corresponding to equality, or in some cases to additivity, is constructed and shown to provide a helpful means of graphical exploration. It is also shown that, with careful choice of smoother, the problem of bias can be eliminated, at least asymptotically.

As well as the usual continuous-response regression context, examples are also considered for binary-response models and survival data (where the Kaplan-Meier estimator is employed). Lastly, the comparison of parametric and nonparametric curves is considered in Section 3.4.

## 3.2 Nonparametric regression curves

In an analysis of covariance experiment, where a regression relationship is investigated across two or more groups, the principal interest is generally in comparing the groups. In addition to a straight test of equality (or parallelism), a useful follow-up would be to identify where any differences might lie. When modelling the regression relationship, assuming a linear, parametric form can often be restrictive, and instead, a more flexible nonparametric relationship might be preferable.

This chapter looks at graphical methods for investigating the differences between nonparametric curves. This does not diminish the need for formal methods of inference, and this is addressed in Chapter 4.

The nonparametric curves are compared by means of a "reference band", a graphical technique based on the idea of a $t$-test at each point of interest. Multiple comparisons mean this is, of course, an inappropriate method of primary analysis, but it can be very useful in interpreting data, particularly when used in conjunction with the tests of Chapter 4.

As seen in Chapter 1, in nonparametric regression an underlying relationship $y = g(x) + \varepsilon$ is estimated by smoothing observed data $(x_j, y_j)$, $j = 1, \ldots, n$, with perhaps the simplest kernel smoother being the Nadaraya-Watson (1964) method. Two alternatives, also introduced earlier, are the Gasser & Müller (1979) smoother, and the method of local line fitting. As seen in Section 1.2.5, the latter two methods have certain useful bias properties.

In all of the kernel smoothing methods, the estimator has the form

$$\hat{g}(x) = \sum_{j=1}^{n} w_j y_j$$

where the weights $w_j$ depend on $x$, the $x_j$'s, and a smoothing parameter $h$.

Considerable attention has been devoted to the problem of constructing confidence bands for the true curve $g$. Härdle (1990) constructs pointwise confidence intervals for a curve using estimates of the true curve and the design density, and also by a bootstrap method, but concedes that the bias of the smoother is not taken into account. A third method, involving a "wild bootstrap" (Härdle & Mammen (1993)), is considered, and is discussed more fully in Härdle & Marron (1991). The method is based on resampling from the estimated residuals and involves two smoothing applications and two corresponding smoothing parameters. The resulting error bars are non-symmetrical, as a consequence of attempting to adjust for bias. Hall & Titterington (1988) construct a confidence band in the context of calibration of radiocarbon dating. Sun & Loader (1994) use an approximation to the tube formula, to produce simultaneous confidence regions, valid for multiple covariates. The method assumes Normal errors, and considers bounds on the bias in order to make appropriate adjustments.

A major difficulty is that the process of smoothing used in the construction of the smooth estimate $\hat{g}$ generally introduces bias. The principal bias term may involve the curvature of the true curve, expressed in $g''(x)$, the design pattern of the $x_i$'s, and a power of the smoothing parameter $h$. The specific form of the bias will depend on the type of estimator used, as seen in Section 1.2.5. Härdle & Bowman (1988) describe a bootstrap procedure intended to construct a confidence band, but this requires the explicit estimation of the second derivative $g''$ and so becomes complex.

It is shown in this chapter that in the case of comparing two curves these issues can be avoided, and the bias problem can be handled by choosing an appropriate smoother.

### 3.2.1 A reference band for equality

Ratkowsky (1983) describes regression data originally collected by I.S. Rogers of the South Australian Department of Agriculture and Fisheries. Figure 3.2.1 displays one of these datasets, which shows the relationship between yield (in *g/plant*, on a log scale) and density (in $plants/m^2$) for Brown Spanish Onions grown in two localities in South Australia.

Ratkowsky describes a variety of nonlinear regression models but here, and in Chapter 4, following earlier work by Hall & Hart (1990) and King, Hart & Wehrly (1991), the merits of allowing the regression effect to be estimated nonparametrically are demonstrated. An important step is the use of the Gasser-Müller or local linear estimator, both of which have the very helpful property that the bias is independent, at least asymptotically, of the design pattern of the $x$'s, as shown in Section 1.2.5.

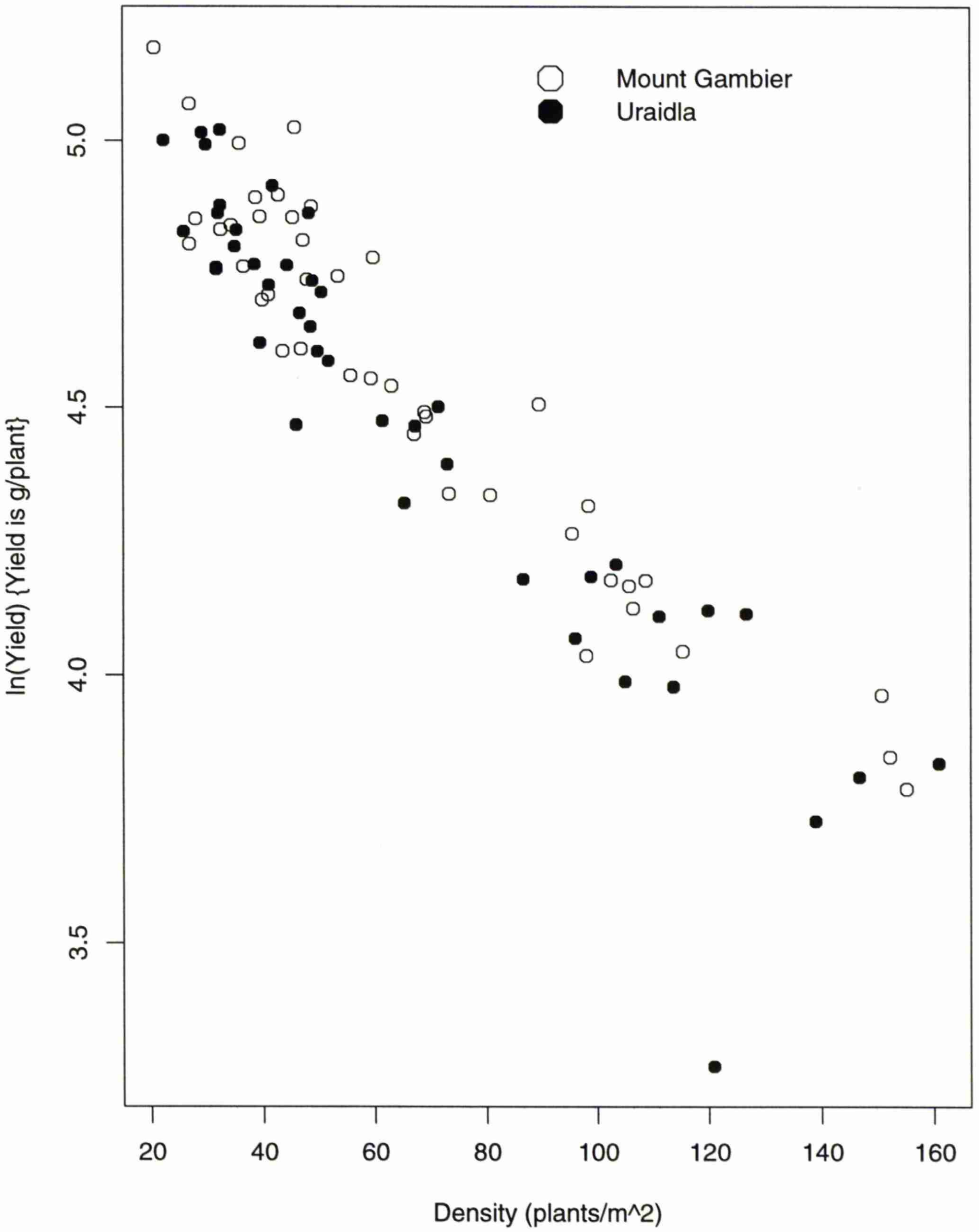
In the Gasser-Müller and local linear cases the bias of the estimator is

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\} g_i'',$$

where $i = 1, 2$ refers in the present context to the two different groups of data.

This expression is independent of the design densities, and depends only on the second derivative of the true functions $g_1$ and $g_2$. A consequence of this expression is that if the two underlying regression curves are identical then so, approximately, is the bias of the estimators. It is then valid to contrast the curves simply as $\{\hat{g}_1(x) - \hat{g}_2(x)\}$ in a test of equality, since the expectation of this contrast is approximately zero when

Figure 3.2.1: Brown Spanish Onions data



the true curves are identical. It is clear that this is not the case if the Nadaraya-Watson estimator is used, unless the underlying densities of the design points are the same.

The variance of the difference between the curves is straightforward to evaluate since, if a common error variance is assumed,

$$var\{\hat{g}_1(x) - \hat{g}_2(x)\} = \left(\sum w_{1j}^2 + \sum w_{2j}^2\right)\sigma^2 \tag{3.2.1}$$

where $w_{ij}(i = 1, 2; j = 1, \ldots, n_i)$ are the weights applied by the smoothing procedure to $y_{ij}$, the $j$-th observation in group $i$. These weights are all known constants. The error variance $\sigma^2$ can be estimated in a variety of ways. Here a simple first-order differencing procedure described by Rice (1984) is used. For convenience it is assumed that the data in each group are ordered by increasing $x$ value. For each group, an estimator of the variance is given by

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i - 1)} \sum_{j=1}^{n_i-1} (y_{i,j+1} - y_{ij})^2.$$

An estimator of the common variance is then constructed as

$$\hat{\sigma}^2 = \frac{1}{2(n_1 + n_2 - 2)} \left\{ \sum_{j=1}^{n_1-1} (y_{1,j+1} - y_{1,j})^2 + \sum_{j=1}^{n_2-1} (y_{2,j+1} - y_{2,j})^2 \right\}$$

by pooling across the two groups. Hall, Kay & Titterington (1990) discuss estimates based on higher-order differencing.

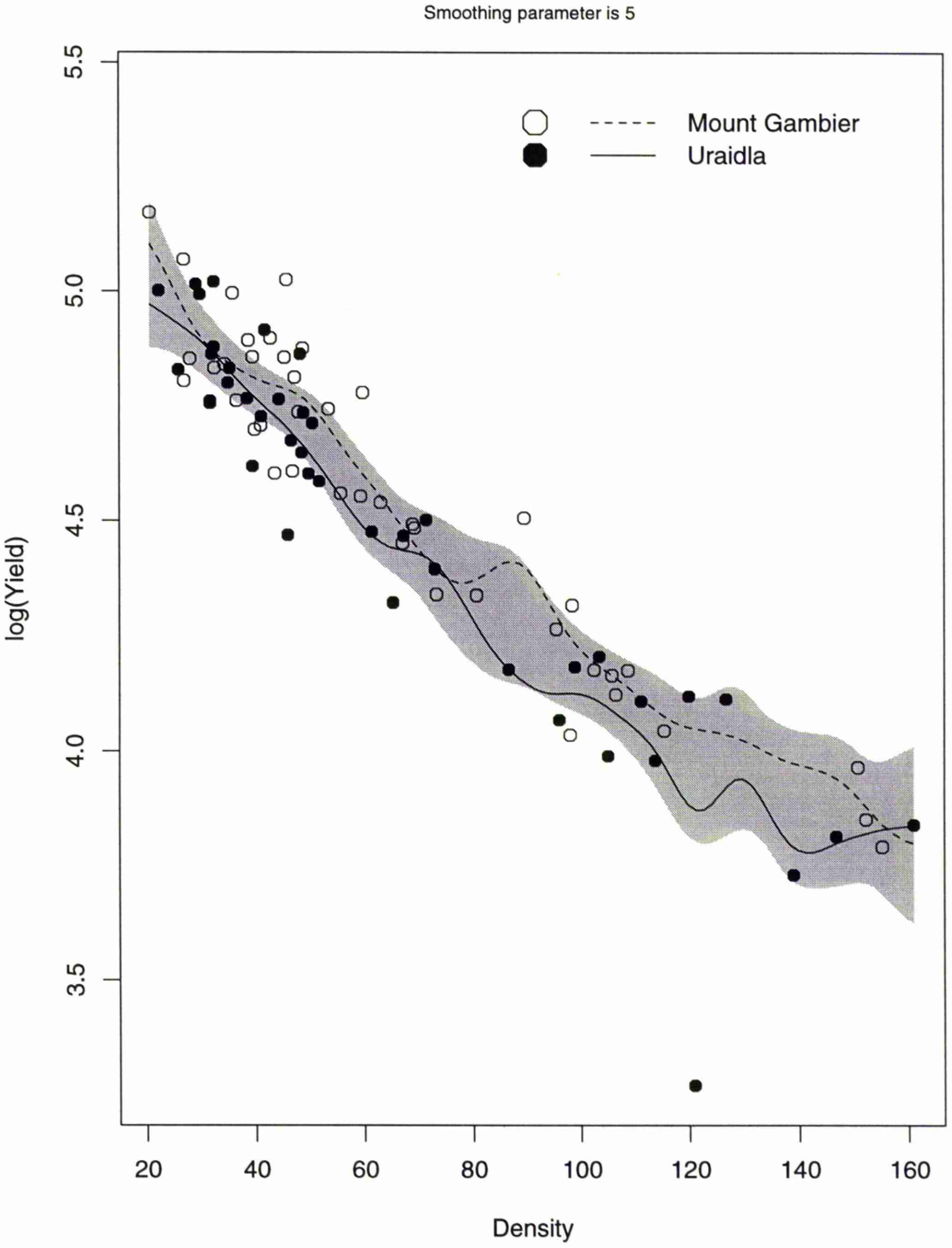
In Chapter 4 a formal test of equality is derived for the entire curves under the assumption of Normal errors. This test returns a $p$-value greater than 0.2 for all reasonable values of smoothing parameter. As a graphical procedure, a reference band for equality is superimposed on the onions data in Figure 3.2.2, along with the estimated unconstrained nonparametric curves. The Gasser-Müller smoother is used here.

At each value of $x$, the band is centred at the average of the two curves and has width $2s.e.(\hat{g}_1(x) - \hat{g}_2(x))$, calculated through formula (3.2.1) with estimated variance $\hat{\sigma}^2$. The result of the global test is confirmed by the fact that the smoothed curves lie within the reference band for all values of $x$.

Had the plotted curves exceeded the band at any point, this would have indicated that the difference was greater than two standard errors, and hence that there was a signficant difference between the groups. A more precise procedure would perhaps be to use the percentage point of, for example, a $t$-distribution instead of the coefficient 2, and introduce a Bonferroni, or other, adjustment, to reflect the multiple comparisons which are implicitly taking place. This might apply especially when the sample sizes are small. However, in the nonparametric context, exact distributional results are not available, and instead the straightforward use of the value 2 provides a helpful simplification, reflecting a popular "rule of thumb" in inference.

The simplest interpretation of the band is as an "acceptance region" for the hypothesis of no difference between the curves at each point in the covariate space. For this reason, the terminology *reference band* is helpful as it stresses the comparison of the observed data with a proposed model, in this case of a common regression

Figure 3.2.2: Reference band for equality of nonparametric regression curves (Brown Spanish Onions data)



relationship for each group.

Figures 3.2.2 and 3.2.3 illustrate the effect of changing the smoothing parameter $h$. This is a very important issue in the estimation of the true underlying curves. However, in the present case, where interest centres on the existence of differences between the curves, the smoothing parameter has remarkably little effect on the conclusions. Small values of $h$ produce curves which are less smooth, but this increased variability is also reflected in the width of the reference band. Indeed, for all realistic values of $h$ the curves lie within the reference band. There is therefore no need in this case to consider more carefully the most appropriate smoothing parameter to use. As in Chapter 2, with the density contour, the question of how to select a smoothing parameter can be sidestepped to a large extent.

### 3.2.2 Regression with binary responses

Hosmer & Lemeshow(1989) describe data on the occurrence of low birthweight collected in Massachussetts. Figure 3.2.4 plots the low birthweight indicator (0 represents $\geq 2500gms$, 1 represents $< 2500gms$) against the mother's weight at her last menstrual period, which is clearly an important variable to take into account when assessing the weight of the baby.

Two groups of data are presented, corresponding to mothers who did, or did not, smoke during pregnancy. The vertical positions of the data have been randomly perturbed by small amounts in order to indicate the patterns in the data more clearly.

Figure 3.2.3: Another reference band for equality for the Brown Spanish Onions data

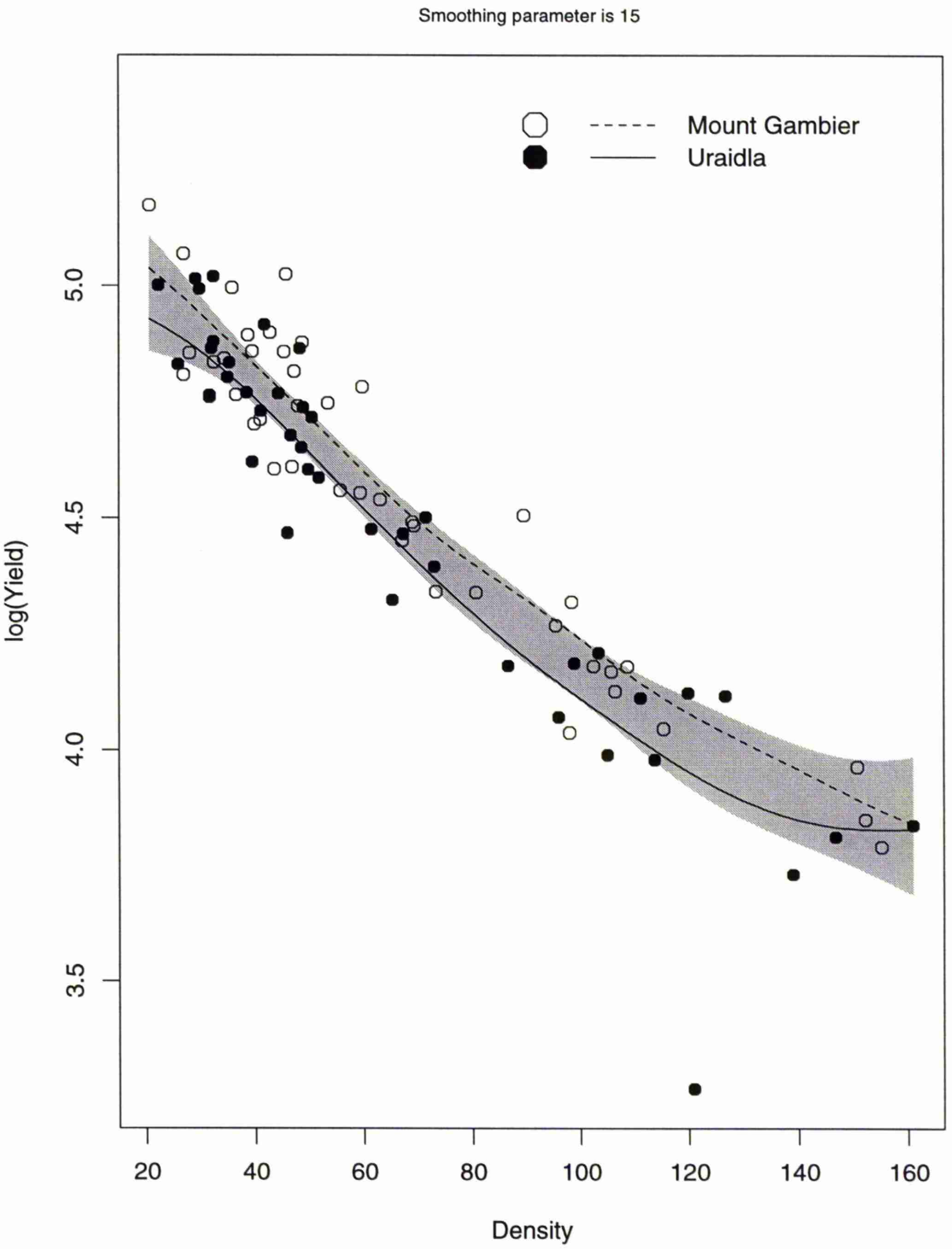
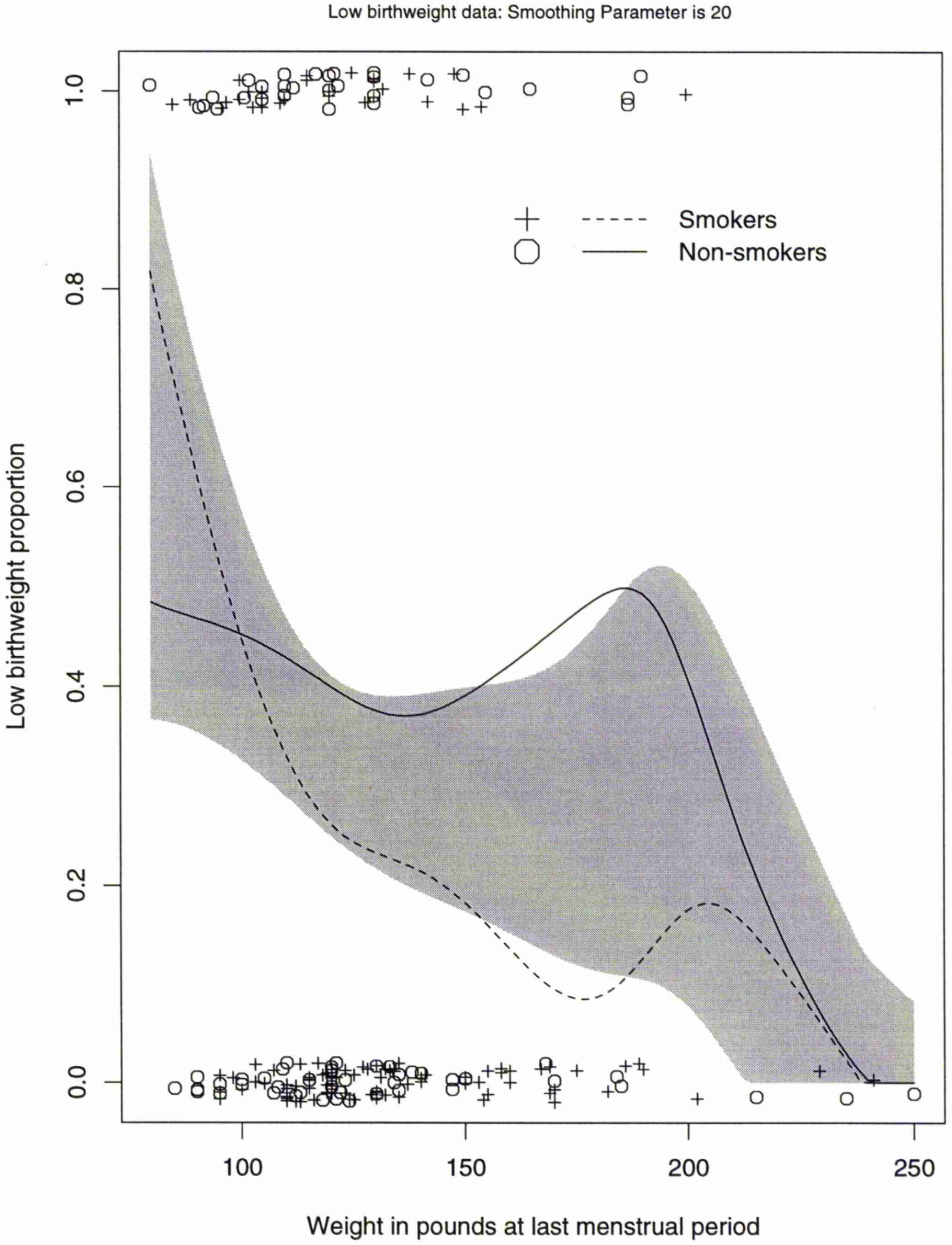


Figure 3.2.4: Reference band for equality of nonparametric regression with binary data



When the response is binary, it is helpful to use nonparametric regression to display the form of the relationship between the response and a continuous covariate, as discussed in Copas (1983). The local line fitting method of smoothing was applied to each group to create the smooth curves displayed in Figure 3.2.4. As in Section 3.2.1, any bias terms arising from the nonparametric smoothing cancel, at least asymptotically, when two curves are contrasted. This is discussed more fully in Chapter 5, when a formal test of equality is derived.

Returning to the current example, a smoothing parameter of 20 was used in Figure 3.2.4. As expected, the incidence of low birthweight generally decreases with increasing mother's weight. However, there is a marked difference between the two curves for intermediate values of mother's weight.

It is helpful to assess this difference through the use of a reference band for equality. In this case, $var(y) = p(x)\left\{1 - p(x)\right\}$, and so the variance of $\hat{p}_i = \sum_{j=1}^{n_i} w_{ij} y_{ij}$ can be estimated as

$$v\hat{a}r\{\hat{p}_i(x)\} = \sum_{j=1}^{n_i} w_{ij}^2 \hat{p}_i(x_{ij})\left\{1 - \hat{p}_i(x_{ij})\right\},$$

where the $w_{ij}$'s are the weights as previously defined, and the response $y_{ij}$ is binary. The estimated standard error of the difference $\{\hat{p}_1(x) - \hat{p}_2(x)\}$ can therefore be easily constructed, as before. Figure 3.2.4 displays the reference band, centred on the average as before, and truncated where necessary at 0 and 1. The plotted curves lie outwith the band from around 155 to 190 pounds. However, the high variance in this region of the plot means that the curves only just depart from the reference band, as the differences are not as marked as might be indicated from a simple plot of the curves alone.

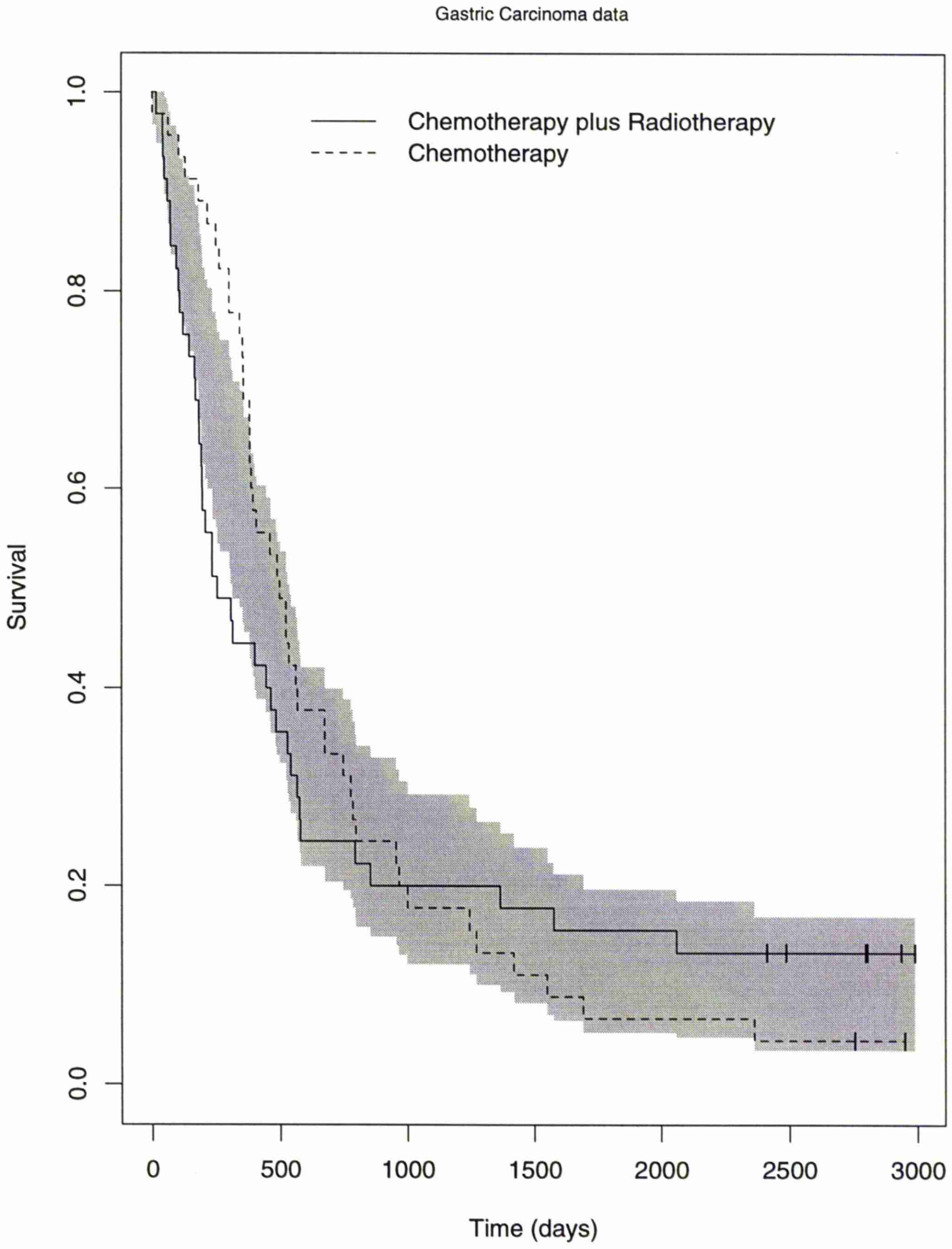
It appears from the reference band that there is a difference between the two groups over the range indicated, but it is only marginal. This is underlined by the results of a formal nonparametric test on these data, in Section 5.3.

As in the previous examples, the conclusion drawn from the reference band is unchanged, for a wide range of smoothing parameter.

## 3.3 Survival curves

The extension of the ideas of the previous sections to the case of Kaplan-Meier survival curves is straightforward. Figure 3.3.5 displays estimated survivor functions constructed from two groups of gastric carcinoma patients undergoing different treatments, using data analysed by Stablein & Koutrouvelis (1985). A log rank test (Cox & Oakes (1984), p.104) produces a large $p$-value, whereas the test proposed by Stablein & Koutrouvelis, which is sensitive to crossing hazards, identifies the two survivor functions as significantly different. Greenwood's formula (Cox & Oakes (1984), p.50) allows approximate standard errors to be added to the estimated log survivor function at any point. A more helpful direct assessment of the relative performances of the two treatments comes through adding the variances, to construct the standard error of the difference of the log survivor functions at each point. This is used to construct a reference band, centred between the two log survivor functions. The band is then

Figure 3.3.5: Reference band for equality of survivor functions



re-expressed and plotted on the original scale. As a result of the transformation, the displayed band is in this case not centred on the average of the two survivor functions. However, any departure of one curve from the reference band will, as usual, be mirrored by a departure of the other at the same points.

The band again provides a helpful indication of where the performances of the treatment groups are identifiably different. In the present case the two survivor functions are seen to differ at values around one year, and production of the band acts as a caution against routine application of a log rank test in this setting.

### 3.3.1 Reference bands for parallelism

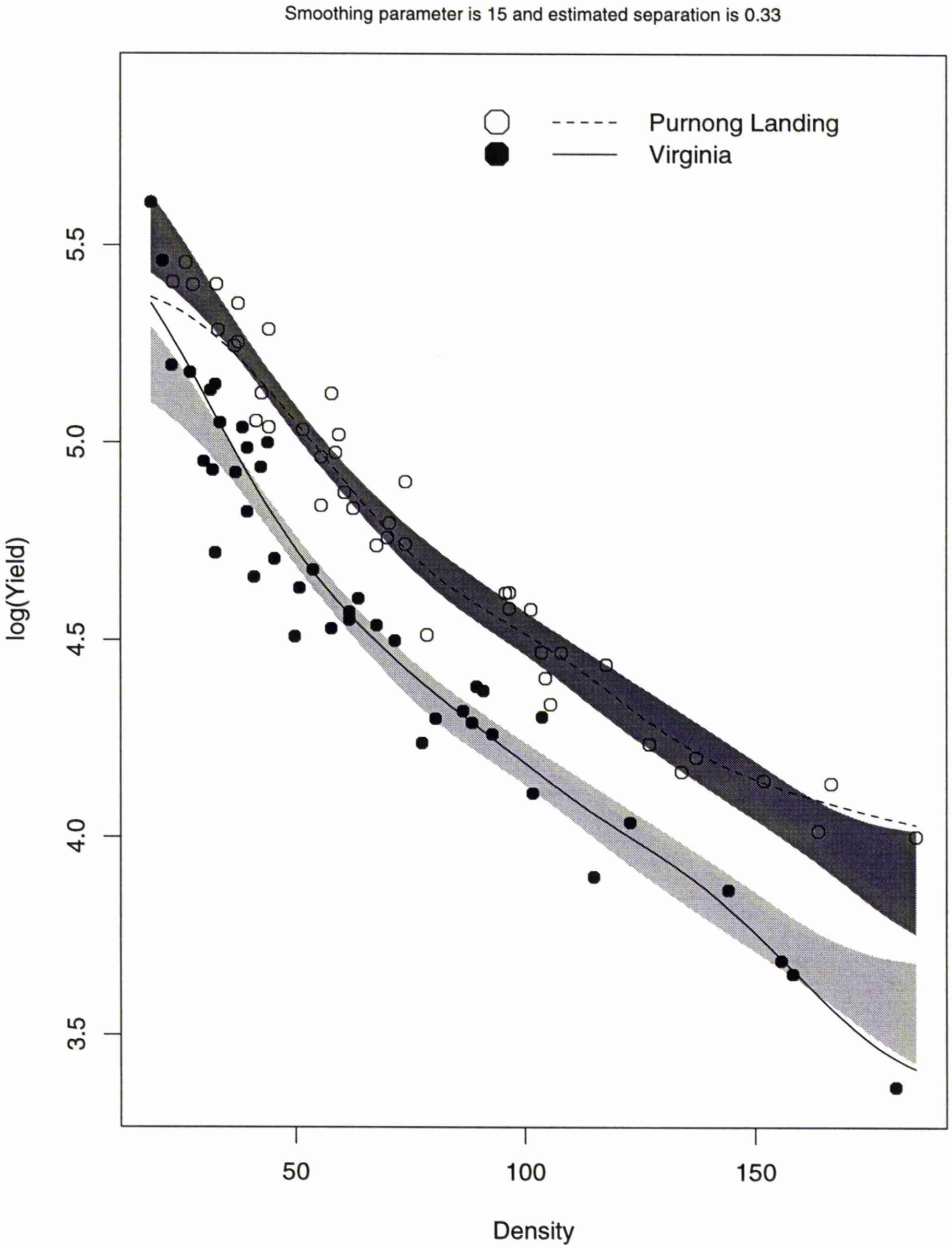
The contexts discussed so far have all been concerned with a null hypothesis of equality. It is also possible to create reference bands which express where curves are expected to lie under a hypothesis of *parallelism*. Figure 3.3.6 displays a second set of onion data, this time of White Spanish Onions.

It is immediately clear from the plot of the data that the two groups, again corresponding to two different localities, are producing different yields. A natural question to ask is whether the underlying regression curves are parallel, which enables a particularly simple interpretation of the differences between the groups.

Chapter 4, extending the work of Speckman (1988), considers the analysis of covariance model

$$y_{ij} = \alpha_i + g(x_{ij}) + \varepsilon_{ij} \tag{3.3.2}$$

Figure 3.3.6: Reference bands for parallelism (White Spanish Onions data)



within the more general model which allows a separate curve $g_i(x)$ for each group. Here $y_{ij}$ and $x_{ij}$ are the response and covariate values of observation $j$ in group $i$, and $g_i$ is the regression curve for group $i$. For identifiability, the intercept term $\alpha_1$ is set to zero. Chapter 4 discusses the estimation of $\alpha_i$ and of $g$, which is by least squares and nonparametric smoothing respectively, and introduces a test of parallelism ($g_i = g$) based on a comparison of the estimated individual curves $\hat{g}_i$ with an estimated common curve $\hat{g}$, at all design points. With the present data, this test produces a significant result for a wide range of smoothing parameters. This differs from a parametric non-linear model, such as a reciprocal quadratic curve, which appears to capture the broad trend in the data but which does not identify differences in the shapes of the curves for each group.

In order to identify more clearly the features of the curves which provide evidence against the parallel hypothesis, Figure 3.3.6 has two reference bands for parallelism superimposed. These are constructed by estimating the difference $\alpha_2$ between the parallel curves of the model in (3.3.2) and translating the second curve by this distance so that a reference band for equality may be constructed, using a width of two standard errors as before. A copy of this band is then translated, along with the fitted curve, back to its original position. This process does not allow for the variability in the estimation of $\alpha_2$ but it does give a useful approximation, from which it is apparent that the two curves are too close together for small densities, and too far apart for high densities. The observations with the two smallest values of density appear to have rather high values of yield. If these two observations are excluded and the model is refitted, then the hypothesis of parallelism is not rejected, and the smooth curves lie almost entirely within the reference bands.

The process of translation described above works well when the two curves are some distance apart. In cases where the curves are closer, and the plotted bands would overlap, it is sufficient to plot one band only. The symmetry of the construction of the bands ensure that where one curve exceeds the band the other curve will do so too.

## 3.4 Comparing nonparametric and parametric curves

The examples of previous sections dealt with a variety of cases where interest is in the comparison of two nonparametric curves. It is also common to compare a nonparametric curve with a parametric one, in assessing the goodness-of-fit of a parametric model. Azzalini, Bowman & Härdle (1989) discussed logistic and Poisson regressions, while Eubank & Hart (1993) review a number of tests which have been developed in this general area.

### 3.4.1 A reference band for linearity

Azzalini & Bowman (1993) described a formal test, based on the residual sums of squares associated with a linear and a nonparametric model, which can be applied to regression data to assess linearity, under the assumption of Normal errors. This test is described more fully in Chapter 6. With simple linear regression, the idea of a reference band is applied most naturally with local linear smoothing, since this estimator tends to the fitted linear model as the smoothing parameter increases. In addition, the bias of the nonparametric estimator is also close to zero when the true model is linear. The fitted linear regression at a point $x$ can be expressed as

$$\hat{\alpha} + \hat{\beta}x = \sum l_i y_i$$

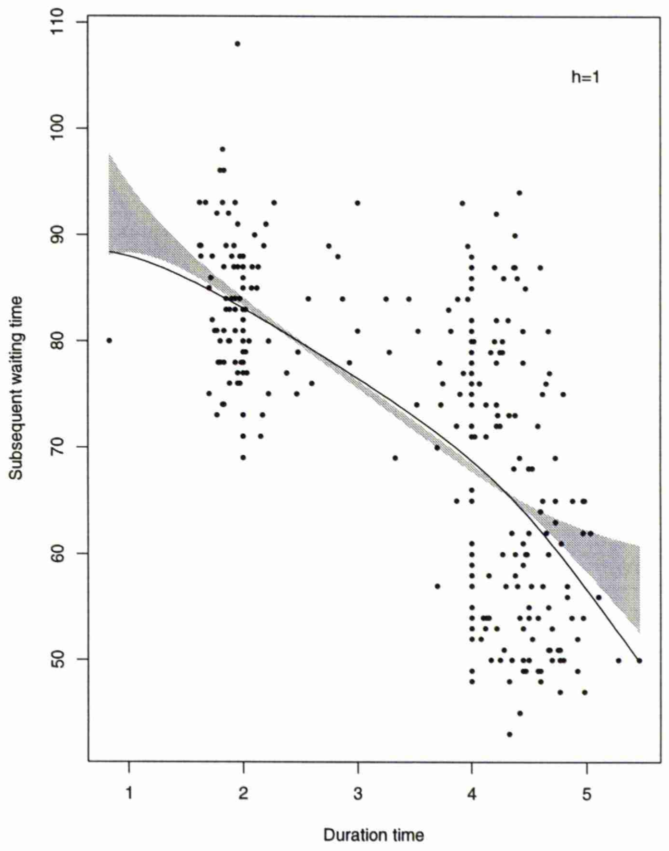
for known weights $l_i$, and so the variance of the comparison between the parametric and nonparametric curves at $x$ is

$$var\left\{\hat{g}(x) - \hat{\alpha} - \hat{\beta}x\right\} = var\left\{\sum (w_i - l_i)y_i\right\} = \sum (w_i - l_i)^2\sigma^2.$$

The standard error of this comparison can therefore be constructed easily. The error variance $\sigma^2$ can be estimated by the differencing procedure described in Section 3.2.1. In this case it is wise to reduce bias by removing the linear trend and applying the procedure to the residuals from the linear regression.

Figure 3.4.7 displays data on duration times of eruptions of the Old Faithful Geyser, as a predictor of subsequent waiting times until the next eruption. This is the same data set encountered in Chapter 2. Data of this type were examined by Cook & Weisberg (1982) and used by Silverman (1985) to illustrate how nonlinearity can be highlighted by nonparametric smoothing. The data are presented here with a reference band for linearity superimposed. This band is centred at the fitted linear regression and extends two $e.s.e.(\hat{g}(x) - \hat{\alpha} - \hat{\beta}x)$ above and below. The fact that

Figure 3.4.7: A reference band for linearity on the geyser data



the smooth nonparametric regression curve strays outside this band at several points, especially larger duration times, indicates why the formal test rejects linearity.

### 3.4.2 A reference band for residual plots

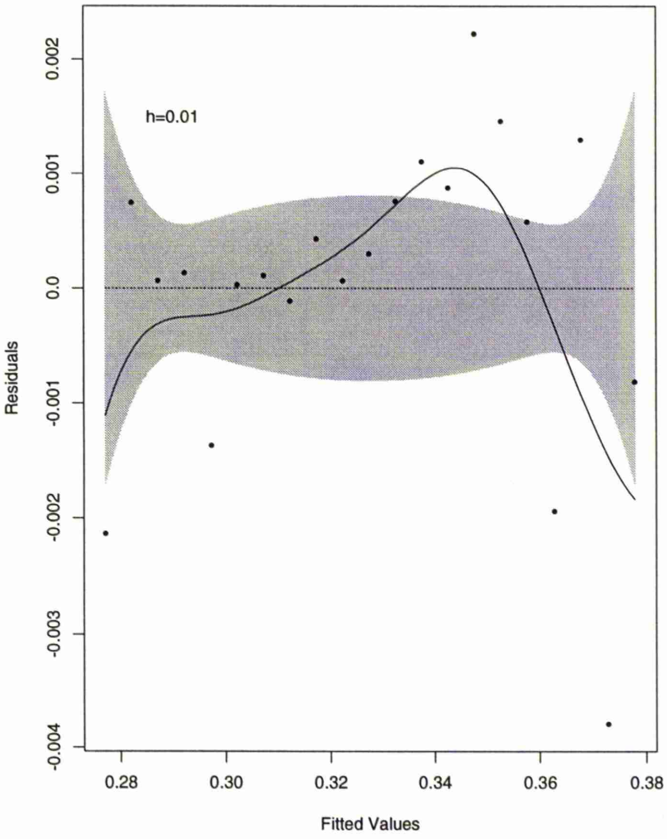
An alternative to the approach of Section 3.4.1 is to carry out an assessment of linearity with a residual plot. Figure 3.4.8 displays a plot of residuals against fitted values from a linear regression model on data reported by Yanagimoto & Yanagimoto (1987). These data are of high precision, and interest centres on whether a linear regression model is appropriate. In common with other procedures which have been applied to these data, the pseudo-likelihood ratio test of Azzalini & Bowman (1993) identified significant nonlinearity, and yielded $p$-values around 0.02 for a wide range of smoothing parameters.

A reference band can be constructed in a manner very similar to that described in Section 3.4.1. To give a general description, suitable for any type of plot of raw residuals, the fitted model is $\hat{\boldsymbol{\beta}}^T\mathbf{X}$ where $\hat{\boldsymbol{\beta}}$ is a vector of parameter estimates and $\mathbf{X}$ is a design matrix. The nonparametric regression is constructed from the plot of residuals $e_i$ against fitted values, or some other variable of interest, denoted here by $z_i$. The variance of the comparison of this nonparametric curve with the fitted model, as expressed by the horizontal axis, is therefore

$$var\left\{\hat{g}(z)\right\} = var\left\{\sum w_i e_i\right\} = var\left\{\mathrm{w}^T\mathrm{e}\right\} = \mathrm{w}^T\left[\mathrm{I_n} - \mathrm{X}(\mathrm{X^T X})^{-1}\mathrm{X^T}\right]\mathrm{w}\sigma^2$$

where $\mathbf{w}$ and $\mathbf{e}$ represent the weights and residuals in vector form. The estimate of

Figure 3.4.8: A reference band for trend in a residual plot

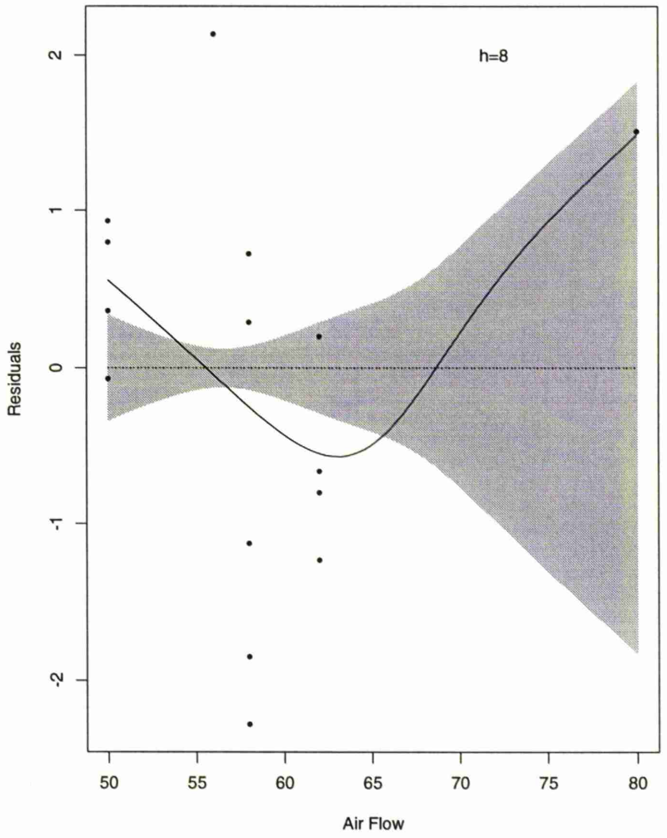


the error variance $\sigma^2$ can be constructed in this case from successive differences of the residual plot.

The strength of the evidence of nonlinearity from the plot of raw residuals, even with the nonparametric regression curve superimposed, is unclear. However, the addition of the reference band confirms that the curvature in the second half of the plot does indeed indicate the linear regression model to be an inadequate description of the data.

An advantage of the method of residual smoothing is that it can be applied when there is more than one explanatory. The residuals can be plotted against a single explanatory of interest, or against the fitted values, as a general check on the model chosen. As a second example, and one with several covariates, the Stack Loss data set of Chapter 2 is considered again. In Chapter 2, it was commented that Daniel & Wood (1980), in their analysis of these data, concluded that while a linear term in water temperature was sufficient, a quadratic term in air flow might be required. In this example, this is investigated further, and the residuals from a Normal linear regression are smoothed against the explanatory of interest (air flow). Figure 3.4.9 shows a reference band superimposed on a plot of the residuals against air flow. The paucity of data at the upper end of the covariate scale is reflected in the very large variance over that area. It is clear from the figure that the pattern in the residuals indicates that a simple linear regression in air flow is inappropriate. It would appear that a polynomial of higher degree, or a smooth function, is required, thus confirming

Figure 3.4.9: A reference band for the stack loss data



the observations made by Daniel & Wood. This is investigated more formally in Chapter 6, where a modified version of the test proposed by Azzalini & Bowman (1993) is applied to these data.

## 3.5 Discussion

The use of reference bands has been explored in a wide variety of simple but common problems where nonparametric curves arise. These bands express a hypothesis, of equality, parallelism, or parametric shape, and allow a helpful graphical assessment of whether the data support the hypothesis. The bands should not be used as an inferential tool on their own. They are intended to support the results of more formal tests, such as those proposed in Chapter 4, by indicating the features of the curves which are likely to be the cause of a small $p$-value, or by indicating why apparent features do not contribute strongly to statistical significance.

In those cases where nonparametric smoothing techniques have been employed, there is a remarkable stability of the results over a wide range of smoothing parameters. Where use of a single smoothing parameter is preferred, the "plug-in" methods, briefly introduced in Section 1.2.3 of Chapter 1, offer very effective choices. The difficulties of bias in the construction of the estimates have also been avoided here, since the focus of interest is on the contrast of two curves. These issues demonstrate the important difference between estimation and inference in this context.

Several examples of reference bands have been illustrated but the principle can be applied to a variety of other contexts. In particular, all the examples discussed in this chapter dealt with the comparison of two groups. If more groups were present, it would be reasonable to compare each pair of groups in turn. In this case, a Bonferroni-type adjustment could be made to the width of each reference band, to compensate for the multiple comparisons.

Bowman & Young (1996) extend the ideas of this chapter further by demonstrating how a reference band for Normality can be produced, and compared with a (nonparametric) density estimate.

# Chapter 4

# Nonparametric analysis of covariance

## 4.1 Introduction

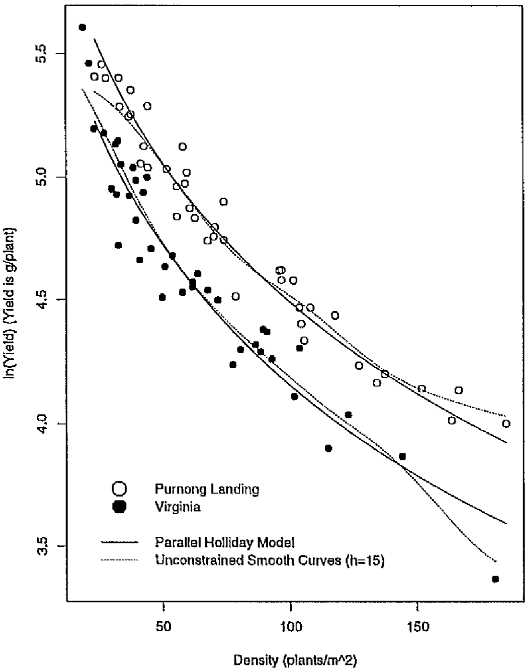
As already demonstrated in Chapter 3, analysis of covariance is a very useful and common technique for comparing the values of a response variable $y$ across several groups in the presence of a covariate effect. We can write this model in the following general notation

$$y_{ij} = m_i(x_{ij}) + \varepsilon_{ij} \quad \text{where} \quad i = 1, \ldots, p; \quad j = 1, \ldots, n_i. \tag{4.1.1}$$

In the simple linear case, $m_i(x) = \alpha_i + \beta_i x$. In some situations there is a parametric, but nonlinear, candidate for $m_i$. For example, Figure 4.1.1 displays again the White Spanish Onions data from Chapter 3. As before, Yield is on the log scale, in order to stabilise the variance.

For these data, Ratkowsky (1983) proposes fitting a Holliday (1960) yield-density

Figure 4.1.1: White Spanish Onions data



model, which has deterministic components

$$log(y) = -log(\alpha + \beta x + \gamma x^2),$$

and these are displayed in Figure 4.1.1 with the (dotted) nonparametric estimates. Ratkowsky's approach leads to a nonlinear analysis of covariance in order to identify differences between the groups. This involves nonlinear regression and the resulting inferences are necessarily approximate.

An extension of the general model in (4.1.1) is available through the semiparametric approach of Speckman (1988). Here specific shape assumptions are relaxed and $m_i(x) = \alpha_i + g(x)$, where $g(x)$ is assumed only to be a smooth function, estimated nonparametrically. Speckman discusses a variety of issues of estimation and describes approximate $F$-tests which can be applied in the present context to identify differences among groups. These tests are based on asymptotic Normality.

Hall & Hart (1990) use a bootstrap approach to test the equality of two smooth curves when the design points of the two groups are identical. This assumption enables bias to be eliminated. The test statistic is essentially a scale-adjusted version of $\sum(\hat{g}_1 - \hat{g}_2)^2$, where $g_1$ and $g_2$ are the true regression curves. A later generalisation drops the assumption of identical designs, by pairing off "similar" design points. Some loss of accuracy is observed, due to errors arising from the mismatch of the two designs. The test also generalises to the comparison of more than two groups. As will be seen, the tests presented in this chapter take a slightly different approach in the construction of a test statistic, and do not require common design densities. In all the tests here, Normal errors are assumed, allowing the calculation of a chi-squared null distribution, as opposed to using the bootstrap method.

King, Hart & Wehrly (1991) also propose a test of equality for two nonparametric regression curves. Design points are identical, in order to eliminate bias. Normal errors are assumed, although the error variances are not constrained to be equal across groups. The type of smoother is not stated explicitly, although the Gasser-Müller (1979) estimator is recommended as its properties make the test asymptotically valid when the Normality assumption fails. The test statistic is of a similar form to Hall

& Hart (1990), based on the sum of squared differences between the two estimates. In deriving an expression for a $p$-value, King et al (1991) follow a similar argument to that presented in this chapter, but use Monte Carlo methods to approximate the null distribution.

An alternative method of inference is the heuristic $F$-test approach of Hastie & Tibshirani (1990). By analogy with parametric modelling, this involves contrasting residual sums of squares under competing models, and comparing the resulting test statistic against an appropriate $F$ distribution. Degrees of freedom for the tests are derived in a similar manner to that discussed in Section 1.2.4 of Chapter 1. The authors note that although exact distributional results are not available, the approximate $F$-test approach provides at least a simple rough guide.

The aim of this chapter is to explore the use of a general nonparametric model in the ANCOVA setting, in particular allowing the assumption of additivity to be tested within the more general model of a different covariate effect for each group. The approach taken is analogous to that used by Azzalini & Bowman (1991) in the context of repeated measurements, although different problems of estimation and testing arise in the present context. Highly accurate moment-based approximations to the null distributions of test statistics will be used as in Azzalini & Bowman (1993). The test statistics are described in Section 4.2 along with issues of estimation of $g$ and $\alpha_i$. In particular, as in Chapter 3, the Gasser-Müller (1979) or local linear estimators can be employed to reduce problems of bias in estimation of $g$. The Spanish Onion data sets are used as examples in Section 4.3.

The power of the proposed tests is estimated through a simulation study and contrasted with a linear parametric approach in Section 4.4. Some final conclusions are drawn in Section 4.5.

## 4.2 Tests of equality and parallelism

### 4.2.1 A test statistic for equality

We consider first the simple case of testing the equality of two or more smooth curves. The hypotheses in this case are most clearly formulated as

H0: $y_{ij} = g(x_{ij}) + \varepsilon_{ij}$

H1: $y_{ij} = g_i(x_{ij}) + \varepsilon_{ij}, \quad g_i$'s not all equal

where the $\varepsilon_{ij}$ are independent N(0,$\sigma^2$) random errors. By analogy with one-way analysis of variance, a natural test statistic is

$$TS = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i} \{\hat{g}_i(x_{ij}) - \hat{g}(x_{ij})\}^2}{\hat{\sigma}^2} \tag{4.2.2}$$

where individual estimates of each curve are contrasted with a common estimate under the assumption of equality. This comparison is made at all points $x_{ij}$ in the design space. In order to remove the effects of scale, an estimate of the error variance $\sigma^2$ appears in the denominator. This is of exactly the same form as the test statistic proposed by Azzalini & Bowman (1991) in a repeated measurements model where several groups are compared and the time effect is assumed to vary smoothly.

However, different issues arise in the present context where covariate values can be irregularly spaced.

A simple nonparametric curve estimator $\hat{g}_i$ is provided by the kernel method. As in Chapter 3, the Nadaraya-Watson (1964) estimator is unsuitable because the bias depends on the underlying design density. The Gasser-Müller (1979) and local linear methods of smoothing, introduced in Section 1.2 of Chapter 1, share the attractive property that the bias is asymptotically independent of the design density, and this is exploited again in this chapter. The form of the bias is also given in Chapter 1.

Under the null hypothesis that the curves $g_i$ are identical, the asymptotic biases in estimation of the unconstrained curves and the common curve are identical. The bias terms in the numerator of the test statistic (4.2.2) therefore cancel. This would not be the case with Nadaraya-Watson (1964) smoothing unless the underlying design densities were the same, or there was a fixed design which was the same for all groups.

To complete the test statistic an estimator of $\sigma^2$ is required. As in Chapter 3, the first-order difference approach of Rice (1984) is used and a pooled estimator is constructed as

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{p}(n_i - 1)\hat{\sigma}_i^2, \quad \text{where} \quad n = \sum_{i=1}^{p} n_i,$$

under the assumption that error variance is constant across groups.

### 4.2.2 Distribution of the test statistic

Azzalini & Bowman (1991, 1993) examined test statistics of a similar form to (4.2.2) and made use of the fact that both the numerator and denominator are quadratic forms in the data. Here, the vector of fitted values for $\{g_i(x_{ij})\}_{j=1}^{n_i}$ can be written in vector-matrix notation as $\mathbf{g_i} = \mathbf{S_i y_i}$, where $\mathbf{S_i}$ is an $n_i \times n_i$ matrix of known weights. The entire collection of these fitted values can be represented as

$$\hat{\mathbf{g}} = \mathbf{S_d y}$$

where $\mathbf{S_d}$ is an $n \times n$ matrix. The vector of fitted values derived from the assumption of a single curve for all the data may be represented as $\mathbf{S_s y}$, where $\mathbf{S_s}$ is a different $n \times n$ matrix of weights.

The numerator of test statistic (4.2.2) can then be written as

$$\mathbf{y}^T[\mathbf{S_d} - \mathbf{S_s}]^T[\mathbf{S_d} - \mathbf{S_s}]\mathbf{y}.$$

For simplicity, the matrix $[\mathbf{S_d} - \mathbf{S_s}]^T[\mathbf{S_d} - \mathbf{S_s}]$ shall be represented by $\mathbf{Q}$. When $\mathbf{y}$ is replaced by $\mathbf{g} + \varepsilon$, several terms in the expansion of this numerator disappear, at least asymptotically, as a result of the bias properties of the Gasser-Müller (or local linear) estimator and it simplifies to

$$\varepsilon^T \mathbf{Q} \varepsilon.$$

The formula for $\hat{\sigma}^2$ can be expressed as

$$\hat{\sigma}^2 = \mathbf{y}^T \mathbf{B} \mathbf{y}$$

with expectation

$$E(\hat{\sigma}^2) = E(\varepsilon^T \mathbf{B} \varepsilon) + \mathbf{g}^T \mathbf{B} \mathbf{g}$$

where $\mathbf{B}$ is an $n \times n$ symmetric matrix. The last term is the sum of successive squared differences of the true function $\mathbf{g}$. This term is small relative to the first term in $\varepsilon$ and it is easy to show that ignoring this term has the effect of making the test conservative.

Putting these two results together gives the following approximate expression for a $p$-value for the test:

$$p = P\left\{\frac{\varepsilon^T \mathbf{Q} \varepsilon}{\varepsilon^T \mathbf{B} \varepsilon} > Obs\right\} = P\left\{\varepsilon^T (\mathbf{Q} - \mathbf{B} \times Obs) \varepsilon > 0\right\} \qquad (4.2.3)$$

where $Obs$ is the observed value of the test statistic. This is now a quadratic form in Normal variables (not necessarily positive) of the type $\mathbf{z}^T \mathbf{A} \mathbf{z}$ where $E(\mathbf{z}) = \mathbf{0}$ and $\mathbf{A}$ is an $n \times n$ symmetric matrix.

Johnson & Kotz (1970b) show that the $s$-th cumulant of the distribution of such a quadratic form is given by

$$K_s = 2^{(s-1)}(s-1)!\, tr\{(\mathbf{VA})^s\}, \quad \text{where} \quad \mathbf{V} = Cov(\mathbf{z}). \qquad (4.2.4)$$

In this case, the equation in (4.2.3) is scale invariant, and so $Cov(\mathbf{z}) \equiv Cov(\varepsilon)$ can be set to the identity matrix, without loss of generality. Result (4.2.4) is used to calculate an accurate approximation to the null distribution. The approach taken is to match the first three moments of the test statistic with those of a suitable chi-squared distribution. Pearson (1963) and Solomon & Stephens (1978) provide some evidence of the suitability of this technique.

From Johnson & Kotz (1970a), the $s$-th cumulant of a $\chi^2(b)$ distribution is given by

$$K_s = 2^{(s-1)}(s-1)!b. \tag{4.2.5}$$

For this test, an $a\chi^2(b)+c$ distribution is used as the approximate null distribution. Formula (4.2.5) gives the mean of an $a\chi^2(b)+c$ distribution as $ab+c$, the variance as $2a^2b$, and the skewness as $\sqrt{\frac{8}{b}}$. Now, formula (4.2.4) gives cumulants $K_1$, $K_2$ and $K_3$ for the quadratic form in formula (4.2.3), and hence the mean, variance and skewness can be found ($\mu = K_1$, $\sigma^2 = K_2$ and $\alpha_3 = \frac{K_3}{\sigma^3}$ respectively).

Thus, in order to match the mean, variance and skewness of the quadratic form with that of an $a\chi^2(b)+c$ distribution, it is necessary to take

$$\begin{aligned} a &= \frac{K_3}{4K_2} \\ b &= \frac{8K_2^3}{K_3^2} \\ c &= K_1 - \frac{2K_2^2}{K_3}. \end{aligned}$$

In other words, given the mean, variance and skewness of the distribution of the test statistic (using formula (4.2.4)), $a$, $b$ and $c$ are calculated such that an $a\chi^2(b)+c$ distribution has the same mean, variance and skewness. This $a\chi^2(b)+c$ distribution is then used as an accurate approximation to the null distribution, to calculate a $p$-value in the usual manner.

### 4.2.3 Estimation and inference in the parallel model

The hypotheses of interest in this case are

H0: $y_{ij} = \alpha_i + g(x_{ij}) + \varepsilon_{ij}, \quad \alpha_1 = 0$

H1: $y_{ij} = g_i(x_{ij}) + \varepsilon_{ij}, \quad g_i$'s not all equal or parallel

where the $\varepsilon_{ij}$ are independent N(0,$\sigma^2$) errors. The proposed test statistic is

$$TS = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\{\hat{\alpha}_i + \hat{g}(x_{ij}) - \hat{g}_i(x_{ij})\}^2}{\hat{\sigma}^2} \qquad (4.2.6)$$

Again, the properties of the Gasser-Müller (1979), or local linear, estimator ensure that any bias terms in the numerator disappear. The common smooth function $g$ is obtained by subtracting $\alpha_i$ from each $y_{ij}$, giving a sample $(x_{ij}, y'_{ij})$ from a single population. These adjusted responses are used to estimate $g$ (in practice, the estimated $\alpha_i$ are used). This is the partial residual method described by Speckman (1988) and this approach gives

$$\hat{\mathbf{g}} = \mathbf{S}_s(\mathbf{y} - \mathbf{D}\boldsymbol{\alpha})$$

where $\mathbf{D}$ is a design matrix for the parametric part of the model, and $\mathbf{S}_s$ is as described in Section 4.2.2. The model can therefore be expressed as

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha} + \mathbf{S}_s(\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}) + \textit{error terms}$$

which can be rearranged as

$$(\mathbf{I}_n - \mathbf{S}_s)\mathbf{y} = [(\mathbf{I}_n - \mathbf{S}_s)\mathbf{D}]\boldsymbol{\alpha} + \textit{error terms.} \qquad (4.2.7)$$

Applying the familiar least-squares formula to (4.2.7) leads to the following estimator for $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}} = [\mathbf{D}^T(\mathbf{I}_n - \mathbf{S}_s)^T(\mathbf{I}_n - \mathbf{S}_s)\mathbf{D}]^{-1}\mathbf{D}^T(\mathbf{I}_n - \mathbf{S}_s)^T(\mathbf{I}_n - \mathbf{S}_s)\mathbf{y}$$

Thus the $\boldsymbol{\alpha}$ term is estimated parametrically, via a least-squares argument which is commonly applied to semiparametric models, of which the present situation is an example. Speckman (1988) notes that this estimator is asymptotically Normally distributed, with negligible bias.

Since $\hat{\boldsymbol{\alpha}}$ involves the smoother $\mathbf{S}_{\mathrm{s}}$, it therefore depends on the choice of smoothing parameter. Although the tests themselves are carried out over a range of smoothing parameters, as described in Section 4.3, simulations indicated that the best approach was to use the same $\hat{\boldsymbol{\alpha}}$ throughout, regardless of the smoothing parameter used when estimating the actual curves. A satisfactory choice of smoothing parameter for estimating $\boldsymbol{\alpha}$ was found to be $\frac{2R}{n}$, where $R$ is the range of the design points.

As in the case of the test of equality, the test statistic can be expressed as a quadratic form, of the appropriate type. The numerator of (4.2.6) reduces to the quadratic form shown below

$$\begin{bmatrix} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \cdots\cdots \\ \boldsymbol{\varepsilon} \end{bmatrix}^T \begin{bmatrix} (\mathbf{I}_n - \mathbf{S}_{\mathrm{s}})\mathbf{D} & \vdots & (\mathbf{S}_{\mathrm{s}} - \mathbf{S}_{\mathrm{d}}) \end{bmatrix}^T \begin{bmatrix} (\mathbf{I}_n - \mathbf{S}_{\mathrm{s}})\mathbf{D} & \vdots & (\mathbf{S}_{\mathrm{s}} - \mathbf{S}_{\mathrm{d}}) \end{bmatrix} \begin{bmatrix} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \cdots\cdots \\ \boldsymbol{\varepsilon} \end{bmatrix}$$

where the square brackets and dotted lines denote partitioned matrices. The derivation of this expression can be found in the Appendix. Since the expectation of $(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is asymptotically negligible, the quadratic form results can again be applied.

## 4.3 A practical example

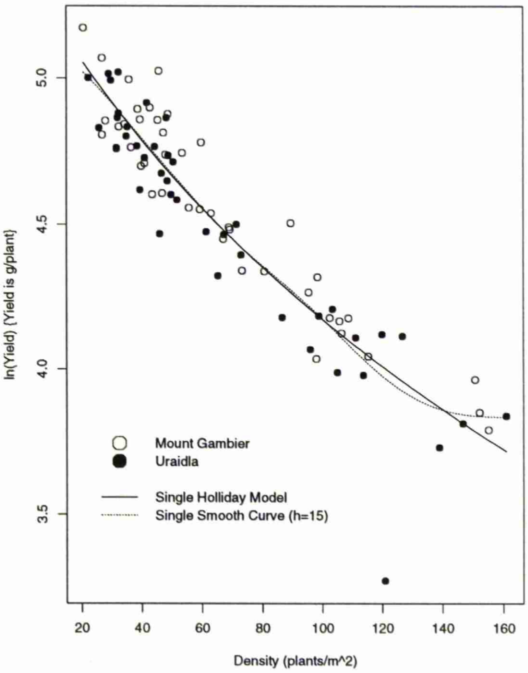
### 4.3.1 Introduction

In this section these methods are applied to the agricultural yield-density data introduced in Chapter 3. Comparisons are also made with the nonlinear models suggested for the data. Figure 4.1.1 showed a plot of the White Spanish Onions data. As a reminder, the data consist of the Yield ($g/plant$) and Density ($plants/m^2$) of White Spanish Onions from two South Australian localities, namely Purnong Landing and Virginia, and the interest is in regressing yield on density. Yield is on the log scale, as a variance stabiliser. The second data set refers to Brown Spanish Onions, from a further two locations, namely Mount Gambier and Uraidla, as illustrated in Figure 4.3.2. As indicated earlier, Ratkowsky (1983) proposes fitting a Holliday (1960) yield-density model to these data. These are displayed in Figure 4.1.1 along with an alternative method of modelling the relationship, namely the nonparametric estimates. Natural questions which arise in this context are (i) whether the yield-density relationship is the same for the two localities, and (ii) whether the yield-density relationships are different but parallel for the different localities. These questions can be answered in a nonparametric way through the tests outlined in Section 4.2.

### 4.3.2 White Spanish Onions

Once again, it is of interest to avoid the contentious issue of choice of smoothing parameter. In this case, the tests are carried out over a wide range of smoothing

Figure 4.3.2: Brown Spanish Onions data



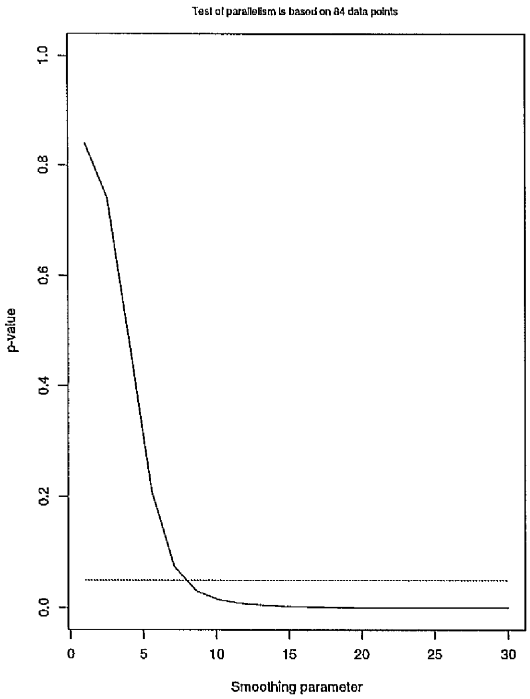
values, and a $p$-value calculated for each. The results are then presented in the form of a *significance trace*, first introduced by Azzalini & Bowman (1991), which plots these $p$-values as a function of smoothing parameter. It is often the case that the corresponding curve is consistent across a wide range of possible smoothing parameters, either always below, or always above, the 5% significance level. This leads to the same conclusion about the data regardless of choice of smoothing parameter. In cases where the results are not so consistent, the shape of the plot is still highly informative.

If a more precisely defined procedure is preferred then a specific method of smoothing parameter selection can be used to identify a single smoothing parameter at which significance can be evaluated. Such a procedure would, however, ignore the effect of the variability in smoothing parameter selection on the $p$-value obtained. For this reason, the full picture provided by the significance trace is very helpful. In particular, it can be useful to know whether the test is significant at *any* value of smoothing parameter.

For the test of equality on the White Spanish Onions data the $p$-values lie below 0.001 for all choices of smoothing parameter, leading to the clear conclusion that the two curves are different. If an approximate $F$-test is carried out on the residual-sums-of-squares resulting from the non-linear parametric Holliday models fitted to the data, a $p$-value of around $10^{-6}$ is obtained, leading to the same conclusion.

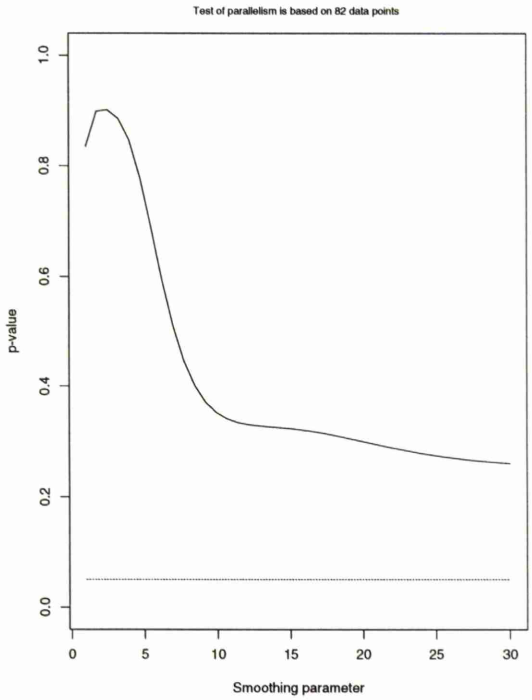
After the rejection of equality, it is natural to investigate parallelism, and Figure 4.3.3 shows the significance trace resulting from the test of the null hypothesis of parallel curves within the alternative of unrestrained curves. Here it can be seen that, for all but the lowest values of smoothing parameter (and these values are unrealistic in practice), the curve lies below the 5% line, leading to rejection of the hypothesis of parallelism. The approximate $F$-test on the Holliday models however, produces a $p$-value of 0.1521, too large to reject parallelism at the 5% level. This suggests that the nonparametric approach is picking up a feature of the data which the more rigid parametric model is unable to detect. Another look at Figure 4.1.1

Figure 4.3.3: Significance trace for the test of parallelism on White Onions data



suggests that there may be two outlying Virginia observations, at the lowest values of density and with the highest yield. The significance trace obtained when these points are removed is displayed in Figure 4.3.4. Here the conclusion is quite different, as the trace remains above the 5% level, indicating that parallel curves describe the data well. When the Holliday models are fitted to this reduced data set, a $p$-value of 0.2523 is obtained for the test of the parallel model. The nonparametric and parametric tests are therefore in agreement when these two observations are removed.

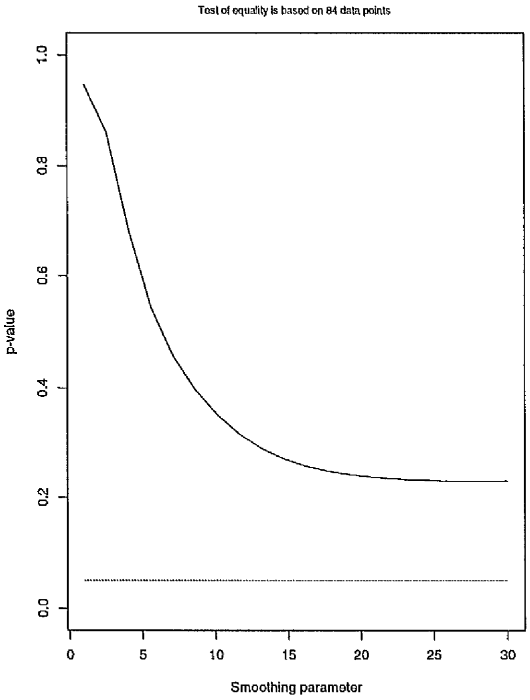
Figure 4.3.4: Significance trace for the test of parallelism on White Onions data with outliers removed



### 4.3.3 Brown Spanish Onions

We now carry out the test of equality on the Brown Spanish Onions dataset displayed in Figure 4.3.2, and the resulting significance trace is shown in Figure 4.3.5. Here the $p$-values are clearly above 0.05, leading to the conclusion that there is no difference between the yield-density relationships of Brown Spanish Onion obtained in Mount Gambier and Uraidla. The approximate $F$-test based on the parametric model returns a $p$-value of 0.0551, and is therefore in agreement with the semiparametric test, although the $p$-value in the parametric case is much closer to the 5% mark.

Figure 4.3.5: Significance trace for the test of equality on Brown Onions data



## 4.4 A simulation study

### 4.4.1 Introduction

In this section the power of the nonparametric approach is studied through simulation. Three tests are considered, namely

N1: the nonparametric approach of Section 4.2.

N2: the nonparametric approach of Section 4.2, but with $\sigma^2$ assumed to be known.

L: a linear analysis of covariance model.

The inclusion of N2 allows us to examine the effect of estimating $\sigma^2$, and L allows comparison with the common parametric analysis of covariance approach, based on straight lines, which might be used if a nonlinear covariate effect is not marked.

Three sets of design points were used, namely:

(i) both groups equally spaced on the interval (0, 1).

(ii) both groups identical, as determined by a single random sample from a Un(0,1) distribution.

(iii) each group determined by a different sample from a Un(0,1) distribution.

Each group was of size 30. Three different regression relationships were used, as defined below.

| | Group 1 | Group 2 | |
|---|---|---|---|
| (a) | $y = x$ | $y = \beta x;$ | $\beta = 1.0, 0.9, 0.8$ |
| (b) | $y = 0$ | $y = \beta sin(2\pi x);$ | $\beta = 0.0, 0.1, 0.2$ |
| (c) | $y = 0$ | $y = \beta(x^2 - x + 0.15);$ | $\beta = 0.0, 0.5, 1.0$ |

In all cases, random error was added, drawn from a $N(0, \sigma^2)$ distribution. Two values of $\sigma$ were used, namely 0.05 and 0.1. The tests were carried out with a smoothing value of $\frac{2R}{n}$, $R$ being the range of the design points, and using the Gasser-Müller (1979) smoother.

Figure 4.4.6 shows one set of sample simulation plots for each of the three functions used, for the case of the equally-spaced design. For each combination, 500 simulations were carried out.

Figure 4.4.6: A set of sample simulation plots

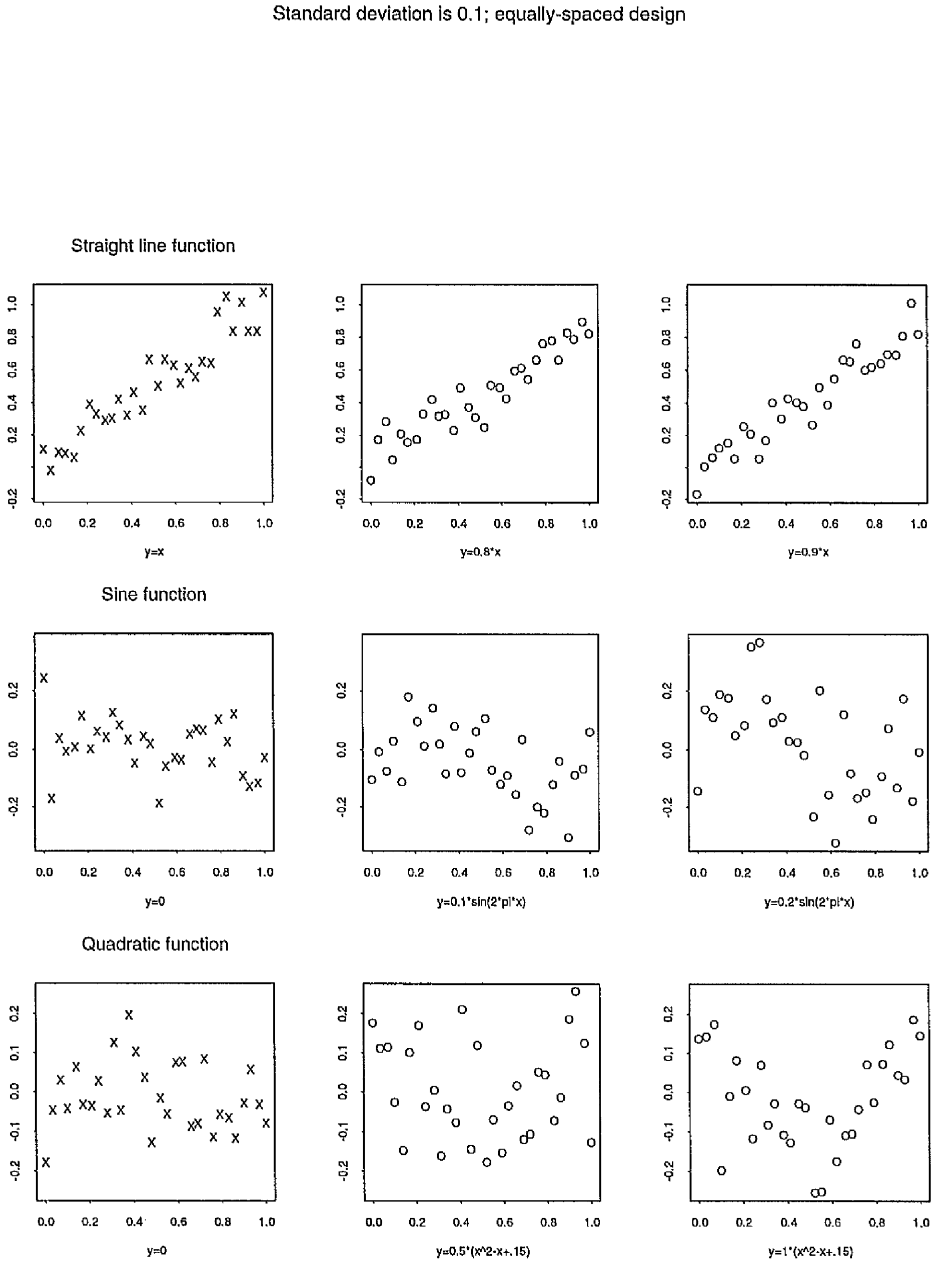


Table 4.4.1: Powers for the test of equality for the underlying linear function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 1.0$ | | | $\beta = 0.9$ | | | $\beta = 0.8$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | N1 | N2 | L | N1 | N2 | L | N1 | N2 | L |
| | Equally-spaced design | | | | | | | | |
| $\sigma = 0.05$ | .02 | .05 | .05 | .76 | .95 | .99 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .03 | .04 | .08 | .23 | .32 | .46 | .86 | .94 | .98 |
| | Un(0,1) design (same for both groups) | | | | | | | | |
| $\sigma = 0.05$ | .01 | .06 | .05 | .45 | .81 | .92 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .04 | .05 | .06 | .15 | .24 | .35 | .68 | .78 | .93 |
| | Un(0,1) design (different for each group) | | | | | | | | |
| $\sigma = 0.05$ | .01 | .04 | .05 | .44 | .75 | .92 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .02 | .07 | .05 | .17 | .20 | .40 | .64 | .74 | .94 |

### 4.4.2 Test of equality

Table 4.4.1[1] shows the powers obtained from simulations with the linear function. Since the linear regression model is the correct one in this case, it would be expected to have superior efficiency. Predictably, the results are better when the known value of $\sigma^2$ is employed, but the difference is not large, except when the error variance is small and the curves differ by a small amount. This reflects the inaccuracy of a first-order difference approach to estimating $\sigma^2$, when the underlying relationship is linear.

The powers obtained for the sine function are given in Table 4.4.2. Here, although

[1] Within each group, the column labelled N1 represents the semiparametric test, the column labelled N2 represents the semiparametric test with known $\sigma^2$, and the column labelled L represents the results of a simple linear regression.

Table 4.4.2: Powers for the test of equality for the underlying sine function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 0$ | | | $\beta = 0.1$ | | | $\beta = 0.2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | N1 | N2 | L | N1 | N2 | L | N1 | N2 | L |
| | Equally-spaced design | | | | | | | | |
| $\sigma = 0.05$ | .04 | .04 | .06 | .96 | .99 | .84 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .05 | .04 | .03 | .46 | .46 | .34 | .97 | .98 | .82 |
| | Un(0,1) design (same for both groups) | | | | | | | | |
| $\sigma = 0.05$ | .04 | .06 | .04 | .93 | .98 | .88 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .06 | .04 | .04 | .37 | .41 | .36 | .95 | .98 | .89 |
| | Un(0,1) design (different for each group) | | | | | | | | |
| $\sigma = 0.05$ | .04 | .04 | .05 | .92 | .97 | .94 | 1 | 1 | 1 |
| $\sigma = 0.1$ | .04 | .03 | .05 | .32 | .37 | .42 | .91 | .95 | .95 |

the linear regression model is clearly not the correct one, it still picks up the differences on most occasions, but is generally out-performed by the nonparametric test.

The results for the quadratic function are shown in Table 4.4.3. In this case the linear regression model is hopelessly inadequate, and is dramatically outperformed by the nonparametric test.

In Tables 4.4.1 to 4.4.3 the slightly conservative nature of the nonparametric test, due to the estimation of $\sigma^2$, is illustrated in those cases where the null hypothesis is correct and the probability of rejection should be 0.05. The properties of the test also depend of course on the suitability of the chi-squared null distribution, and this appears to be satisfactory.

Table 4.4.3: Powers for the test of equality for the underlying quadratic function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 0$ | | | $\beta = 0.5$ | | | $\beta = 1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | N1 | N2 | L | N1 | N2 | L | N1 | N2 | L |
| | Equally-spaced design | | | | | | | | |
| $\sigma = 0.05$ | .04 | .04 | .06 | .49 | .61 | .02 | 1 | 1 | .00 |
| $\sigma = 0.1$ | .06 | .03 | .05 | .14 | .16 | .04 | .46 | .61 | .04 |
| | Un(0,1) design (same for both groups) | | | | | | | | |
| $\sigma = 0.05$ | .04 | .06 | .05 | .48 | .55 | .29 | 1 | 1 | .76 |
| $\sigma = 0.1$ | .05 | .04 | .03 | .15 | .16 | .10 | .49 | .61 | .28 |
| | Un(0,1) design (different for each group) | | | | | | | | |
| $\sigma = 0.05$ | .04 | .04 | .05 | .46 | .52 | .08 | .98 | 1 | .13 |
| $\sigma = 0.1$ | .06 | .03 | .05 | .13 | .13 | .06 | .43 | .52 | .09 |

### 4.4.3 Test of parallelism

In this case, another parameter $\alpha$, describing the shift between the curves for each group, must be chosen. In all simulations $\alpha = 0.5$ has been added to the responses for the second group. For example, the linear function now becomes a comparison between $y = x$ and $y = \alpha + \beta x$ (ignoring the stochastic terms) for $\beta = 0.8$, 0.9 and 1.

Powers are given for the semiparametric test, and for a simple linear analysis of covariance. If the parameters (this time $\sigma^2$ and $\alpha$) are known, then the test reduces to a test of equality with known $\sigma^2$, and the results are as for N2 in Tables 4.4.1 to 4.4.3. Hence, there is no N2 in our tables in this section. However, the effect of estimating $\alpha$ can be assessed by comparing the N1 results from Section 4.4.2 with the

Table 4.4.4: Powers for the test of parallelism for the underlying linear function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 1.0$ | | $\beta = 0.9$ | | $\beta = 0.8$ | |
|---|---|---|---|---|---|---|
| | N1 | L | N1 | L | N1 | L |
| | Equally-spaced design | | | | | |
| $\sigma = 0.05$ | .02 | .05 | .16 | .65 | .85 | 1 |
| $\sigma = 0.1$ | .02 | .05 | .09 | .21 | .29 | .66 |
| | Un(0,1) design<br>(same for both groups) | | | | | |
| $\sigma = 0.05$ | .01 | .06 | .11 | .56 | .73 | .99 |
| $\sigma = 0.1$ | .03 | .04 | .07 | .19 | .24 | .56 |
| | Un(0,1) design<br>(different for each group) | | | | | |
| $\sigma = 0.05$ | .004 | .04 | .13 | .58 | .61 | .99 |
| $\sigma = 0.1$ | .03 | .04 | .05 | .17 | .20 | .58 |

N1 results here, as will be seen.

Table 4.4.4[2] shows the results for the linear function. As before, the linear regression model performs well when the true model is linear. The non-parametric test performs reasonably for $\beta = 0.8$ and $\sigma = 0.05$, but exhibits a drop in power elsewhere, when compared with the N1 results in Table 4.4.1. Since the essential difference between these power studies is the requirement to estimate $\alpha$, it is there that power is being lost. Indeed, additional simulations suggest that $\hat{\alpha}$ is sometimes subject to bias when the underlying model is linear. However, the nonparametric approach may hold less attractions anyway when the data are very close to linearity.

[2] Within each group, the column labelled N1 represents the semiparametric test and the column labelled L represents the results of a simple linear regression.

Table 4.4.5: Powers for the test of parallelism for the underlying sine function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 0$ | | $\beta = 0.1$ | | $\beta = 0.2$ | |
|---|---|---|---|---|---|---|
| | N1 | L | N1 | L | N1 | L |
| | Equally-spaced design | | | | | |
| $\sigma = 0.05$ | .05 | .05 | .97 | .94 | 1 | 1 |
| $\sigma = 0.1$ | .06 | .07 | .42 | .49 | .97 | .93 |
| | Un(0,1) design<br>(same for both groups) | | | | | |
| $\sigma = 0.05$ | .05 | .06 | .95 | .96 | 1 | 1 |
| $\sigma = 0.1$ | .02 | .05 | .40 | .48 | .96 | .96 |
| | Un(0,1) design<br>(different for each group) | | | | | |
| $\sigma = 0.05$ | .05 | .05 | .93 | .98 | 1 | 1 |
| $\sigma = 0.1$ | .06 | .03 | .32 | .55 | .93 | .98 |

The powers for the sine function are given in Table 4.4.5. Here the semiparametric test performs satisfactorily, and the estimates of $\alpha$ are better than in the case of the linear function. The powers are consequently comparable with those for N1 in Table 4.4.2.

Table 4.4.6 has the powers for the quadratic function. As with the test of equality, the parametric test is generally inadequate, and is outperformed by the semiparametric test. Again, good estimates of $\alpha$ lead to N1 results comparable to Table 4.4.3.

Table 4.4.6: Powers for the test of parallelism for the underlying quadratic function, based on 500 simulations and with $n_i = 30$, at the 5% level.

| | $\beta = 0$ | | $\beta = 0.5$ | | $\beta = 1$ | |
|---|---|---|---|---|---|---|
| | N1 | L | N1 | L | N1 | L |
| | Equally-spaced design | | | | | |
| $\sigma = 0.05$ | .05 | .05 | .52 | .03 | 1 | .00 |
| $\sigma = 0.1$ | .06 | .07 | .14 | .04 | .50 | .02 |
| | Un(0,1) design<br>(same for both groups) | | | | | |
| $\sigma = 0.05$ | .05 | .06 | .51 | .40 | 1 | .83 |
| $\sigma = 0.1$ | .02 | .05 | .15 | .14 | .51 | .41 |
| | Un(0,1) design<br>(different for each group) | | | | | |
| $\sigma = 0.05$ | .05 | .05 | .47 | .06 | .98 | .08 |
| $\sigma = 0.1$ | .06 | .03 | .14 | .06 | .48 | .06 |

## 4.5 Discussion

The examples and simulation study show that the nonparametric approach offers a very useful alternative to parametric methods of analysis of covariance. Parametric methods will clearly be preferable where there is confidence that the adopted model provides an accurate description of the data. The flexibility of the nonparametric approach is particularly effective when a parametric model is clearly inappropriate, or of doubtful validity.

Much of the work in this chapter can be found in the paper by Young & Bowman (1995).

# Chapter 5

# Analysis of covariance with binary response data

## 5.1 Introduction

This chapter extends the ideas presented in Chapter 4 by considering binary response data, where typically the interest is in modelling the probability that an event occurs. Such models are often used in biostatistics, for example, to model the probability of suffering from a disease.

This situation has already been considered in Section 3.2.2, where the idea of a reference band for equality was applied to binary response data. For the continuous-response regression model, Chapter 4 proposed a formal method of inference, to support the graphical techniques presented in Chapter 3. In the same way, in this chapter a formal test is derived to accompany the reference bands for binary response data.

A traditional parametric approach to such data is to fit a linear logistic regression, which is an example of a generalized linear model. McCullagh and Nelder (1989) give a good introduction to this class of model. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent observed pairs where, for each $i$, $Y_i$ is a random binary response variable, and $X_i$ is a covariate value. Let $(X, Y)$ denote a generic member of the sample. In generalized linear modelling it is usual to model as linear a transformation of the regression function $\mu(x) = E(Y|X = x) = P(Y = 1|X = x) = p_x$, say, giving

$$g(p_x) = \alpha + \beta x$$

$g$ is the *link* function, and in linear logistic regression, is taken to be the logit function

$$g(p_x) = log_e \left( \frac{p_x}{1 - p_x} \right).$$

In other words, with Y taking the value 0 or 1, the model is

$$P(Y = 1|X = x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}.$$

Other common link functions are the probit function, and the complementary log-log function.

This chapter, as in Section 3.2.2, and in an analogous manner to Chapter 4, addresses the case where the data consist of a binary response, and two covariates, one of which is continuous, and the other discrete. The discrete covariate is therefore a grouping factor, and interest centres on comparing the groups. A nonparametric approach to the logistic regression is adopted, and a formal test of equality of groups derived. The non-Normality of the errors in this situation means that the quadratic form theory of Chapter 4 cannot be applied here. Instead the first two moments of the proposed test statistic are derived exactly, and these are used to provide an approximate null distribution for the test. In Section 5.3 the test is applied to the Pregnancy data introduced in Chapter 3. The power of the proposed test is assessed through some simulation studies, and contrasted with a parametric approach, in Section 5.4. Lastly, some final conclusions are drawn in Section 5.5.

## 5.2 A nonparametric test statistic

The motivation for the proposed test statistic is the same as in Section 4.2.1, even though, as will be shown, the quadratic form theory cannot be applied here.

Before the test statistic is introduced, the model and smoothing method employed are clarified. From Section 5.1, the linear logistic model is

$$P(Y=1|X=x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = p_x$$

for a binary response variable Y, and continuous covariate X. This gives

$$log_e\left(\frac{p_x}{1-p_x}\right) = \alpha + \beta x. \qquad (5.2.1)$$

Copas (1983) showed that the parametric assumption in (5.2.1) can be relaxed, and instead the binary responses can be smoothed, and $p_x$ modelled nonparametrically. In that case, the model can be expressed as

$$\hat{p}_x = \sum_{i=1}^{n} w_i y_i$$

where the $w_i$ are kernel weights. Copas used the Nadaraya-Watson (1964) estimator, as defined in Chapter 1, but, as in earlier chapters, bias considerations dictate that the methods of local line fitting, or Gasser-Müller (1979) estimation, are more suitable.

Fan, Heckman and Wand (1995) use local line fitting in this context, and the method employed here, where $p_x$ is estimated directly, is equivalent to applying their techniques with the link equal to the identity.

This chapter considers the case where there are $p$ groups, and for each group $i$, a set of observations of the form $(x_1, y_1), \ldots, (x_{n_i}, y_{n_i})$ is available. In addition, let $n = \sum_{i=1}^{p} n_i$, and $\mathbf{y}$ be the $n$-vector of responses. The proposed test statistic, in matrix notation, is

$$TS = \mathbf{y}^T(\mathbf{S_d} - \mathbf{S_s})^T(\mathbf{S_d} - \mathbf{S_s})\mathbf{y} \qquad (5.2.2)$$

where $\mathbf{S_s}$ and $\mathbf{S_d}$ are smoother matrices corresponding to the hypotheses of equality and unconstrained curves, respectively. Hence, equation (5.2.2) consists of the sum of squared differences between a single, common logistic curve, and an unconstrained curve for each group. This comparison is made at all points in the design space. In this respect, the test statistic is similar in nature to those introduced in Chapter 4.

Large values of this test statistic would lead to a rejection of the null hypothesis of equality.

Fan, Heckman and Wand (1995) show that, if local line fitting is adopted, the bias in smoothing, for a single curve, is, asymptotically,

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\}\eta''(x)$$

where $\eta = g(\mu(x))$. Here $g$, the link function, is the identity, and so the asymptotic bias is

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\}\mu''(x)$$

or alternatively,

$$\frac{h^2}{2}\left\{\int t^2 k(t)dt\right\}p_x''$$

in the earlier notation.

This is the same form as the bias in previous cases, where the response was continuous, so the result is the same; namely, that the bias does not depend on the density of the design points, but on the shape of the true curve. This result ensures that in equation 5.2.2, under the null hypothesis of equal curves, any bias terms cancel, at least asymptotically.

### 5.2.1 Obtaining a null distribution

In Chapter 4, quadratic form theory was used to match the first three moments of the distribution of the proposed test statistic to a suitable chi-squared distribution. This chi-squared distribution was then used as a null distribution to calculate a $p$-value for the test.

The quadratic form theory applies to terms of the form

$$\mathbf{z}^T\mathbf{A}\mathbf{z}$$

where $E(\mathbf{z}) = \mathbf{0}$, $\mathbf{A}$ is an $n \times n$ symmetric matrix, and $\mathbf{z}$ consists of independent Normal random variables. The assumption of Normal errors in the models of Chapter

4 allowed the quadratic form results to be applied. With a binary response variable however, Normality of errors does not hold, and so another approach must be adopted.

Le Cessie and van Houwelingen (1991) proposed a goodness-of-fit test for binary response models, based on nonparametric kernel smoothing. In order to obtain a null distribution, a two-moment approximation was used, and matched to a chi-squared distribution. A similar strategy is employed here, in the absence of the quadratic form results.

The procedure therefore involves deriving the mean and variance of the distribution of the proposed test statistic, then finding $a$ and $b$ such that an $a\chi^2(b)$ distribution has the same first two moments. This chi-squared distribution is then used as a null distribution to provide a $p$-value for the test.

### 5.2.2 Expectation of the test statistic

Consider first, the test statistic defined by equation (5.2.2) and reproduced below:

$$TS = \mathbf{y}^T(\mathbf{S}_\mathrm{d} - \mathbf{S}_\mathrm{s})^T(\mathbf{S}_\mathrm{d} - \mathbf{S}_\mathrm{s})\mathbf{y}.$$

This can be simplified to $\mathbf{a}^T\mathbf{a}$, where $\mathbf{a} = (\mathbf{S}_\mathrm{d} - \mathbf{S}_\mathrm{s})\mathbf{y}$. Remember that $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ are smoother matrices corresponding to the hypotheses of equality and unconstrained curves, respectively.

To find the expectation of the test statistic, $TS$, it is necessary to evaluate $E(\mathbf{a}^T\mathbf{a})$ in the notation above. Applying the bias results of Section 5.2 gives $E(\mathbf{a}) = \mathbf{0}$, asymptotically at least, and so

$$E(TS) = E(\mathbf{a}^T\mathbf{a}) = \sum_{i=1}^{n} V(a_i).$$

Now, suppose $Cov(\mathbf{a}) = \mathbf{\Sigma}$, then

$$E(TS) = tr(\mathbf{\Sigma})$$

and

$$\mathbf{\Sigma} = (\mathbf{S_d} - \mathbf{S_s})^T Cov(\mathbf{y})(\mathbf{S_d} - \mathbf{S_s}).$$

At this point it is necessary to introduce some further notation. Let $\mathbf{S_s} = [s_{ij}]_{n\times n}$ and $\mathbf{S_d} = [d_{ij}]_{n\times n}$ and so

$$a_i = \sum_{j=1}^{n} (d_{ij} - s_{ij})y_j$$

where each $y_j$ is a realisation of an independent $\mathrm{Bi}(1, p_{x_j})$ random variable. Hence

$$Cov(\mathbf{y}) = diag\left(p_{x_1}(1 - p_{x_1}), \ldots, p_{x_n}(1 - p_{x_n})\right)$$

and so the expectation of the test statistic can be expressed as

$$E(TS) = \sum_{i=1}^{n}\sum_{j=1}^{n} (d_{ij} - s_{ij})^2 p_{x_j}(1 - p_{x_j}). \qquad (5.2.3)$$

The true $p_{x_j}$ are unknown, and are replaced by their estimates $\hat{p}_{x_j}$.

### 5.2.3 Variance of the test statistic

The next step is to derive the variance of the test statistic, and use this, together with the mean calculated in Section 5.2.2, to calculate an approximate null distribution.

Retaining the notation of the previous section, the test statistic is of the form $\mathbf{a}^T\mathbf{a}$, where $\mathbf{a} = (\mathbf{S_d} - \mathbf{S_s})\mathbf{y}$. It is required to calculate

$$V\left(\mathbf{a}^T\mathbf{a}\right) = V\left(\sum_{i=1}^{n} a_i^2\right)$$

where $E(\mathbf{a}) = 0$ and $Cov(\mathbf{a}) = \mathbf{\Sigma}$. Also, from before, $a_i = \sum\limits_{v=1}^{n}(d_{iv} - s_{iv})y_v$.

It is required to find

$$V\left(\sum_{i=1}^{n} a_i^2\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(a_i^2, a_j^2\right).$$

Now,

$$Cov\left(a_i^2, a_j^2\right) = E\left(a_i^2 a_j^2\right) - E\left(a_i^2\right)E\left(a_j^2\right).$$

To simplify matters, let $c_{iv} = d_{iv} - s_{iv}$, so that $a_i = \sum\limits_{v=1}^{n} c_{iv}y_v$. Hence

$$Cov\left(a_i^2, a_j^2\right) = E\left\{\left(\sum_{k=1}^{n} c_{ik}y_k\right)^2 \left(\sum_{l=1}^{n} c_{jl}y_l\right)^2\right\} - E\left\{\left(\sum_{k=1}^{n} c_{ik}y_k\right)^2\right\} E\left\{\left(\sum_{l=1}^{n} c_{jl}y_l\right)^2\right\}.$$

Each $Y_r$ is an independent $\mathrm{Bi}(1, p_{x_r})$ random variable, giving $E(Y_r^p) = p_{x_r}$. Using this result, and after some algebraic work, the covariance can be re-expressed as

$$\begin{aligned}
Cov\left(a_i^2, a_j^2\right) &= \sum_{r=1}^{n} c_{ir}^2 c_{jr}^2 p_{x_r} + 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s}\left(c_{ir}^2 c_{jr}c_{js} + c_{ir}c_{is}c_{jr}^2\right)p_{x_r}p_{x_s} \\
&+ 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s} c_{ir}c_{is}c_{jr}c_{js}p_{x_r}p_{x_s} + 4\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}}_{r\neq s,t;s\neq t} c_{ir}c_{is}c_{jr}c_{jt}p_{x_r}p_{x_s}p_{x_t} \\
&- \sum_{r=1}^{n} c_{ir}^2 c_{jr}^2 p_{x_r}^2 - 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s}\left(c_{ir}^2 c_{jr}c_{js} + c_{ir}c_{is}c_{jr}^2\right)p_{x_r}^2 p_{x_s} \\
&- 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s} c_{ir}c_{is}c_{jr}c_{js}p_{x_r}^2 p_{x_s}^2 - 4\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}}_{r\neq s,t;s\neq t} c_{ir}c_{is}c_{jr}c_{jt}p_{x_r}^2 p_{x_s}p_{x_t}
\end{aligned}$$

This simplifies to

$$\begin{aligned} Cov\left(a_i^2, a_j^2\right) &= \sum_{r=1}^{n} c_{ir}^2 c_{jr}^2 p_{x_r}\left(1-p_{x_r}\right) \\ &+ 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s}\left(c_{ir}^2 c_{jr} c_{js} + c_{ir} c_{is} c_{jr}^2\right) p_{x_r}\left(1-p_{x_r}\right) p_{x_s} \\ &+ 2\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}}_{r\neq s} c_{ir} c_{is} c_{jr} c_{js} p_{x_r} p_{x_s}\left(1-p_{x_r} p_{x_s}\right) \\ &+ 4\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}}_{r\neq s,t;s\neq t} c_{ir} c_{is} c_{jr} c_{jt} p_{x_r}\left(1-p_{x_r}\right) p_{x_s} p_{x_t} \end{aligned} \tag{5.2.4}$$

Returning to the test statistic, $TS = \mathbf{a}^T\mathbf{a}$, it follows that

$$V(TS) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(a_i^2, a_j^2\right) \tag{5.2.5}$$

where $Cov\left(a_i^2, a_j^2\right)$ is as given in equation (5.2.4). As was the case with the expectation of the test statistic, the true $p_{x_r}$ are unknown, and so these are replaced by their estimates $\hat{p}_{x_r}$.

### 5.2.4 Using the two-moment approximation

Now that the mean, (5.2.3), and variance, (5.2.5), of the test statistic have been derived, the only remaining step is to match these with an appropriate chi-squared distribution.

An $a\chi^2(b)$ distribution has mean $ab$ and variance $2a^2b$. Given the mean $E$, and variance $V$, of the test statistic, this gives

$$b = \frac{2E^2}{V};\, a = \frac{E}{b} = \frac{V}{2E}.$$

Hence, a $p$-value for the test of equality, in the binary response case, can be obtained as follows:

$$p = P(TS > Obs) \approx P(a\chi^2(b) \text{ r.v.} > Obs) = 1 - P\left(\chi^2(b) \text{ r.v.} < \frac{Obs}{a}\right)$$

where $Obs$ is the observed value of the test statistic.

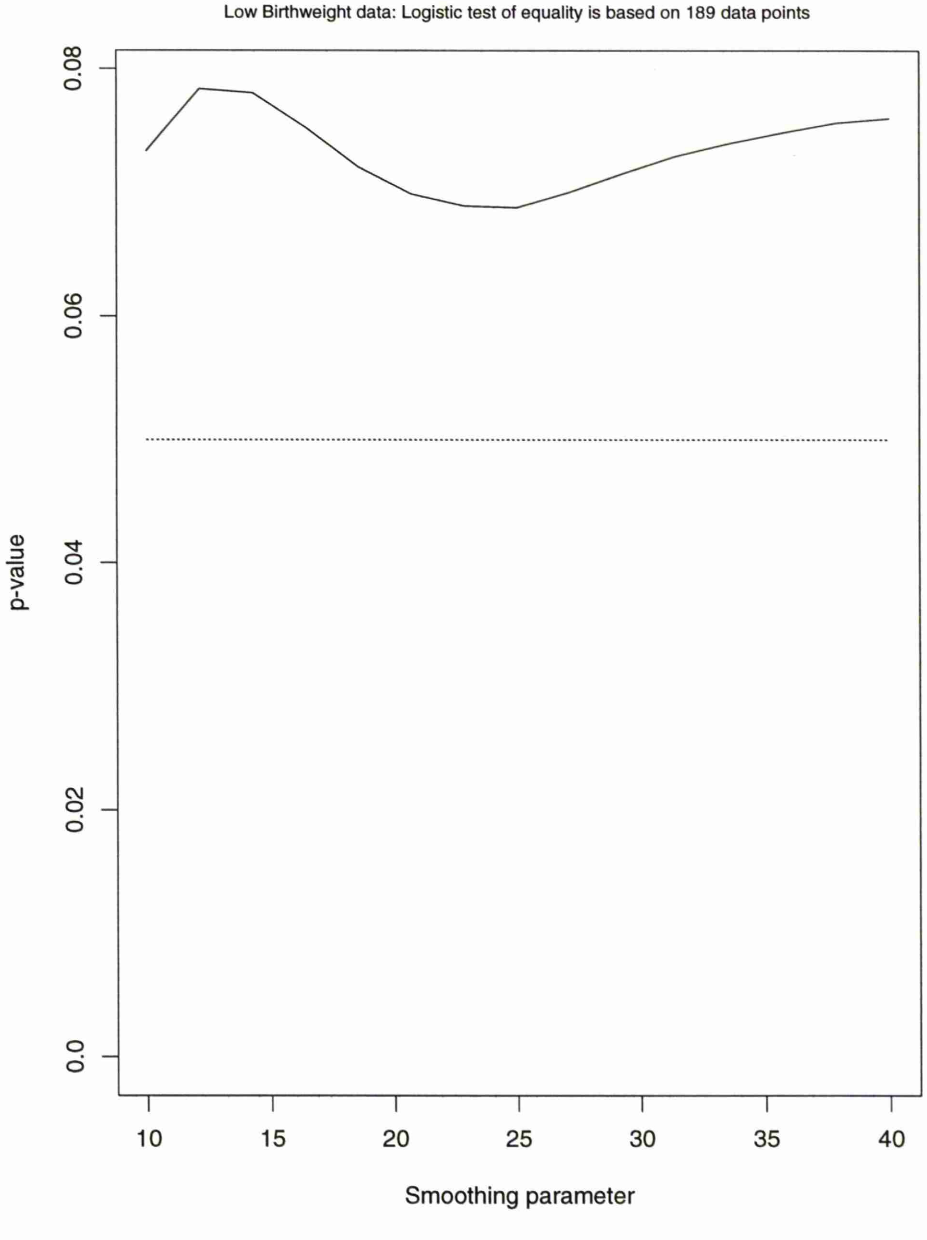
## 5.3 A practical example

In Section 3.2.2, a birthweight example was introduced. This dataset consists of observations on 189 mothers, collected in Massachussetts. The response consists of a low birthweight indicator (0 represents $\geq 2500gms$, 1 represents $< 2500gms$), with the mother's weight at her last menstrual period as a covariate. The mothers are grouped according to whether they did, or did not, smoke during pregnancy. It is of obvious interest to establish whether there is evidence that smoking during pregnancy affects the probability of having a low birthweight baby. In Section 3.2.2, a reference band was produced for these data, and it suggested there was some evidence of a slight difference between the two groups, over a limited range of the covariate space.

In this section, that claim is investigated formally, using the test derived earlier. As in Chapter 4, the test results are presented in the form of a significance trace, which reports the $p$-value obtained for each of a range of smoothing parameters.

The significance trace for this example is given in Figure 5.3.1. In the figure, all of the $p$-values are just over 0.05, indicating a non-significant result at the 5% level. The closeness of the $p$-values to 0.05 reflects the marginal nature of the difference

Figure 5.3.1: Significance trace for equality of nonparametric regression with binary data



suggested by the reference band in Figure 3.2.4 of Chapter 3.

An approximate likelihood ratio test, based on a parametric logistic model with $logit(p_x) = \alpha + \beta x$, returns a $p$-value of 0.042. Plots of the smoothed data on the logistic scale suggest a quadratic relationship, i.e. $logit(p_x) = \alpha + \beta x + \gamma x^2$, or even a cubic, may be more appropriate. If a quadratic model is assumed, the $p$-value is 0.069, and a cubic function returns a $p$-value of 0.113. This would suggest that the nonparametric approach is giving useful results, without the need to choose a rigid parametric model.

## 5.4 Some simulation studies

In this section the power, and size, of the proposed test of equality are assessed by simulation. In each case, the results of the nonparametric test are compared to those of a parametric approach, based on a Generalised Likelihood Ratio Test. This approximate test is based on fitting linear logistic models to the simulated data.

In these studies, a Normal null distribution was also tested, as an alternative to the chi-squared distribution. Again, the two-moment approximation was used, and a Normal with mean and variance matching the test statistic was taken as the null distribution.

### 5.4.1 Power

To assess the power first of all, data were generated from the following two models

$$\text{Group One}: \ \text{logit}(p_x) = -3 + 3x + x^2$$

$$\text{Group Two}: \ \text{logit}(p_x) = -3 + 7x$$

Figure 5.4.2 illustrates these two models, on the logistic scale. Notice that over the covariate space used, 0 to 1, the quadratic function (Group One) appears reasonably linear.

Overall, 1000 simulations were carried out, using two sets of covariates. In one set, there were 25 observations in each group, and in the other, there were 50 in each group. In both cases, the covariate values were equally-spaced over the range (0, 1).

To give an indication of the effect of the choice of smoothing parameter, four values of $h$ were used, namely 0.05, 0.1, 0.15 and 0.2. Figure 5.4.3 shows one set of sample simulation data for each combination. Superimposed are the smooth estimates of the curves, for each choice of smoothing parameter.

Finally, the tests were carried out at four significance levels, namely 10%, 5%, 2.5% and 1%.

The results for the case with 25 observations in each group are shown in Table 5.4.1. As can be seen from the table, the parametric test very slightly outperforms the nonparametric one in general, although that might be expected from the almost linear nature of the model for Group One. Nevertheless, the nonparametric test

Figure 5.4.2: Plot of logistic models for the simulation study

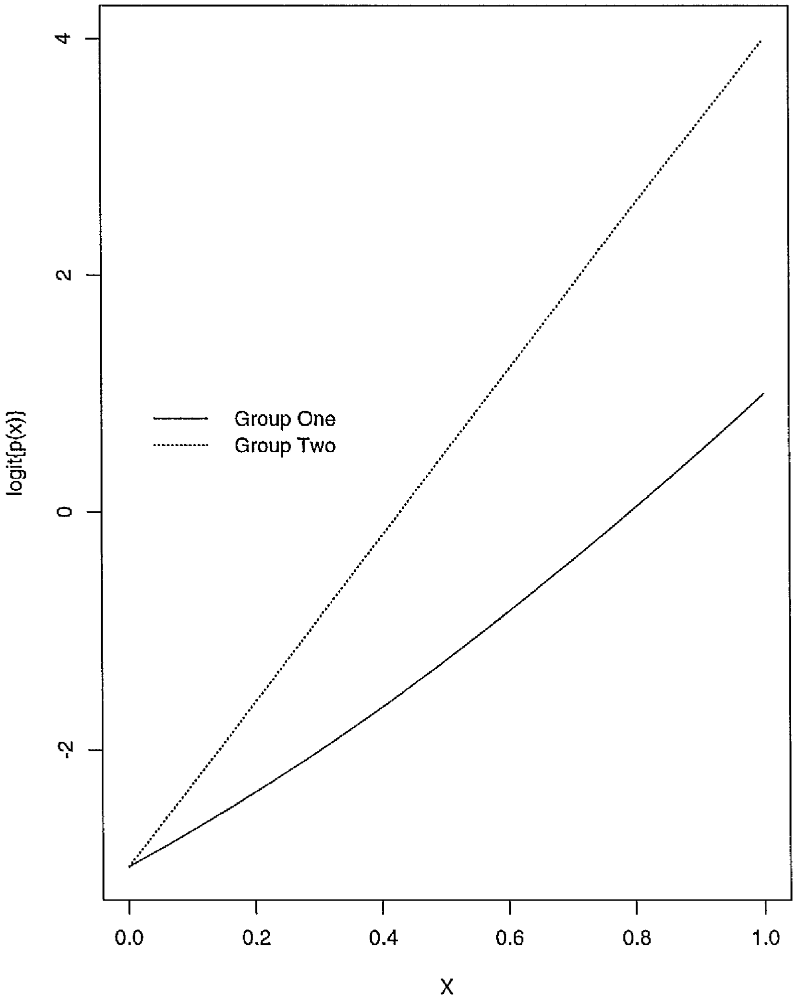


Figure 5.4.3: A set of sample simulation plots

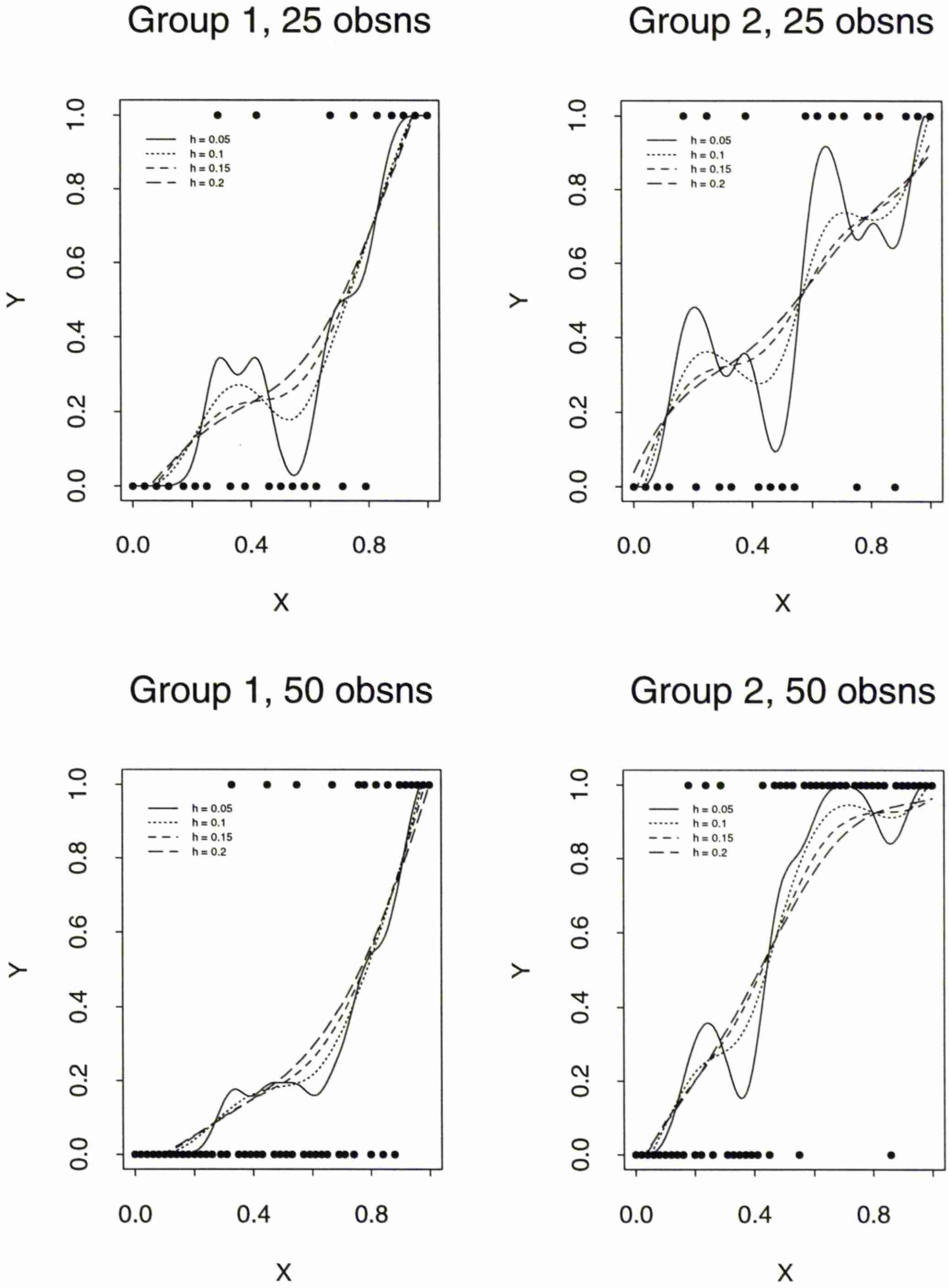


Table 5.4.1: Powers for the test of equality based on 1000 simulations with $n_i = 25$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.745 | 0.627 | 0.506 | 0.361 |
| | $h$=0.1 | 0.742 | 0.625 | 0.521 | 0.385 |
| | $h$=0.15 | 0.764 | 0.649 | 0.536 | 0.404 |
| | $h$=0.2 | 0.773 | 0.655 | 0.536 | 0.412 |
| Normal null | $h$=0.05 | 0.761 | 0.681 | 0.611 | 0.518 |
| | $h$=0.1 | 0.752 | 0.677 | 0.620 | 0.557 |
| | $h$=0.15 | 0.771 | 0.700 | 0.649 | 0.592 |
| | $h$=0.2 | 0.778 | 0.712 | 0.655 | 0.589 |
| Parametric Test | | 0.789 | 0.692 | 0.593 | 0.459 |

performs well in comparison, particularly with the Normal null. It exhibits reasonable power over the range of smoothing parameters chosen, and indeed outperforms the parametric test on occasion. With regard to the choice of null distribution for the nonparametric test, the Normal and chi-squared distributions perform similarly, although the former does better at the lower significance levels.

Table 5.4.2 shows the results for the case where there are 50 observations in each group. Predictably, the powers are much higher here, since there are twice the number of observations in each group. The nonparametric approach performs very creditably, even doing better than the parametric test in a number of cases. In addition, a similar pattern to Table 5.4.1 is observed for the chi-squared and Normal null distributions.

It was noted in Chapter 4 that the parametric approach was outperformed by the nonparametric tests when the linear model was inadequate. A second simulation study, which follows, indicates that this is also clearly the case in this setting.

Table 5.4.2: Powers for the test of equality based on 1000 simulations with $n_i = 50$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.930 | 0.875 | 0.811 | 0.721 |
| | $h$=0.1 | 0.958 | 0.916 | 0.860 | 0.789 |
| | $h$=0.15 | 0.975 | 0.945 | 0.898 | 0.827 |
| | $h$=0.2 | 0.971 | 0.933 | 0.889 | 0.818 |
| Normal null | $h$=0.05 | 0.934 | 0.903 | 0.868 | 0.823 |
| | $h$=0.1 | 0.960 | 0.934 | 0.914 | 0.885 |
| | $h$=0.15 | 0.975 | 0.966 | 0.944 | 0.914 |
| | $h$=0.2 | 0.973 | 0.952 | 0.935 | 0.910 |
| Parametric Test | | 0.967 | 0.942 | 0.897 | 0.838 |

In this second simulation, the models are

$$\text{Group One}: \ \text{logit}(p_x) = -1.5 + 16(x - 0.5)^2$$

$$\text{Group Two}: \ \text{logit}(p_x) = 0$$

In all other respects, the simulation study is the same as before. The results for the case with 25 observations in each group are given in Table 5.4.3. Here the linear logistic model is totally inadequate, and the parametric test is clearly outperformed by the nonparametric approach. It is noticeable that the choice of smoothing parameter is more important in this example, with $h$=0.2 (a rather large choice, in any case) performing poorly in comparison to the other values.

Table 5.4.4 displays the results for the simulations with 50 observations in each group. Again, the increased sample size results in increased power in general, although the parametric test still performs very poorly. As in Table 5.4.3, the highest value of

Table 5.4.3: A second simulation study for the test of equality based on 1000 simulations with $n_i = 25$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.500 | 0.365 | 0.254 | 0.155 |
| | $h$=0.1 | 0.495 | 0.347 | 0.223 | 0.118 |
| | $h$=0.15 | 0.423 | 0.266 | 0.180 | 0.101 |
| | $h$=0.2 | 0.364 | 0.191 | 0.107 | 0.042 |
| Normal null | $h$=0.05 | 0.517 | 0.419 | 0.341 | 0.258 |
| | $h$=0.1 | 0.508 | 0.416 | 0.333 | 0.252 |
| | $h$=0.15 | 0.438 | 0.337 | 0.261 | 0.207 |
| | $h$=0.2 | 0.372 | 0.269 | 0.190 | 0.140 |
| Parametric Test | | 0.083 | 0.038 | 0.018 | 0.006 |

Table 5.4.4: A second simulation study for the test of equality based on 1000 simulations with $n_i = 50$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.691 | 0.585 | 0.477 | 0.336 |
| | $h$=0.1 | 0.684 | 0.571 | 0.450 | 0.303 |
| | $h$=0.15 | 0.665 | 0.501 | 0.376 | 0.252 |
| | $h$=0.2 | 0.572 | 0.424 | 0.271 | 0.148 |
| Normal null | $h$=0.05 | 0.702 | 0.630 | 0.562 | 0.485 |
| | $h$=0.1 | 0.695 | 0.630 | 0.560 | 0.482 |
| | $h$=0.15 | 0.677 | 0.581 | 0.496 | 0.432 |
| | $h$=0.2 | 0.587 | 0.496 | 0.423 | 0.323 |
| Parametric Test | | 0.092 | 0.043 | 0.019 | 0.006 |

$h$ returns relatively disappointing results.

In the second simulation study, as with the first, the Normal null distribution gives better results than the chi-squared null. The next section however, which looks at the size of the proposed test, suggests that the chi-squared approach may be preferable in practice.

### 5.4.2 Size

In this section, the size of the proposed test is assessed through simulations in which both groups are generated from the same model, namely

$$\text{logit}(p_x) = -3 + 6x.$$

As in the previous section, this was done for both 25 and 50 observations in each group.

Looking at 25 observations first of all, the results of those simulations are given in Table 5.4.5. In these examples, the values should of course mirror the significance level. The nonparametric test with a chi-squared null produces very satisfactory results, and seems to be giving a good approximation to the true null distribution. The Normal null on the other hand, has significance levels higher than the nominal ones, particularly at the smaller values. This would indicate that a Normal null distribution is not fitting adequately at the tails. Also, it is noticeable that the parametric test also produces higher values than expected; often higher indeed than the nonparametric test with a Normal null. This reflects the approximate nature of

Table 5.4.5: Assessing the size of the test of equality based on 1000 simulations with $n_i = 25$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.137 | 0.064 | 0.025 | 0.009 |
| | $h$=0.1 | 0.092 | 0.059 | 0.033 | 0.017 |
| | $h$=0.15 | 0.086 | 0.046 | 0.023 | 0.009 |
| | $h$=0.2 | 0.106 | 0.048 | 0.015 | 0.006 |
| Normal null | $h$=0.05 | 0.148 | 0.085 | 0.056 | 0.030 |
| | $h$=0.1 | 0.104 | 0.071 | 0.059 | 0.038 |
| | $h$=0.15 | 0.091 | 0.057 | 0.046 | 0.030 |
| | $h$=0.2 | 0.108 | 0.079 | 0.051 | 0.027 |
| Parametric Test | | 0.130 | 0.079 | 0.049 | 0.025 |

the Generalised Likelihood Ratio Test.

Lastly, the figures for the simulations with 50 observations in each group are given in Table 5.4.6. The pattern here is similar to before, although the parametric test performs much more satisfactorily. For the nonparametric test, the chi-squared null is again the better of the two. The use of different significance levels allows assessment of the fit of the null distribution at the tails, and Tables 5.4.5 and 5.4.6 would suggest that the chosen chi-squared null distribution models the true distribution of the test statistic reasonably well. The Normal null does not fit as well, and gives too many ”false positive” results. For that reason, despite the slight superiority of the Normal null in the power studies of Section 5.4.1, it would appear that the chi-squared distribution is the better option.

Table 5.4.6: Assessing the size of the test of equality based on 1000 simulations with $n_i = 50$

| | | Significance Level | | | |
|---|---|---|---|---|---|
| | | 10% | 5% | 2.5% | 1% |
| $\chi^2$ null | $h$=0.05 | 0.109 | 0.061 | 0.023 | 0.008 |
| | $h$=0.1 | 0.100 | 0.055 | 0.027 | 0.012 |
| | $h$=0.15 | 0.089 | 0.035 | 0.025 | 0.011 |
| | $h$=0.2 | 0.088 | 0.039 | 0.023 | 0.009 |
| Normal null | $h$=0.05 | 0.117 | 0.080 | 0.052 | 0.027 |
| | $h$=0.1 | 0.108 | 0.074 | 0.054 | 0.035 |
| | $h$=0.15 | 0.094 | 0.062 | 0.037 | 0.029 |
| | $h$=0.2 | 0.090 | 0.058 | 0.039 | 0.028 |
| Parametric Test | | 0.107 | 0.053 | 0.026 | 0.012 |

## 5.5 Discussion

The examples and simulation studies show that the nonparametric approach offers a flexible alternative to parametric analysis of covariance, when the response is binary. This extends the work of Chapter 4, which dealt with the standard regression model. As was illustrated in that chapter, the nonparametric approach is particularly useful if the chosen parametric model is clearly inappropriate, as demonstrated by the simulation studies of Section 5.4.

The work presented in this Chapter can be extended to more than one covariate. Fan, Heckman and Wand (1995) report that the bias results hold for several covariates, and so the methods here could be extended, for example, to compare surfaces. Smoothing with more covariates is covered in Chapter 6, focussing on the standard regression models.

# Chapter 6

# Introducing more covariates

## 6.1 Introduction

This chapter looks at extending the previous results to situations where there are several continuous covariates. The nonparametric models considered so far are all examples of the class known as Generalised Additive Models. A thorough introduction to this class of models is given by Hastie & Tibshirani (1990). The Generalised Additive Model takes the form

$$y_j = \alpha + \sum_{i=1}^{v} g_i(x_{ij}) + \varepsilon_j$$

for $j = 1, \ldots, n$ sets of observations, each of which consists of a single continuous response variable, $y$, and $v$ continuous explanatories, $x_1, \ldots, x_v$. The errors are distributed with mean 0 and variance $\sigma^2$. The traditional parametric approach would be to model each $g_i$ as $g_i = \beta_i x_{ij}$ (or a polynomial of higher order). The Generalised Additive Model approach allows for any number of the $g_i$'s to be smooth, nonparametric functions. Thus a semiparametric modelling structure is created, where models containing both parametric and smooth components can be fitted. In order to fit these models, Hastie & Tibshirani (1990) use a technique called *backfitting.*

When more covariates are introduced into the model, the topic of inference becomes less straightforward. As indicated in Chapter 4, Hastie & Tibshirani (1990) advocate an heuristic $F$-test approach, mimicking that of the parametric ANOVA-ANCOVA test. It is possible however, to apply some of the methods already introduced in this thesis, to more complicated models involving several covariates.

This chapter begins by considering existing literature on the subject, and using the Stack Loss data, introduced earlier, to illustrate one possible test, applicable to multiple regression models with a single nonparametric component.

Moving on from smoothing in a single covariate, the topic of bivariate smoothing is considered. A particular application in image analysis is discussed, and a nonparametric test of equality derived. This work is then extended, in Section 6.4, to the more general bivariate analysis of covariance, and a test of equality is derived, using the same quadratic form theory as in Chapter 4.

Some final comments about this chapter, and the thesis in general, are given in Section 6.5.

## 6.2 Testing a model with a single nonparametric component

Previous chapters dealing with nonparametric regression have concentrated on the case where there is a single continuous covariate. In many practical applications however, it is common to want to identify how a response variable $y$ is related to a number of continuous explanatory variables. If there are $n$ sets of observations, and $v$ explanatory variables, then the data for the $j$-th case $(j = 1, \ldots, n)$ are of the form $(x_{1j}, x_{2j}, \ldots, x_{vj}, y_j)$.

The Normal linear model for such regression data would take the form

$$y_j = \alpha + \sum_{i=1}^{v} \beta_i x_{ij} + \varepsilon_j,$$

where the $\varepsilon_j$'s are independent N(0,$\sigma^2$) random variables. Suppose it was required to test the parametric assumption in one of the covariates. That is, the question of interest is whether the term $\beta_i x_i$ is sufficient to model the relationship of the response to the covariate $x_i$, or whether an unspecified smooth function $g(x_i)$ is required. This involves testing a parametric model within a Generalised Additive Model with a single nonparametric component.

Azzalini & Bowman (1993) proposed a pseudo-likelihood ratio test, which can be applied in this case. This test is based on the residuals from the parametric linear regression, which are smoothed against the covariate of interest, $x_i$. If the parametric model holds for that covariate, there ought to be no pattern in the residuals, and this would be reflected in the smoothed curve lying close to zero. A discernible pattern in the residuals would suggest that the parametric assumption is invalid, and a smooth term, $g(x_i)$, is perhaps required. Note that under the null hypothesis, $E(\mathbf{e}) = \mathbf{0}$, and so the bias in smoothing is zero.

Azzalini & Bowman's test is based on a contrast between the residual sum of squares from the regression ($R_0 = \mathbf{e}^T\mathbf{e}$) and $R_1$. $R_1$ is given by the sum of the squared differences between the residuals and the smoothed residuals. If there is no pattern in the residuals, $R_1$ and $R_0$ should be similar in value, and the value of the test statistic small. Larger values of the test statistic would lead to rejection of the null hypothesis.

Formally, let $\mathbf{S}_h$ be a smoother matrix for the residuals $\mathbf{e}$, for a given smoothing parameter $h$. The method of local-line fitting is used here. Then

$$R_1 = (\mathbf{e} - \mathbf{S}_h\mathbf{e})^T(\mathbf{e} - \mathbf{S}_h\mathbf{e}) = \mathbf{e}^T(\mathbf{I}_n - \mathbf{S}_h)^T(\mathbf{I}_n - \mathbf{S}_h)\mathbf{e} = \mathbf{e}^T\mathbf{M}_1\mathbf{e}$$

where $\mathbf{M}_1 = (\mathbf{I}_n - \mathbf{S}_h)^T(\mathbf{I}_n - \mathbf{S}_h)$.

The pseudo-likelihood ratio test statistic takes the form

$$TS = \frac{R_0 - R_1}{R_1} = \frac{\mathbf{e}^T\mathbf{e} - \mathbf{e}^T\mathbf{M}_1\mathbf{e}}{\mathbf{e}^T\mathbf{M}_1\mathbf{e}} = \frac{\mathbf{e}^T(\mathbf{I}_n - \mathbf{M}_1)\mathbf{e}}{\mathbf{e}^T\mathbf{M}_1\mathbf{e}}$$

which leads to the following expression for a p-value for the test:

$$p = P\{TS > Obs\} = P\left\{\mathbf{e}^T\left(\mathbf{I}_\mathrm{n} - (1 + Obs)\mathbf{M}_1\right)\mathbf{e} > 0\right\}.$$

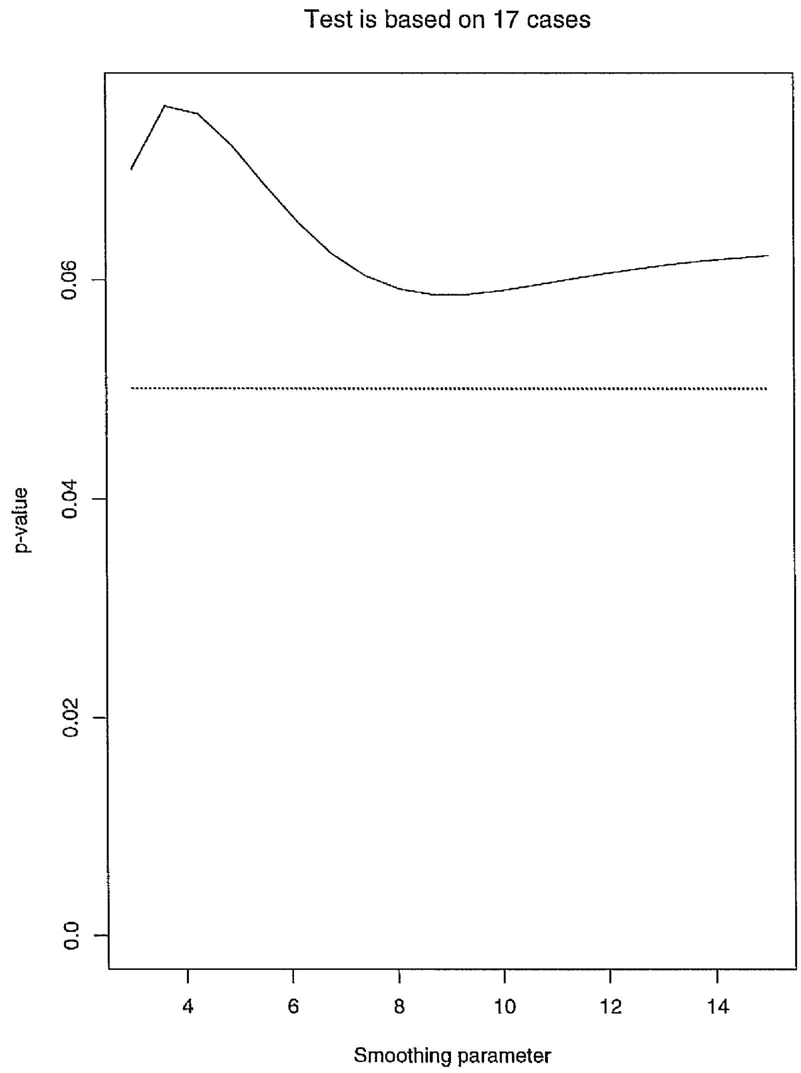
This is a quadratic form in Normal variates, of the same form as those met in Chapter 4. Azzalini & Bowman fitted the Johnson family of curves to this quadratic form, matching the first four moments, to give a null distribution. Here, the familiar three-moment chi-squared approximation is again used.

### 6.2.1 Illustrating the method of residual smoothing by example

To illustrate the technique of residual smoothing, the Stack Loss data are again considered. In Chapter 2, the three-dimensional Figure 2.3.6 suggested that a linear term in the covariate air flow was inadequate, and this agreed with comments made by Daniel & Wood (1980). Section 3.4.2 of Chapter 3 investigated this further by means of a reference band superimposed on a plot of the smoothed residuals. This too suggested the linear model was inappropriate. In this section, the formal test of Section 6.2 is applied to these data.

As with the other nonparametric tests already encountered, the result is reported in the form of a significance trace. Figure 6.2.1 displays the significance trace for the test based on a linear term in air flow. The figure shows that, for a wide range of smoothing parameters, the test returns a $p$-value close to 0.06. While not significant at the 5% level, this is clearly small enough to cause concern, and confirms the impressions gained earlier, that a linear term is inappropriate. The next step would be to either adopt a smooth function in air flow, or determine an appropriate higher-order polynomial. Often, the flexibility of a smooth function, leading to a generalised additive model, would suffice, as a suitable parametric formulation may not be obvious. In this case however, Daniel & Wood (1980) suggested that a quadratic term might

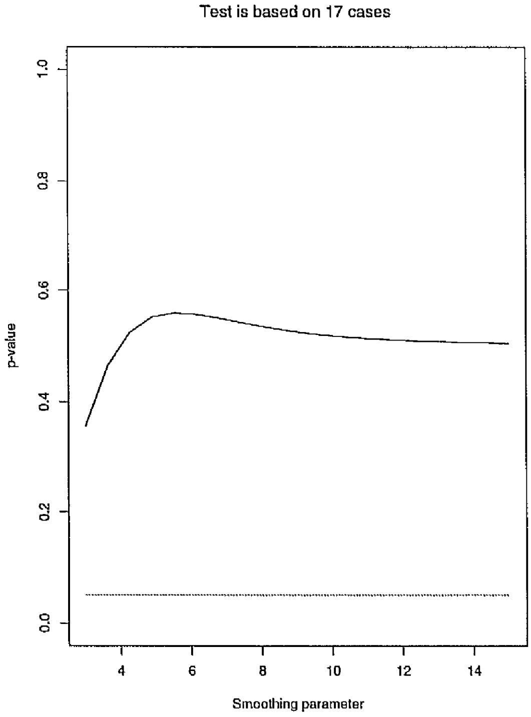
Figure 6.2.1: Significance trace for test of a linear term in air flow



suffice, so that is tested here also.

Figure 6.2.2 shows the Significance Trace obtained when the test is applied to the residuals from a regression where a quadratic term in air flow is fitted. The second covariate, water temperature, remains as a linear term in the model. Here the $p$-values are much higher, around the 0.5 mark, indicating that the quadratic term is much more acceptable. For comparison, an $F$-test of the linear model within the quadratic one returns a $p$-value of 0.051, which is broadly in agreement with Figure 6.2.1.

Figure 6.2.2: Significance trace for test of a quadratic term in air flow



This data set has been used to illustrate several techniques, because it is a familiar one which has been analysed many times before. However, the nature of the data is such that, as Daniel & Wood (1980) comment, even the quadratic model is "not a very nice fit". Nevertheless, a number of features are apparent in the data, and they have been brought out in the methods of Chapters 2 and 3, and again here.

## 6.3 Multivariate smoothing

This section considers the topic of multivariate smoothing, and as with previous sections, addresses the important area of inference. A particular application in image analysis is used in Section 6.3.1 to motivate a test which is an extension of the work in Section 6.2, but involves bivariate smoothing. In Section 6.4, a test is proposed for a more general analysis of covariance, with two continuous explanatories, which extends the work of Chapter 4.

Section 6.2 looked at the case where there are $v$ explanatory variables, and it is required to test the suitability of a parametric model against a semiparametric model with a smooth term in a single covariate. This section considers another aspect of the multivariate case, where it is required to carry out a multiple regression, and produce a smooth, regression *surface*.

The method of local line fitting extends naturally to the multivariate setting. When there are $v$ explanatory variables (let $\mathbf{X} = [x_1, \ldots, x_v]^T$), the method involves fitting a locally-weighted hyperplane with $v + 1$ parameters. In the simplest case, with two explanatories, this is *local plane* fitting.

Ruppert & Wand (1994) show that, if $g(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, the multivariate regression analogue of the local line fitting estimator is

$$\hat{g}(\mathbf{x};\mathbf{H}) = \mathbf{e}_1^T(\mathbf{X_x}^T\mathbf{W_x}\mathbf{X_x})^{-1}\mathbf{X_x}^T\mathbf{W_x}\mathbf{Y}$$

where

$$\begin{aligned}
\mathbf{X_x} &= \begin{bmatrix} 1 & (\mathbf{X_1}-\mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X_n}-\mathbf{x})^T \end{bmatrix}, \\
\mathbf{W_x} &= diag\{k_{\mathbf{H}}(\mathbf{X_1}-\mathbf{x}),\ldots,k_{\mathbf{H}}(\mathbf{X_n}-\mathbf{x})\}, \\
\mathbf{Y} &= [Y_1,\ldots,Y_n]^T,
\end{aligned}$$

and $\mathbf{e}_1$ is the $(v+1)\times 1$ vector with 1 in the first entry and 0 elsewhere. In addition, $k$ is a $v$-variate kernel function (in the two-dimensional case, this is taken to be a bivariate Normal), and $\mathbf{H}$ is a bandwidth matrix, which for simplicity is taken to be $diag(h_1^2,\ldots,h_p^2)$. The question of bias in the multivariate case is addressed later, in Section 6.4.1.

Another estimator for $g(\mathbf{x})$ is the multivariate version of the Nadaraya-Watson (1964) estimator. The basic estimator extends quite intuitively to more than one explanatory, and, as in the method above, a multivariate Normal kernel is used, with a diagonal bandwidth matrix. Although the "local hyperplane" method shares with its univariate counterpart several advantages over the Nadaraya-Watson (1964) estimator, the latter approach is quicker to calculate, and it is for this reason that it is mentioned here, and indeed adopted for the test in Section 6.3.1.

### 6.3.1 An application in image analysis

Figure 6.3.3 shows two sample images which have been generated. Both images are of the same object, but with different random error, or ”white noise”, added. It is of interest to be able to test whether, allowing for the noise, the underlying images are the same. This section proposes a technique, based on the same principle as that of Section 6.2, to test for equality. The data for the images are in the form of two covariates, specifying a point on the image, and a response, which indicates the ”brightness”, or intensity, at that point. The first step is to calculate the vector of differences, **d**, between the two images as observed. If the underlying images are the same, then this vector **d** would consist of random observations with mean zero, displaying no discernible pattern. Any indication of a trend in the differences would suggest that the underlying images differ. Figure 6.3.4 shows perspective plots of the observed differences between the two images, and of the smoothed differences, both on the same scale. For the smooth plot, the bandwidth matrix is $\mathbf{H} = diag(1.5, 1.5)$. It

Figure 6.3.3: Plots of the two sample images

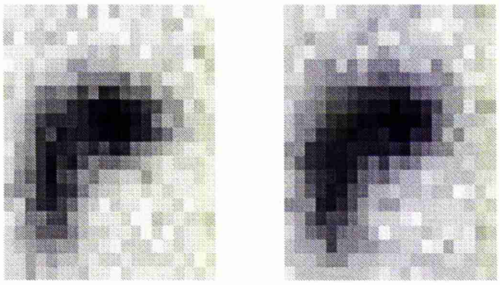
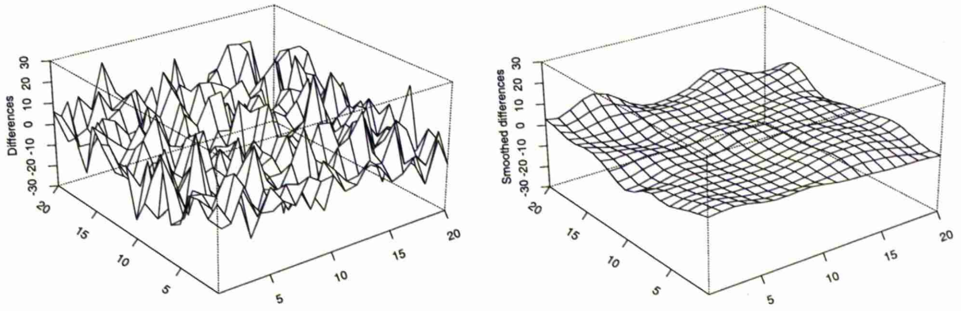


Figure 6.3.4: Plots of the observed and smoothed differences



appears from these plots, that there is no obvious pattern in the differences, indicating that the underlying images appear to be the same.

Note that this approach assumes that the differences, and therefore any errors in the model, are uncorrelated and Normally distributed, under the null hypothesis of equality. Also, the zero expectation of the differences eliminates any bias, even if the Nadaraya-Watson (1964) estimator is used.

Thus, in a manner similar to Section 6.2, the following test statistic is proposed:

$$TS = \frac{R_0 - R_1}{R_1}$$

where $R_0 = \mathbf{d}^T\mathbf{d}$, $R_1 = \mathbf{d}^T(\mathbf{I}_n - \mathbf{S_H})^T(\mathbf{I}_n - \mathbf{S_H})\mathbf{d}$, and $\mathbf{S_H}$ is an $n \times n$ bivariate smoother matrix, for bandwidth matrix $\mathbf{H} = diag(h_1, h_2)$, and $n$ equal to the total number of observed data points.

Now, if $\mathbf{M} = (\mathbf{I}_n - \mathbf{S_H})^T(\mathbf{I}_n - \mathbf{S_H})$, then

$$TS = \frac{\mathbf{d}^T(\mathbf{I}_n - \mathbf{M})\mathbf{d}}{\mathbf{d}^T\mathbf{M}\mathbf{d}}. \tag{6.3.1}$$

This gives the following expression for a $p$-value

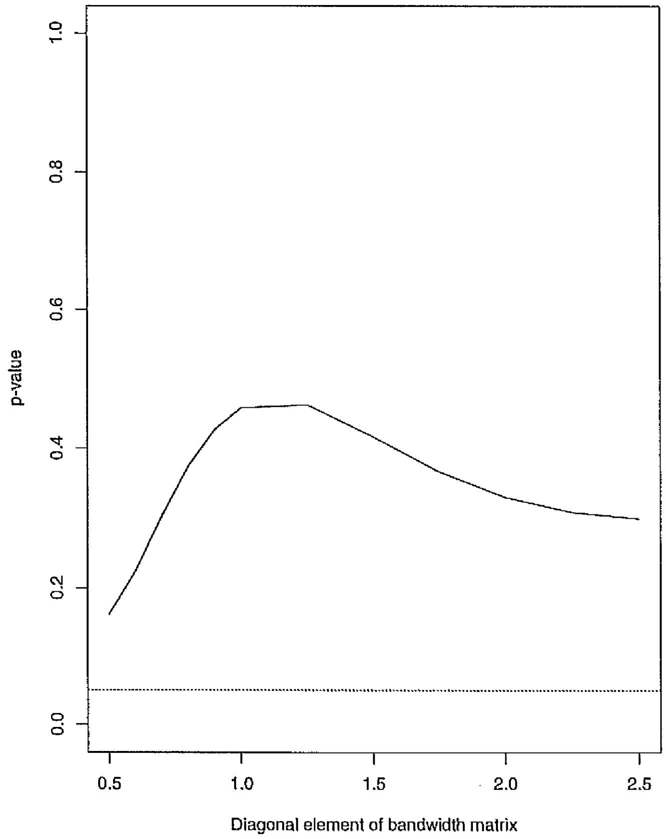
$$\begin{aligned} P(TS > Obs) &= P\left\{\mathbf{d}^T(\mathbf{I}_n - \mathbf{M})\mathbf{d} > \mathbf{d}^T\mathbf{M}\mathbf{d} \times Obs\right\} \\ &= P\left\{\mathbf{d}^T(\mathbf{I}_n - (1 + Obs)\mathbf{M})\mathbf{d} > 0\right\}. \end{aligned}$$

As before, this is a quadratic form in Normal variates, of the type $\mathbf{z}^T\mathbf{A}\mathbf{z}$ where $E(\mathbf{z}) = \mathbf{0}$ and $\mathbf{A}$ is an $n \times n$ symmetric matrix, and so the three-moment approximation to a chi-squared distribution can again be used. Note also that the expression in Formula 6.3.1 is scale invariant, and so $Cov(\mathbf{d})$ can be set equal to the identity matrix, without loss of generality.

### 6.3.2 Applying the test to the example data

In the same manner as in previous chapters, the results can be displayed via a significance trace. This example differs slightly from the previous ones, in that there are now two smoothing parameters to choose (when the bandwidth matrix is diagonal). However, the symmetry of this application means that it is reasonable to assume the two parameters to be equal, thus allowing the familiar signifance trace to be plotted. The results for the image data in Section 6.3.1 are shown in Figure 6.3.5. It is clear from the plot that, over the range of bandwidth matrices chosen, the $p$-values for the

Figure 6.3.5: Significance trace for the test of equality of images



test are all much larger that 0.05, indicating that there is no evidence to reject the null hypothesis that the underlying images are the same.

## 6.4 Multivariate analysis of covariance

The previous section dealt with a specific situation where two surfaces were being compared, and where the design spaces were identical. This section extends that to the general case where there are several surfaces to be compared, and where the design densities are not constrained to be the same.

The method is similar to that employed in Chapter 4, and uses the same quadratic form theory, and an extension of the univariate bias results.

### 6.4.1 Bias in the multivariate case

In this section, the method of local plane fitting is adopted. Predictably, this is because of its superior bias properties, which can be utilised here in much the same way as in the univariate case.

Ruppert & Wand (1994) show that in the multivariate case, the asymptotic bias of the local plane smoother, with standard bivariate Normal kernel and bandwidth matrix $\mathbf{H}$, is given by

$$\frac{1}{2}tr\{\mathbf{H}\mathcal{H}(x)\}, \qquad (6.4.2)$$

where $\mathcal{H}$ is the Hessian matrix of the true surface. This is a natural extension of the univariate result given in Chapter 1, and again has the property that the bias is independent of the design density.

Although, for simplicity, only the bivariate case is considered here, the bias result in equation (6.4.2) holds for more than two covariates, and so the test which follows can be extended to compare hyperplanes in more than two dimensions. However, as explained in Section 6.4.3, care should be taken when smoothing in higher dimensions.

### 6.4.2 A test of bivariate equality

In this section, the test of equality from Section 4.2.1 of Chapter 4 is extended to the bivariate case, and instead of comparing two (or more) smooth curves, the problem involves comparing smooth surfaces. In this case the data are $(y_{ij}, x1_{ij}, x2_{ij})$ for $i = 1, \ldots, p; j = 1, \ldots, n_i$, corresponding to observations on a response $y$ and each of two continuous explanatory variables, $x1$ and $x2$, for each subject in $p$ groups. The hypotheses of interest are:

H0: $y_{ij} = g\,(x1_{ij}, x2_{ij}) + \varepsilon_{ij}$

H1: $y_{ij} = g_i\,(x1_{ij}, x2_{ij}) + \varepsilon_{ij}, \quad g_i$'s not all equal

where the $\varepsilon_{ij}$ are independent N(0,$\sigma^2$) random errors.

The test statistic takes the same form as in Section 4.2.1, namely:

$$TS = \frac{\mathbf{y}^T(\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d})^T(\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d})\mathbf{y}}{\hat{\sigma}^2}$$

where $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ are smoother matrices under H0 and H1 respectively. By a similar argument to before, utilising the bias property of the smoother, the numerator can be expressed, at least asymptotically, as $\varepsilon^T\mathbf{Q}\varepsilon$, where $\mathbf{Q} = (\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d})^T(\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d})$.

To complete the test statistic, an estimate of the error variance is required. The method used in the univariate case, that of first-order differencing, does not readily extend to two or more covariates. Hastie & Tibshirani (1990) propose an estimator motivated by a residual-sums-of-squares argument, which in this notation is expressed by

$$\widehat{\sigma^2} = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{S}_\mathrm{d})^T(\mathbf{I}_n - \mathbf{S}_\mathrm{d})\mathbf{y}}{df_{err}}$$

where $n = \sum_{i=1}^{p} n_i$, and

$$df_{err} = n - tr(2\mathbf{S}_\mathrm{d} - \mathbf{S}_\mathrm{d}\mathbf{S}_\mathrm{d}{}^T).$$

Now, let $\mathbf{B} = df_{err}^{-1} \times (\mathbf{I}_\mathrm{n} - \mathbf{S}_\mathrm{d})^T(\mathbf{I}_\mathrm{n} - \mathbf{S}_\mathrm{d})$, an $n \times n$ symmetric matrix. Thus, the expectation of $\widehat{\sigma^2}$ is $E(\boldsymbol{\varepsilon}^T\mathbf{B}\boldsymbol{\varepsilon}) + \mathbf{g}^T\mathbf{B}\mathbf{g}$. The term $\mathbf{g}^T\mathbf{B}\mathbf{g}$ is the sum of squared differences between the true surface and the smoothed true surface, which is small relative to the first term in $\boldsymbol{\varepsilon}$. As in Chapter 4, this term is ignored, making the test conservative, and the following expression for a $p$-value emerges:

$$p = P\left\{\frac{\boldsymbol{\varepsilon}^T\mathbf{Q}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T\mathbf{B}\boldsymbol{\varepsilon}} > Obs\right\} = P\left\{\boldsymbol{\varepsilon}^T(\mathbf{Q} - \mathbf{B} \times Obs)\boldsymbol{\varepsilon} > 0\right\}.$$

Once again, this is a quadratic form in Normal variates, with all the required properties. The same theory as before can be applied to fit an approximate chi-squared null distribution, and from there it is a simple matter to obtain the $p$-value for the test.

In the bivariate case, the question of how to display the test results is not as clear as in Chapter 4. Here there are two smoothing parameters to consider, one for each covariate. Let these be $h_1$ and $h_2$. Perhaps the simplest solution would be to produce a table of $p$-values corresponding to various combinations of $h_1$ and $h_2$, or alternatively, to plot a three-dimensional surface representing the $p$-values.

The requirement to calculate, and operate on, the smoother matrices $\mathbf{S}_\mathrm{s}$ and $\mathbf{S}_\mathrm{d}$ mean that this procedure is extremely computationally intensive. With current computer processing capabilities the process is slow, even for small datasets. For that reason, merely the theory is presented here, without example.

### 6.4.3 The curse of dimensionality

It should be noted that smoothing in higher dimensions can be problematic. Unless there are a large number of observations, sparsity of data can make higher-dimensional smoothing difficult. In addition, there is an increase in the number of bandwidth parameters, which often leads to simplifying assumptions.

This complication in multidimensional modelling is often called the *curse of dimensionality.* While bivariate smoothing with practical sample sizes is reasonable, it can be difficult when further covariates are introduced. A number of measures have been proposed to overcome this, and indeed, the Generalized Additive Model is one such measure. By modelling the overall regression as a sum of univariate smooths, one for each covariate, the requirement to smooth multivariately is avoided.

## 6.5 Discussion and final comments

This final chapter has attempted to extend previous results to models with more covariates. When a single smooth term is required, the method of residual smoothing appears to offer an appealing inferential tool. The univariate result of previous chapters extends naturally to multivariate smoothing, leading to a multivariate nonparametric analysis of covariance. The drawback with such an approach is that the computational requirements can be prohibitive, particularly when the "curse of dimensionality" demands large data sets.

It appears that simpler models may be more appealing, one class being Generalised Additive Models, where the overall regression is defined as a sum of univariate smooths (or a combination of smooth and parametric terms). Such models are an intuitive progression from the familiar parametric approach. It is natural therefore, to mimic the parametric methods of inference, and apply an $F$-test to this new nonparametric (or *semiparametric*) class of model, as exemplified by Hastie & Tibshirani (1990). Formal distributional results, however, would be welcome, in order to establish methods of inference with a strong theoretical background.

This thesis has attempted to address the topic of inference in nonparametric regression. There can be no question that nonparametric smoothing offers a simple and flexible graphical tool for examining relationships in data. What is often lacking is a set of inferential tools to accompany the models, and this thesis has attempted to introduce techniques applicable in many common situations.

The tests proposed have focussed on comparing models, and have used this to eliminate bias, so often a problem in the nonparametric field. The appealing properties of local line fitting and the Gasser-Müller (1979) estimator are also significant in the results presented here. In most cases these bias results enabled intuitive test statistics to be expressed as quadratic forms, whose distributional properties are well-known.

An area in nonparametric modelling which attracts a lot of attention is choice of smoothing parameter. This thesis has attempted to demonstrate that this topic can assume less importance where inference is concerned. It is often the case that the result of a test (or even the impression gained from a graphical comparison) is unaffected by choice of smoothing parameter.

The growth of nonparametric modelling can perhaps be attributed to its graphical appeal, and this aspect has certainly not been ignored in this thesis. Many graphical methods have been introduced, again focussing on the comparison of models, and often to be used to complement the corresponding nonparametric tests.

The ongoing developments in computing capabilities would suggest that nonparametric modelling will continue to be a popular topic for research. It offers many opportunities for flexible modelling, and the tools, and software packages, are now readily available to make this attractive technique a standard addition to the statistician's inventory. The obvious graphical appeal of nonparametric modelling, and lack of straightforward distributional properties, have meant that the topic of inference was often neglected. This situation seems to be changing, as the importance of inference in nonparametrics is recognised, and this thesis has attempted to contribute to the body of work now emerging.

# Appendix

This Appendix explains the derivation of the matrix form of the numerator of (4.2.6) in the test of parallelism in Chapter 4. The numerator is defined as

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}\{\hat{\alpha}_i + \hat{g}(x_{ij}) - \hat{g}_i(x_{ij})\}^2$$

In the vector-matrix notation of Chapter 4, this becomes

$$[\{\mathbf{D}\hat{\boldsymbol{\alpha}} + \mathbf{S_S}(\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}})\} - \mathbf{S_d}\mathbf{y}]^T\,[\{\mathbf{D}\hat{\boldsymbol{\alpha}} + \mathbf{S_S}(\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}})\} - \mathbf{S_d}\mathbf{y}] \qquad \text{(A1)}$$

Considering the left-hand side of (A1) only for now, this can be expressed as

$$\begin{aligned}
& [(\mathbf{I}_n - \mathbf{S_s})\mathbf{D}\hat{\boldsymbol{\alpha}} + (\mathbf{S_s} - \mathbf{S_d})\mathbf{y}]^T \\
= \; & [(\mathbf{I}_n - \mathbf{S_s})\mathbf{D}\hat{\boldsymbol{\alpha}} + (\mathbf{S_s} - \mathbf{S_d})\,(\mathbf{D}\boldsymbol{\alpha} + \mathbf{g} + \boldsymbol{\varepsilon})]^T \quad \text{under H0} \\
= \; & [(\mathbf{I}_n - \mathbf{S_s})\mathbf{D}\hat{\boldsymbol{\alpha}} + (\mathbf{S_s} - \mathbf{S_d})\,(\mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon})]^T \quad \text{using the bias property} \\
= \; & [\mathbf{D}\hat{\boldsymbol{\alpha}} - \mathbf{S_d}\mathbf{D}\boldsymbol{\alpha} - \mathbf{S_s}\mathbf{D}\hat{\boldsymbol{\alpha}} + \mathbf{S_s}\mathbf{D}\boldsymbol{\alpha} + (\mathbf{S_s} - \mathbf{S_d})\boldsymbol{\varepsilon}]^T \\
= \; & [\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - \mathbf{S_s}\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + (\mathbf{S_s} - \mathbf{S_d})\boldsymbol{\varepsilon}]^T \quad \text{since } \mathbf{S_d}\mathbf{D}\boldsymbol{\alpha} = \mathbf{D}\boldsymbol{\alpha} \\
= \; & [(\mathbf{I}_n - \mathbf{S_s})\mathbf{D}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + (\mathbf{S_s} - \mathbf{S_d})\boldsymbol{\varepsilon}]^T \qquad \text{(A2)}
\end{aligned}$$

Using dotted lines to denote partitioned matrices, (A2) becomes the transpose of

$$\left[\begin{array}{c:c} (\mathbf{I}_n - \mathbf{S}_\mathrm{s})\mathbf{D} & (\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d}) \end{array}\right] \left[\begin{array}{c} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \cdots\cdots \\ \boldsymbol{\varepsilon} \end{array}\right]$$

Employing a similar argument for the right-hand side of (A1) gives the required expression for the entire numerator of the test statistic:

$$\left[\begin{array}{c} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \cdots\cdots \\ \boldsymbol{\varepsilon} \end{array}\right]^T \left[\begin{array}{c:c} (\mathbf{I}_n - \mathbf{S}_\mathrm{s})\mathbf{D} & (\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d}) \end{array}\right]^T$$

$$\left[\begin{array}{c:c} (\mathbf{I}_n - \mathbf{S}_\mathrm{s})\mathbf{D} & (\mathbf{S}_\mathrm{s} - \mathbf{S}_\mathrm{d}) \end{array}\right] \left[\begin{array}{c} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \cdots\cdots \\ \boldsymbol{\varepsilon} \end{array}\right]$$

# References

Aitchison, J. & Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press: Cambridge.

Angell, I.O. (1981). A practical introduction to computer graphics. London: Macmillan.

Azzalini, A., Bowman, A.W. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Azzalini, A. & Bowman, A.W. (1990). A look at some data on the Old Faithful Geyser. *Applied Statistics* **39**, 357–365.

Azzalini, A. & Bowman, A.W. (1991). Nonparametric methods for repeated measurements. In "Nonparametric functional estimation and related topics", Roussas G. ed., Kluwer Academic Publishers (Dordrecht, Boston, London), *NATO ASI Series C, Mathematical and Physical Sciences*, **335**, 377–87.

Azzalini, A. & Bowman, A.W. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society*, Series B,

**55**, 549–557.

Bowman, A.W. (1992a). Density based tests for goodness-of-fit. *J.Stat.Comp.Sim.* **40**, 1–13.

Bowman, A.W. (1992b). Contribution to discussion of Hall, P. & Johnstone, I. on "Empirical functionals and efficient smoothing parameter selection." *Journal of the Royal Statistical Society*, Series B, **54**, 511–512.

Bowman, A.W. & Foster, P.J. (1993). Density-based exploration of two-dimensional data. *Statistics & Computing* **3**, 171–177.

Bowman, A.W. & Robinson, D.R. (1990). Introduction to Regression and Analysis of Variance: a computer illustrated text. IOP Publishing Ltd.: Bristol.

Bowman, A.W. & Young, S.G. (1996). Graphical Comparison of Nonparametric Curves. *Applied Statistics* **45**, 83–98.

Brownlee, K.A. (1965). Statistical Theory and Methodology in Science and Engineering. Wiley; New York.

le Cessie, S. & van Houwelingen, J.C. (1991). A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods. *Biometrics* **47**, 1267–1282.

Chu, C.K. & Marron, J.S. (1991). Choosing a kernel regression estimator (with discussion). *Statistical Science* **6**, 404–436.

Clark, R.M. (1975). A calibration curve for radiocarbon dates. *Antiquity* **49**, 251–266.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W.S. & 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. (1993). Visualising data. AT&T Bell Laboratories: New Jersey.

Cook, R.D. & Weisberg, S. (1982). Residuals and influence in regression. London: Chapman & Hall.

Copas, J.B. (1983). Plotting p against x. *Applied Statistics* **32**, 25–31.

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, Series B, **34**, 187–220.

Cox, D.R. & Oakes, D. (1984). Analysis of survival data. London: Chapman & Hall.

Daniel, C. & Wood, F.S. (1980). Fitting equations to data: computer analysis of multifactor data. Wiley; New York.

Diggle, P.J. & Fisher, N.I. (1985). SPHERE: a contouring program for spherical data. *Computers and Geosciences* **11**, 725–766.

Eubank, R.L. & Hart, J.D. (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* **80**, 89–98.

Fan, J.Q. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J.Q. (1993). Local linear-regression smoothers and their minimax efficiencies. *Annals of Statistics* **21, No.1**, 196–216.

Fan, J.Q., Gijbels, I. (1995). Data-driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society*, Series B, **57**, 371–394.

Fan, J., Heckman, N.E. & Wand, M.P. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions, *Journal of the American Statistical Association*, **90**, 141–150.

Fisher, N.I., Lewis, T. & Embleton, B.J.J. (1987). Statistical analysis of spherical data. Cambridge University press: Cambridge.

Gasser, T. & Müller, H-G. (1979). Kernel estimation of regression functions. In Lecture Notes in Mathematics **757**, 23–68; eds. T. Gasser & M. Rosenblatt. Springer-Verlag: New York.

Gasser, T., Kneip, A. and Kohler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* **86**, 643–652.

Gower, J. (1990). Three-dimensional biplots. *Biometrika* **77**, 773–785.

Green, P.J. & Silverman, B.W. (1994). Nonparametric regression and generalised linear models. London: Chapman & Hall.

Hall, P. & Hart, J. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* **85**, 1039–1049.

Hall, P., Kay, J.W. & Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

Hall, P. & Titterington, D.M. (1988). On confidence bands in nonparametric density estimation and regression *Journal of Multivariate Analysis* **27**, 228–254.

Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press: London.

Härdle, W. & Bowman, A.W.(1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* **83**, 102–110.

Härdle, W., Hall, P. & Marron, J.S. (1988). How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum? (with discussion). *Journal of the American Statistical Association* **83**, 86–99.

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21**, 1926–1947.

Härdle, W. & Marron, J.S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics* **19**, 778–796.

Härdle, W. & Marron, J.S. (1995). Fast and Simple Scatterplot Smoothing. *Computational Statistics & Data Analysis* **20**, 1–17.

Hastie, T.J. & Tibshirani, R.J. (1990). Generalized Additive Models. Chapman &

Hall: London.

Hinde, J. (1988). A review of "MacSpin". *Applied Statistics* **37**, 124–126.

Holliday, R. (1960). Plant Population and Crop Yield. *Field Crop Abstr.* **13**, 159–167 & 247–254.

Hosmer, D.W. & Lemeshow, S. (1989). Applied logistic regression. Wiley: New York.

Huber, P.J. (1987). Experiences with three-dimensional scatterplots. *Journal of the American Statistical Association* **82**, 448–453.

Johnson, N.L. & Kotz, S. (1970a). Distributions in Statistics: Continuous Univariate Distributions 1. Wiley: New York.

Johnson, N.L. & Kotz, S. (1970b). Distributions in Statistics: Continuous Univariate Distributions 2. Wiley: New York.

Kalbfleisch, J.D. & Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. Wiley: New York.

King, E., Hart, J.D. & Wehrly, T.E. (1991). Testing the equality of two regression curves using linear smoothers. *Statist.Prob.Letters* **12**, 239–247.

McCullagh, P. & Nelder, J.A. (1989). Generalized Linear Models (Second Edition). Chapman and Hall: London.

MacSpin (1986). Graphical data analysis software: version 1.1. $D^2$ Software, Inc.: Austin.

Macauley, F.R. (1931). The Smoothing of Time Series. National Bureau of Economic Research, New York.

Mahalanobis, P.C., Majumdar, D.N. &Rao, C.R. (1949). Anthropometric survey of the United Provinces, 1941: a statistical study. *Sankhya* **9**, 89–324.

Mallows, C.L. (1973). Some comments on $C_p$. *Technometrics* **15**, 661–675.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

Parzen, E. (1979). Nonparametric statistical data modelling. *Journal of the American Statistical Association* **74**, 105–131.

Pearson, E.S. (1963). Some problems arising in approximating to probability distributions using moments. *Biometrika* **50**, 94–111.

Pickering, J.F. (1985). Giving in the Church of England: an econometric analysis. *Applied Economics* **17**, 619–632.

Prentice, R.L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279–288.

Ratkowsky, D.A. (1983). Nonlinear regression modelling (Statistics: Textbooks and Monographs, Volume 48). Marcel Dekker Inc.: New York.

Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Annals of Statistics* **12**, 1215–30.

Ruppert, D., Sheather, S.J. & Wand, M.P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Scott, D.W. (1992). Multivariate density estimation: theory, practice and visualisation. Wiley: New York.

Sheather, S.J. & Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B, **53**, 683–690.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society*, Series B, **47**, 1–52.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman & Hall: London.

Solomon, H. & Stephens, M.A. (1978). Approximation to density functions using Pearson curves. *Technometrics* **73**, 153–160.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society*, Series B, **50**, 413–36.

Stablein, D.M. & Koutrouvelis, I.A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics* **41**, 643–652.

Stone, C.J. (1977). Consistent Nonparametric Regression. *Annals of Statistics* **5**, 595–620.

Sun, J & Loader, C.R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics* **22**, 1328–1345.

Tierney, L. (1990). Lisp-Stat. Wiley: New York.

Upson, Craig, Faulhaber, T., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz, R., van Dam, A. (1989). The Application Visualisation System: a computational environment for scientific visualisation. *IEEE Computer Graphics & Applications* **9**, 30–42.

Wand, M. & Jones, M.C. (1995). Kernel Smoothing. London: Chapman & Hall.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā*, Series A, **26**, 359–372.

Yanagimoto, T. & Yanagimoto, M. (1987). The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model. *Technometrics* **29**, 95–101.

Young, S.G. & Bowman, A.W. (1995). Nonparametric analysis of covariance. *Biometrics* **51**, 920–931.