Limburgs Universitair Centrum
Faculteit Wetenschappen

Smoothing Techniques and Bootstrap Methods for Multiparameter Likelihood Models

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen: Wiskunde
aan het Limburgs Universitair Centrum te verdedigen door

Gerda CLAESKENS

Promotor: Prof. dr. M. Aerts

1999

Smoothing Techniques and Bootstrap Methods for Multiparameter Likelihood Models

Gerda Claeskens

Research Assistant of the Fund for
Scientific Research Flanders – Belgium (F.W.O.)

Center for Statistics
Limburgs Universitair Centrum
B-3590 Diepenbeek, Belgium
Gerda.Claeskens@luc.ac.be

May 28, 1999

Acknowledgements

It is my pleasure to thank everyone who helped me with the realization of this Ph.D. dissertation.

In the autumn of 1995, when I for the first time entered Marc Aerts' office, I was just hoping that I made the right decision. Now, I am sure I did! To Marc, my advisor, many thanks: for guiding me through the entrance to the fascinating world of statistical research, for his enormous patience and understanding and for all the fun we had and still have.

I am also very grateful to Jeff Hart for the many things I learned from him and for numerous good advice.

Certainly I appreciate the stimulating conversations with Geert Molenberghs and enjoyed the cooperation with Matt Wand.

Another important and unique source of support is provided by my mother and father and by An & David. Them I owe many thanks for always being there and for trying to understand my enthusiasm. And, for everything else...

At several occasions and in several ways, I learned from the following persons at the Center for Statistics: Paul Janssen, Noël Veraverbeke and Herman Callaert. I really enjoyed the time spent with (in alphabetical order): Bart, Didier, Heidi, Helena, Herbert, Ingrid, Liesbeth, Lieven, Roel, Thomasz, Ziv.

To all mentioned above, and all I accidently forgot to mention: Thank you so much!!

Gerda

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In statistical literature there has been (and still is) a lot of interest in the "classical" regression model

$$Y_i = \theta(x_i) + \varepsilon_i$$

where the errors ε_i are zero mean (usually continuous and independent) random variables. Mainly because its ease of interpretation and its mathematical attractiveness, statistical aspects of this model are nowadays well understood. In practice however, lots of data sets do not show this simple additive structure: response = mean + error. A first extension leads to the class of generalized linear models (McCullagh and Nelder, 1989). A setting which requires an even more complex approach occurs when the study subjects belong to some "natural groups". For example, in ophthalmology, both eyes of a person form a natural cluster, or in teratology, litters contain several foetuses. Also members of the same family or inhabitants of the same geographical area have to be considered as clusters. Subjects of the same group are likely to show a more similar behavior than subjects belonging to different groups. We might assume independence between clusters, but within clusters, the data are often correlated. For this kind of data we have to use a model with at least *two* "parameters" where one parameter may be the mean response and the other describes the association of subjects within a cluster. We are particularly interested in clustered data with binary response. Several examples of such data sets are given in the first section.

The first challenge we take is to extend existing *smoothing* techniques for the classical model to *multiparameter likelihood* models. We propose local polynomial estimators in a likelihood setting with $\kappa \geq 1$ parameters $(\theta_1(x), \ldots, \theta_\kappa(x))$. Our approach allows for simultaneous estimation of all κ curves $(\theta_1(\cdot), \ldots, \theta_\kappa(\cdot))$. Also extensions to non-full likelihood models and to additive models are studied in detail. It should be stressed that although we focus attention to *clustered binary response* data, the domain of application of these methods is much broader. Indeed, any one, two, three, ... parameter (likelihood) model fits into this framework.

Our second goal is to address the problem of constructing omnibus *lack of fit tests* based on smoothing ideas. The orthogonal series based tests that we propose are applicable in a wide range of statistical problems and are easy to implement using standard statistical software.

A third objective is to define and to study *bootstrap techniques* for both parametric and nonparametric approaches in multiparameter likelihood based regression models. Next to a study of a parametric bootstrap resampling scheme, we will construct one-step bootstrap estimators, which have the advantage that no additional iterative model fitting is required.

After introducing some examples we will take a closer look at each of the main key-words: smoothing, bootstrap, lack of fit tests, multiparameter likelihood models, clustered binary data.

1.1 Data examples

In each of the data examples below the study subjects are clustered.

Low-iron rat teratology data

The first example uses data from the experimental setup from Shepard, Mackler and Finch (1980). A total number of 58 female rats was given different amounts of iron-supplement (ranging from normal to zero level). The rats were made pregnant and sacrificed 3 weeks later. The total number of foetuses (ranging from 1 to 17), the number of dead foetuses and the hemoglobin levels of the mothers were recorded. The proportion of death foetuses as a function of the mother animal's hemoglobin level is shown in Figure 1.1. This plot does not give any information about the correlation structure within a litter. For this kind of clustered binary data we would

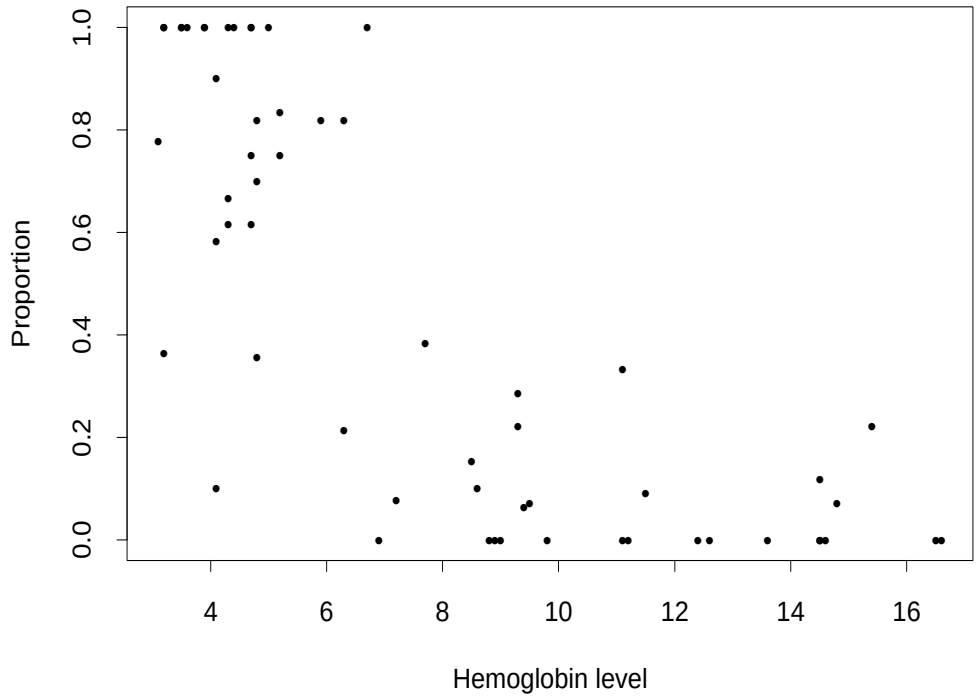


Figure 1.1: Low-iron rat teratology data.

like to construct two curves: the estimated probability of death *and* the estimated correlation within a cluster, both as a function of the covariate (see Section 2.6.2). In Section 3.5.2 we also propose a bootstrap method to construct simultaneous confidence intervals for both curves. An omnibus test for a linear effect of hemoglobin level on the logit of the probability of deaths will be constructed in Section 6.6.1.

POPS data

The Project On Preterm and Small-for-gestational age infants (POPS) collected information on 1338 infants born in the Netherlands in 1983 and having gestational age (x_1) less than 32 weeks and/or birthweight (x_2) less than $1500g$; see Verloove, et al. (1986) for more details. The outcome of interest here concerns the situation after two years. The binary variable Y is 1 if an infant has died within two years after birth or survived with a major handicap, and 0 otherwise. Some of the recorded

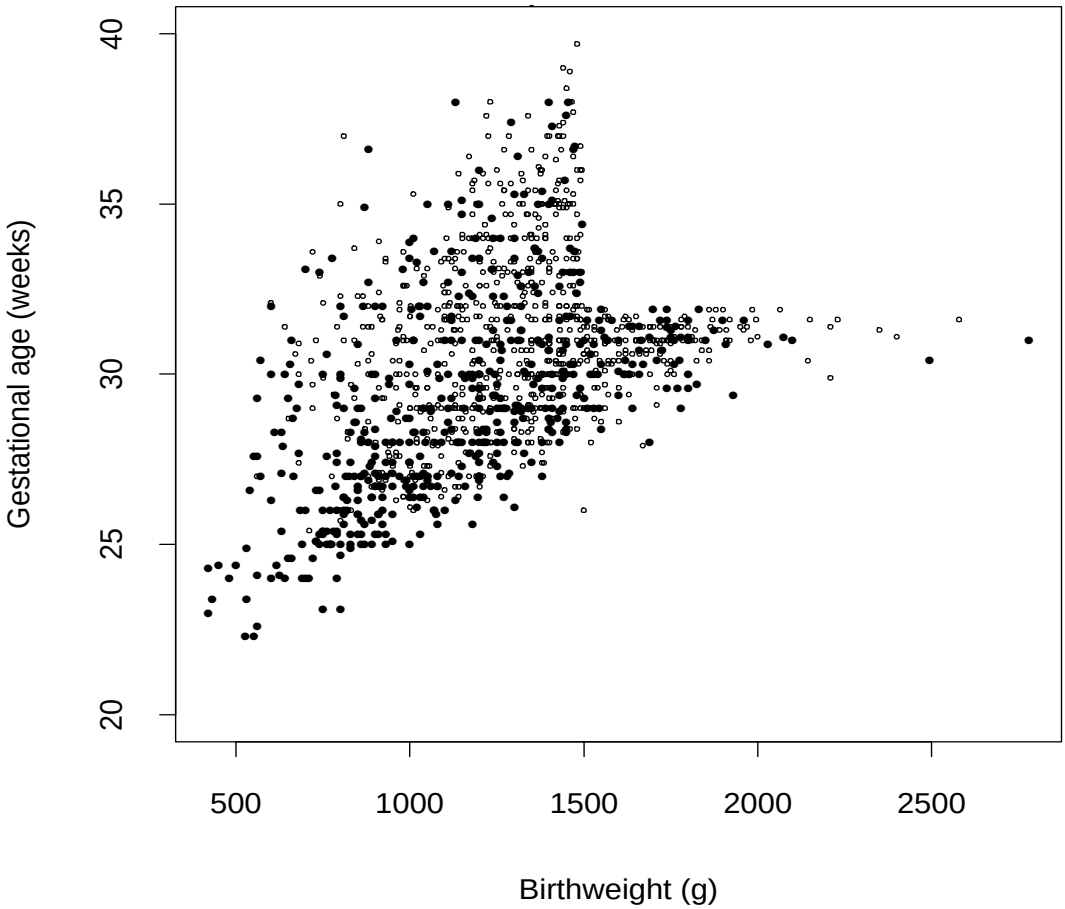


Figure 1.2: POPS data, the open circles correspond to zero outcomes.

observations are from twins or triplets. So, one might have to account for the association between siblings of the same twin (or triplet,...). Another interesting aspect is that there are observations on both cluster and individual level. For example, for a twin, the mother's age and the gestational age is the same for both siblings, while birthweight is subject specific.

An interesting question here is whether local polynomial estimators are still applicable in this multiple regression setting. We will also use the POPS data for illustrating lack of fit tests (see Section 7.6.1).

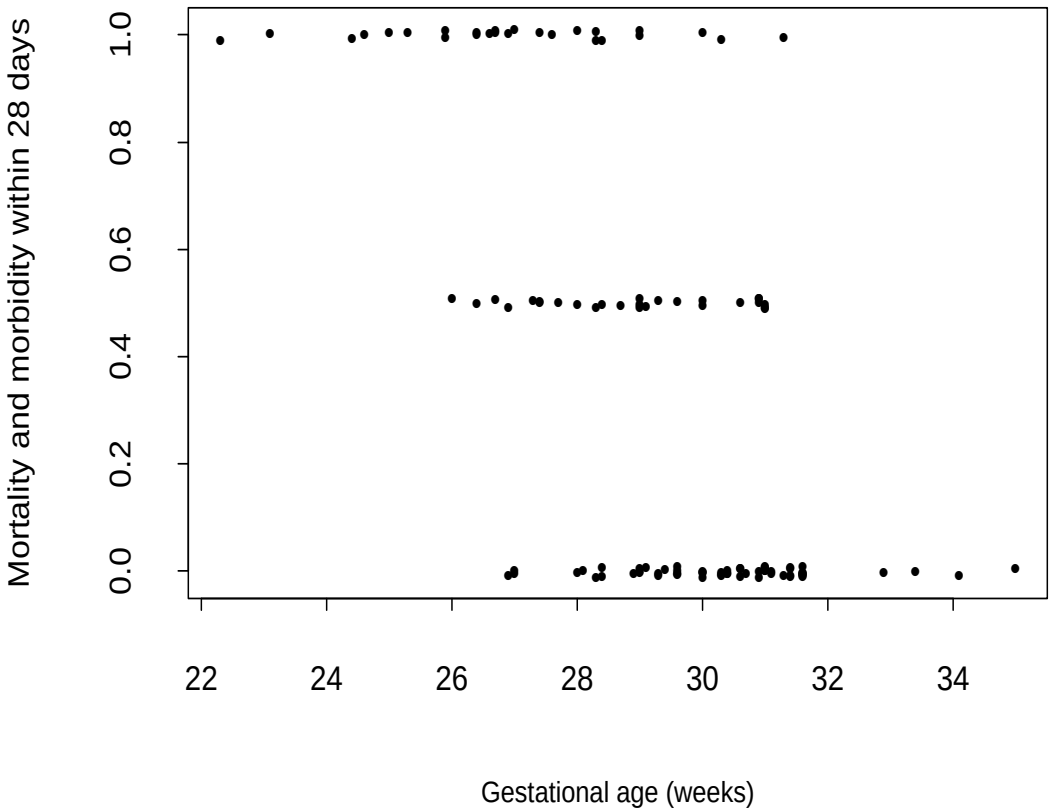


Figure 1.3: Twins data.

The subset of 107 twins is shown graphically in Figure 1.3. In Section 2.6.3 we will estimate the proportion of neonatal mortality and morbidity (i.e. within 28 days after birth) and the correlation between siblings as a function of gestational age. Confidence intervals will be obtained in Section 3.5.3 and an omnibus test for linearity of this proportion (on logit scale) is applied in Section 6.6.2.

The Wisconsin diabetes study [1]

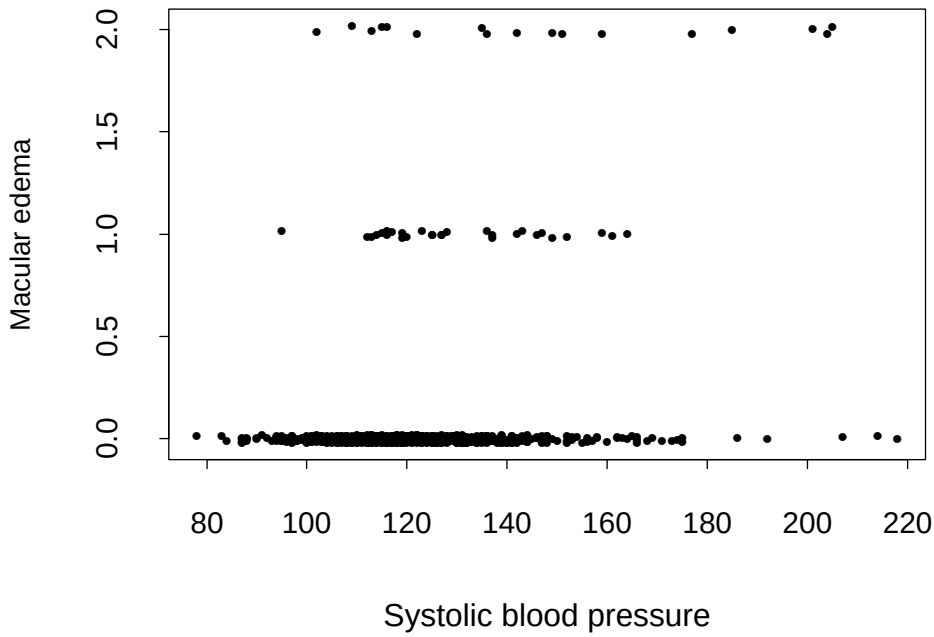


Figure 1.4: Wisconsin diabetes study.

In this data set there are records from 720 younger onset diabetic persons. Both eyes of each person are examined for the presence of macular edema. See Klein, Klein, Moss, Davis, and DeMets (1984) for more details. In the study there were 29 individuals where macular edema is present at only one eye, for 17 of them it is observed at both eyes, and for the remaining 674 persons, it was completely absent. We will study the probability of macular edema as a function of the patient's systolic blood pressure, hereby taking the clustered nature of the data into account, as indeed the response values of both eyes are likely to be correlated. A graphical representation of these data can easily yield some idea of how the proportion of macular edema infected eyes varies with the person's systolic blood pressure.

[1] We thank Professor R. Klein of the University of Wisconsin, Madison, for kindly providing this data set (NIH grant EY 03083, Wisconsin Diabetic Retinopathy Study).

Such a graph however, do not give any information about the correlation between the outcomes of both eyes. Most often, this correlation is just assumed to be some constant, which can be estimated from the data. An omnibus lack-of-fit test would be useful to discuss the validity of this assumption for these data (see Section 6.6). Estimates for the probability of macular edema and for the correlation are obtained in Section 2.6.4 and confidence intervals in Section 3.5.4.

The theophylline data

This data set concerns a study on mice conducted by the Research Triangle Institute under contract to the National Toxicology Program (NTP). It investigates toxic and teratogenic effects of the chemical theophylline (THEO), which is used for example as a drug in the treatment of asthma during pregnancy. It is important to investigate whether therapeutic doses in mothers may be toxic to infants (Lindstrom et.al., 1990). This particular study was designed to determine adverse effects of THEO in pregnant mice. Several dose levels were administered during the period of organogenesis. There was one control group and three treatment groups, who received, respectively, 0, 0.075, 0.15 and 0.20% theophylline in the drinking water. The mice were sacrificed prior to term and the uterine contents were examined for malformations. Since littermates are likely to show the same behavior, this kind of experiment results in cluster correlated binary outcomes (malformation: yes or no).

In Figure 1.5 the proportions of malformed foetuses in a litter are shown for each of the four dose levels. The experimenters have distinguished three types of malformation: external, visceral and skeletal. The occurences of the latter two kinds seems to be rather rare. Observed proportions for a collapsed outcome (any of these malformations) are shown too.

These teratology data are structurally different from e.g. the low-iron rat data. In the THEO example the dose levels (i.e. the values of the covariate) are fixed and restricted to only four different values. No smoothing methods will be applied here. The main question now is whether there is a dose effect. Emphasis can be placed on estimating a dose effect parameter, on testing the null hypothesis of no dose effect, or on determining a benchmark dose. For testing the null hypothesis of no effect, classical tests such as Wald, score or likelihood ratio tests are most often used with P-values obtained from their asymptotic distribution. We want to know whether accuracy can be gained by using bootstrap techniques (see Sections 8.5.2 and 9.4.3).

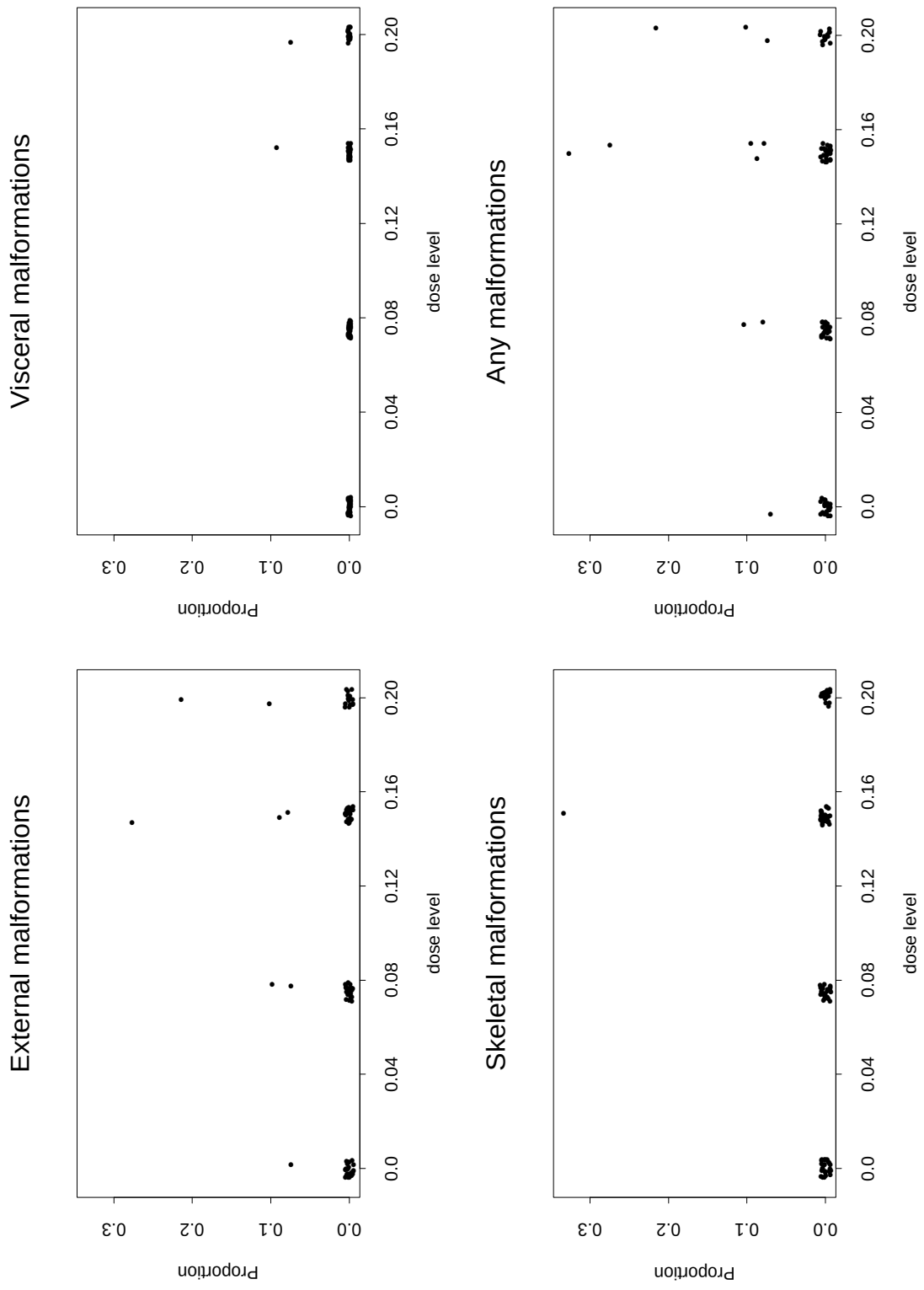


Figure 1.5: Theophylline data (Data values have been jittered to avoid replicate values on the graph).

Study of herbicides on mice

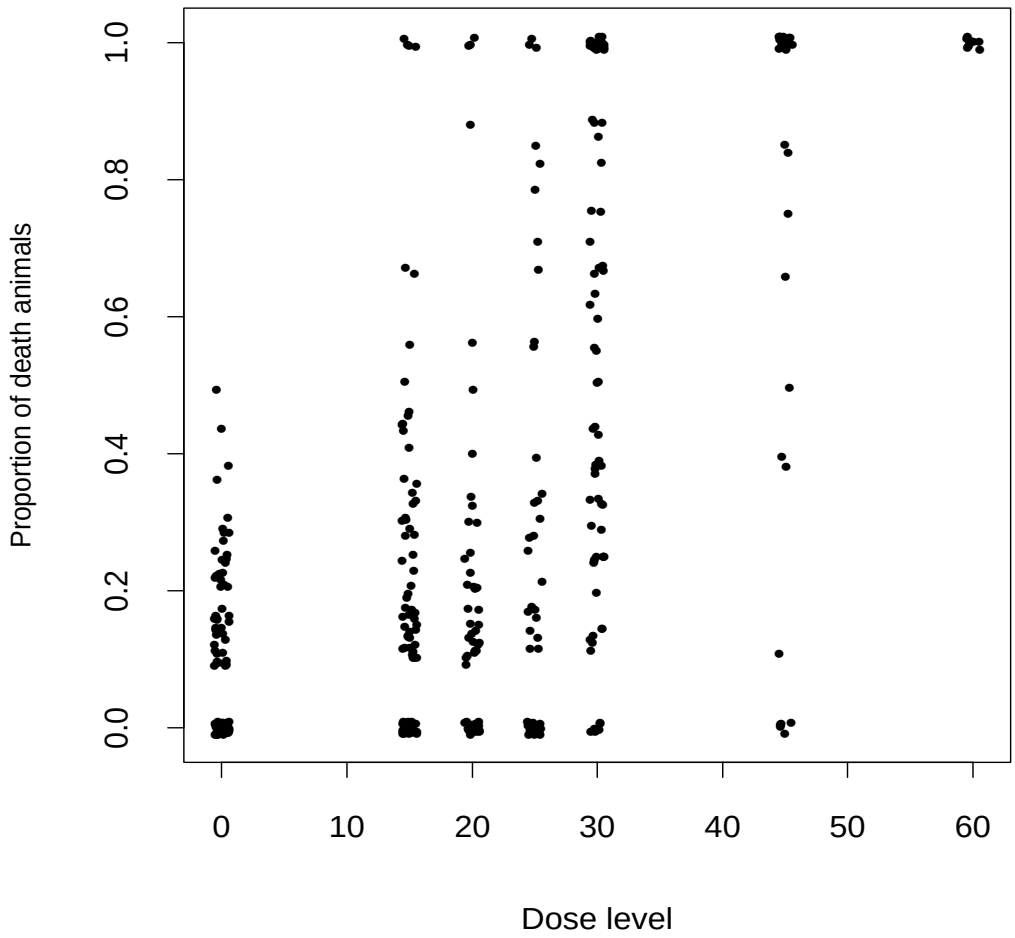


Figure 1.6: Study of herbicides on mice (dose levels have been jittered).

This data set is from a study conducted by the U.S. National Center for Toxicological Research; see Ahn and Chen (1997) for more details. It concerns an experiment on mice, which were exposed to the herbicide 2,4,5-trichlorophenoxyacetic acid. The data set contains seven doses: 0, 15, 20, 25, 30, 45 and 60 mg/kg/day. The number of litters for each dose are 89, 86, 56, 10, 76, 33 and 9 respectively. The observed proportion of death foetuses for each litter is shown in Figure 1.6. Of interest here

is the dose-response relationship (see Section 2.6.5) and the significance of a dose effect (see Section 8.5.3).

1.2 Smoothing techniques

The popularity of *smoothing* methods can partly be explained by their flexibility: data can be evaluated without having to postulate a shape for the curve of interest. Most existing techniques, such as the scatterplot smoothers, focus on the "classical" regression model $Y_i = \theta(x_i) + \varepsilon_i$.

Local polynomial smoothers

In this latter context, *local polynomial* smoothers (of degree p) at a point x are defined as the minimizers with respect to $(\beta_0, \ldots, \beta_p)$ of the following weighted least squares problem:

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (x_i - x)^j \right)^2 K\left(\frac{x_i - x}{h} \right). \tag{1.1}$$

The function K is called the "kernel" and h denotes the "bandwidth". The minimizer β_j is an estimator for the jth derivative of $j!\theta(\cdot)$ at x (e.g., Cleveland, 1979 and Fan, 1992, 1993). Ruppert and Wand (1994) extended this idea to the multiple regression setting.

Although scatterplot smoothers still receive much attention in the statistical literature, their domain of application is restricted. As an example, let us consider the estimation of a "dose-response" curve. Typically, the covariate values $x_1, \ldots, x_n$ denote the dose levels of a chemical which are selected for the experiment, or in the case of the low-iron data, they are the mothers' hemoglobin levels. Staniswalis and Cooper (1988) obtained local constant estimates of the dose- response curve for quantal bioassay studies. Often, in such experiments dose x_i is administered to a number of animals, m_i and, say, $Y_{i\bullet}$ of them show the response of interest (outcome equals 1); the outcome is zero for the remaining $m_i - Y_{i\bullet}$ animals. By calculating the sample proportions $Y_i = Y_{i\bullet}/m_i$ and by assuming relationship $Y_i = \theta(x_i) + \varepsilon_i$, they obtained the estimators via (1.1) with $p = 0$. Other kernel weighted averages of sample proportions have been studied by Copas (1983), Müller and Schmidt (1988), Aragaki and Altman (1997). An important drawback of this method is that one has

to assume that the outcomes are independent, since possible correlations between the binary outcomes are completely ignored by only considering the sample proportions. Another disadvantage of these local averages is that there is no guarantee that the estimators belong to the range of allowable values. It might well happen that an estimated proportion is less than zero or bigger than one. Therefore, a suitable link function should be applied.

Local likelihood estimators

Such a link function can easily be included in the approach of Tibshirani and Hastie (1987) who introduce the *local likelihood estimators*. They explain the concept as follows: "Given a global method for estimating a linear response (e.g., maximum likelihood estimation in the linear logistic model), we apply it locally, estimating a separate line in a window around each x value. The value of the estimated line at x is the estimate of the smooth response function at x ...". Their "running lines" smoother is a local linear estimator in the likelihood model. While Tibshirani and Hastie used symmetric nearest neighborhoods, Staniswalis (1989) defines the local likelihood equations by multiplying the data's contribution to the log likelihood with a kernel weight:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i; x_i, \theta) K\left(\frac{x_i - x}{h}\right) = 0.$$

This equation is solved for θ to obtain the local constant estimator at x. For logistic regression models, the parameter θ might represent the logit of the success probability. This local likelihood estimator is reconsidered by Zhao (1994) and Chaudhuri and Dewanji (1995).

An interesting extension of the local polynomial estimators (in the classical regression model) to the case of generalized linear models (GLM) and quasilikelihood functions, is proposed by Fan, Heckman and Wand (1995). If the probability density function f belongs to a one-parameter exponential family model, the estimators for the parameter $\theta(x)$ and its derivatives are obtained by solving the set of equations

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i; \sum_{j=0}^{p} \beta_j (x_i - x)^j)(x_i - x)^k K\left(\frac{x_i - x}{h}\right) = 0, \qquad k = 0, \ldots, p.$$

This methodology is extended in Kauermann and Tutz (1996) to varying coefficient models. These models result from generalized linear models by allowing the pa-

rameter of the linear predictor to vary across some additional explanatory variable, called the effect modifier.

Although the formerly mentioned classes of models are quite extensive, they are essentially still restricted to the one-parameter case: the mean response is the main function of interest. For clustered binary data, we ideally want to use a *two-parameter* model: one parameter representing the mean response and the other describing the association of members within a cluster. When studying e.g. trends in sample extremes, one might use a *three parameter* likelihood model, where location, scale and shape vary according to smooth functions over time (Davison and Ramesh, 1998).

None of the local polynomial smoothing methods discussed above is able to handle such more complicated data structures. In Aerts and Claeskens (1997), see Chapter 2, we propose local polynomial estimators in a multiparameter likelihood setting, where the probability density function of the response Y might depend on κ parameters: $f(Y_i; \theta_1(x), \ldots, \theta_\kappa(x))$. Each of the curves $\theta_j(\cdot)$ is locally approximated by a polynomial of, say, degree p_j. An important advantage of our approach is that all curves are estimated simultaneously. Even for Gaussian data, this option has not been explored before. For example, Ruppert, Wand, Holst and Hössjer (1997) obtain local polynomial variance function estimators, but only after estimators for the mean are obtained.

Recently, other extensions of the classical techniques have been proposed. Staniswalis and Lee (1998) apply local constant smoothers ($p = 0$) for multiple regression in a longitudinal data setting (observations are correlated over time). Betensky (1997) used nearest neighbor local likelihood techniques to estimate smooth population and individual growth curves also in a longitudinal data setting. Tutz and Kauermann (1997) define local likelihood estimators in multivariate generalized linear models with varying coefficients. Fan, Farmen and Gijbels (1998) is a related reference for local polynomial estimators in one-parameter likelihood models.

Other approaches in multiparameter models

Carroll, Ruppert and Welsh (1998) propose local versions of general estimating equations, which might contain several parameters. Their equations are very similar to those of Aerts and Claeskens (1997). More details are given in Section 2.7 and Chapter 4.

When there is more than one covariate in the data set, all of the above estimation techniques have to deal with the so-called curse of dimensionality. Additive models are a popular alternative, see, e.g., Hastie and Tibshirani (1990). Opsomer and Ruppert (1997) study asymptotic properties of the local polynomial estimators in additive classical regression models. Semiparametric models are studied by Opsomer and Ruppert (1998), in the same context. Alternatives to the additive models are provided by the partially linear single-index models (Carroll, Fan, Gijbels and Wand, 1997). Claeskens and Aerts (1999), see Chapter 4, study local polynomial estimators in multiparameter additive models for general estimating equations, and indicate how to extend the obtained results to multiparameter semiparametric models.

Another extensive class of smoothers are the spline estimators. There are two general approaches: regression splines (where a set of knots has to be specified) and smoothing splines (where the data values themselves are the knots). For a detailed explanation and for more references we refer to Eubank (1988) and to Green and Silverman (1994).

Yee and Wild (1996) and Wild and Yee (1996) define a generalization of the cubic smoothing splines which is applicable in multiparameter models. They consider a particular extension of the generalized additive models. Their vector smoothing splines are also applicable in combination with GEE2 methods (generalized estimating equations, Zhao and Prentice, 1990).

Ruppert and Carroll (1997) combine features of smoothing and regression splines to obtain the penalized regression splines. The jumps at the knots are shrunk towards zero by using a penalty function. This penalization makes the choice of knots less important. While they mainly focus the applications to multivariate (classical) regression models, Aerts, Claeskens and Wand (1999), see Chapter 5, study the asymptotic mean squared error properties (AMSE) of this type of estimator in a wide variety of settings, including semiparametric models, generalized additive models, generalized estimating equations and multiparameter likelihood models.

Model checking by kernel or spline-based smoothers

Often it is stated that smoothing methods can be used as a diagnostic tool to detect departures from a fitted parametric model. For example, Copas (1983) makes a graphical comparison of a parametric regression function with a nonparametric kernel estimate. This informal model checking is sometimes not sufficient and the-

oretical sound lack of fit tests are called for. Since smoothing methods relax the parametric assumptions on the shape of the functional relationships, it is quite natural to base a test on a comparison of parametric and nonparametric results.

There is a large variety of tests available, most of them focus on the classical regression context. We will not even try to give a complete overview, but some relevant references are Yanagimoto and Yanagimoto (1987), Cox, Koh, Wahba and Yandell (1988), Eubank and Spiegelman (1990) and Chen (1994), who all used smoothing splines. Eubank and Spiegelman (1990) and Eubank, Hart and LaRiccia (1993) apply Fourier series estimators. Bjerve, Doksum and Yandell (1985) used nearest neighbors and based a test on simultaneous confidence intervals. There are many references to tests based on kernel smoothers, such as Raz (1990), Hart and Wehrly (1992), Müller (1992) and Li (1998a). King, Hart and Wehrly (1991) and Young and Bowman (1995) test for the equality of regression curves. The distribution of the resulting test statistic is usually quite complicated and often bootstrap techniques are used (see also below for more references on bootstrap methods). An important early reference is Härdle and Mammen (1983), who compared nonparametric and parametric regression fits using the wild bootstrap in combination with local constant estimators.

A test specifically developed for application with (assumed to be independent) binary data is constructed by Azzalini, Bowman, Härdle (1988), who defined a kind of likelihood ratio test to compare the parametric and nonparametric curve. Asymptotic results for this type of test are given by Severini and Staniswalis (1991) in a one-parameter local likelihood framework. Firth, Glosup, Hinkley (1991) also developed a test based on local likelihood estimators. An alternative has been provided by le Cessie and van Houwelingen (1991) whose test statistic is a weighted sum of local constant smoothed residuals. Rodríguez-Campos, González-Manteiga and Cao (1998) define a test applicable in binary response regression in generalized linear models, and they too use bootstrap methods. In le Cessie and van Houwelingen (1993) the problem with the selection of the bandwidth is circumvented by calculating the test statistics for a collection of different bandwidths. For tests specifically in the generalized linear models context, we refer to Li (1998b) and Kauermann and Tutz (1996).

To our knowledge, no tests based on smoothing methods are developed for general multiparameter likelihood models, and none of the smoothing based tests for

binary response data takes explicitly the possible correlation between outcomes into account.

Model checking by orthogonal series estimators

Although the above tests weren't considered in multiparameter likelihood models, it is to be expected that the same difficulties with the bandwidth and the bias will appear in multiparameter models. Therefore we will consider another class of smoothers: the *orthogonal series estimators*. The smoothing aspect comes into play by taking only a finite number of terms of this series. The smoothing parameter is the series truncation point, which is discrete. This discreteness will make the selection of the truncation point somewhat more feasible. For a comprehensive account on orthogonal series estimators and lack of fit tests in classical regression models, see Hart (1997) and references therein.

Eubank and Hart (1992) constructed "order selection tests" based on orthogonal series estimators for testing the fit of a classical regression model. The particular null model of interest is extended with terms from some orthogonal series and the series truncation point is used as the test statistic. The idea of this kind of tests originates from the goodness of fit tests, where one wants to know whether a set of i.i.d. observations arises from a particular class of probability distributions (Neyman, 1937, Kallenberg and Ledwina, 1995).

Aerts, Claeskens and Hart (1999), see Chapter 6, introduce omnibus lack of fit tests based on orthogonal series estimators in multiparameter likelihood models with a single covariate. Several methods for smoothing parameter selection are proposed, some of them are also valid in even more general models, where it is not necessary to specify the full distribution of the response (GEE, misspecification).

In Aerts, Claeskens and Hart (1998), see Chapter 7, the extension to multiple covariates is made.

1.3 Bootstrap methods

As already mentioned before, bootstrap methods are often applied to approximate the distribution of test statistics and to obtain P-values of a test. After the publication of Efron's (1979) article on bootstrap methods, the use of resampling techniques has known an enormous growth. The idea of bootstrap methods is to replace the

classical approximations to biases, variances, distributions,... by simulation based alternatives. We refer to Efron and Tibshirani (1993), Shao and Tu (1995) and Davison and Hinkley (1997) for more information on bootstrap methods and their various applications.

Bootstrap and correlated data

References to bootstrap mechanisms which are applicable in general multiparameter likelihood models do not abound. Although Burke and Gombay (1991) consider likelihood models with several parameters, they still assume the observations to be independent and identically distributed, which excludes the regression context. Considerable interest has been in generalized linear models. Moulton and Zeger (1988, 1991) constructed bootstrap estimators by resampling from the standardized Pearson residuals and then calculating the estimators by only the first step of the iterative estimating procedure. Residual resampling is also used by Simonoff and Tsai (1988) in quasi-likelihood models. For correlated outcomes in generalized linear models, Sherman and le Cessie (1997) do not separate the data of members of the same cluster. By resampling these groups of data, their bootstrap samples retain the correct dependence structure within a cluster.

For clustered binary data, Lockhart, Piegorsch and Bishop (1992) resample for each dose level the couples $(y_{i\bullet}, m_i)$ where $y_{i\bullet}$ is the number of positive binary responses per cluster and m_i the cluster size. Next, they use these resampled data mainly to estimate the variance of the estimated proportions. Also Carr and Portier (1993) ignore the dependence structure once a resampled set of data is obtained. Frangos and Schucany (1995) propose a particular kind of parametric bootstrap, with still a strong nonparametric flavor, to construct confidence intervals in toxicological experiments.

Although testing the null hypothesis of no dose effect is usually of great interest in toxicity studies, in none of these papers the bootstrap methodology is used to approximate P-values of tests. In general, bootstrap methods are hardly ever applied for hypothesis testing in parametric models. The main reason for this is that for the bootstrap to work, data have to be generated under the restrictions imposed by the specific null hypothesis.

A parametric bootstrap

By assuming the likelihood model of the data to be known up to some finite dimensional parameter, the generation of bootstrap data under the null hypothesis is easily taken care of by the parametric bootstrap (Beran, 1988). In this approach one samples from the data's distribution function where estimated null parameters replace the "true" (and unknown) model parameters. This bootstrap method is studied by Aerts and Claeskens (1999), see Chapter 8, in the context of testing hypothesis by means of a likelihood ratio, Wald or score test. Also when pseudolikelihood methods are used, they show this bootstrap technique to be very useful in approximating the distribution of test statistics.

A semiparametric bootstrap

One might expect that a parametric bootstrap will break down if the likelihood model of the data is grossly misspecified. In that case, a nonparametric bootstrap (resampling the data) or a semiparametric bootstrap (resampling residuals) should be preferred. Resampling the data is often not possible, since the resampled set should reflect the null hypothesis. Also resampling residuals can give problems in multiparameter likelihood models, since no simple additive error structure might be available. Even defining residuals which can be used in the bootstrap procedure is not always possible. Rather than considering these approaches, Hu and Zidek (1995) propose a method which resamples summands in the estimating function of the linear regression model, which was used to produce the original estimates.

In Aerts and Claeskens (1998a), see Chapter 9, this idea is extended to general multiparameter likelihood models. In their semiparametric bootstrap, one resamples from the first and second order partial derivatives of the log likelihood function, or equivalently, from the summands of the estimating equations and their respective derivatives. By constructing one-step estimators, motivated by linear and quadratic approximations to the estimators, Aerts and Claeskens (1998a) are able to obtain bootstrap estimators directly (without any additional iterations) which reflect the restrictions imposed by the null hypothesis. The use of a one-step bootstrap method dates back to Schucany and Wang (1991), see also Section 5.4 in Shao and Tu (1995). It is an approximate method to simplify the computations for estimators which have to be computed iteratively, as clearly is the case for most clustered binary data

models.

Aerts and Claeskens (1998b), see Chapter 10, explore the quadratic approximation somewhat further. It can also be an interesting method to obtain bias corrected estimators or to construct confidence intervals for the model parameters.

1.4 Bootstrap and smoothing

The use of bootstrap methods is of course not restricted to parametric models. It gets a bit more complicated though in nonparametric models. The problems arise because of the bias of the estimators, when the so-called optimal bandwidth is used. This bandwidth balances squared bias and variance in such a way that they tend to zero at the same rate. The "naive" bootstrap approach of just resampling the data vectors is inappropriate because the bootstrap bias would be zero.

Resampling methods

Most existing resampling schemes in the context of local estimators make use of some kind of residuals. For classical regression models with independent errors, Härdle and Bowman (1988) resample the centered residuals and explicitly estimate the bias of the Priestley and Chao (1972) estimator. This estimated bias term is then used in the bootstrap construction. Another important example is the wild bootstrap (Wu, 1986, Härdle and Mammen, 1993). A two-point distribution is defined which has mean zero, variance equal to the square of the residual and third moment equal to the cube of the residual. It is called the wild bootstrap because the resampling distribution can be thought of as an attempt to reconstruct the distribution of each residual through the use of a single observation.

A variant on this is the moment-oriented bootstrap (Bunke, 1997) which is also a wild bootstrap, based on local estimators of the first four error moments that are smoothed by Gasser-Müller (1979) kernel estimators. A comparison of these two wild bootstrap methods in the context of a classical regression model with heteroscedastic errors has been made by Sommerfeld (1997). In all cases, bootstrap response values are constructed as the sum of an initial estimator for the mean response $\theta(x_i)$ and the resampled error term. A direct application of this approach for binary data, would result in bootstrapped data which take on values different from zero and one. Rounded values could be used in that case.

The smoothed bootstrap of Cao-Abad and González-Manteiga (1993) is another resampling scheme in classical regression models with a random covariate. In contrast to the wild bootstrap which is conditional on the original covariate values, the smoothed bootstrap also resamples the explanatory variable.

A parametric bootstrap is used by Aerts and Veraverbeke (1995) in polytomous regression models, where Gasser-Müller kernel estimators are used to obtain the vector of smoothed probabilities.

To avoid the explicit definition of exchangeable residuals, Kauermann and Tutz (1996) define, in the context of local likelihood estimation in generalized linear models, a bootstrap estimator based on the first step in a Fisher scoring algorithm. This one-step estimator does not need any additional iterative model fitting in the bootstrap simulation. In Claeskens and Aerts (1998), see Chapter 3, we propose to sample directly from the set of first and second partial derivatives of the kernel weighted log likelihood contributions. Next, these bootstrapped quantities are used to construct the bootstrap estimators in only one step. This method can be seen as an extension of the one-step bootstrap estimator in the fully parametric likelihood models.

Confidence intervals

As a by-product of the resampling methods, simultaneous bootstrap confidence intervals over a finite grid can be obtained. See Kauermann and Tutz (1996) for confidence intervals in varying coefficient models and Claeskens and Aerts (1998) for application to multiparameter likelihood models, and in particular to clustered binary data models.

Some relevant references concerning the construction of confidence intervals and bands in classical regression models are Härdle and Bowman (1988), Härdle and Marron (1991), Xia (1998) and Neumann and Polzehl (1998). Rodríguez-Campos and Cao-Abad (1993) construct confidence intervals for the mean in independent binary outcome regression models, by applying the simple local constant estimator (1.1) directly to the observed binary variables.

1.5 Outline

Chapter 2 defines local polynomial estimators in multiparameter likelihood models. We show the weak consistency and joint asymptotic normality of the estimators of the unknown curves and of all their derivatives up to the order of the polynomial chosen. A likelihood-based cross-validation method provides a data-driven bandwidth.

Chapter 3 builds further on this idea and extends the previous results in several directions. Under the appropriate regularity conditions, we now show that the estimators are within a more general framework strongly consistent: we no longer assume that the likelihood model is correctly or fully specified. We propose and study a bootstrap resampling scheme based on a one step approximation of the estimators. Next to point estimators we also construct simultaneous confidence intervals for $\{\theta_j^{(k)}(x) : j = 1, \ldots, \kappa; x \text{ belongs to a finite grid}\}$. The construction of these intervals follows naturally from the one-step bootstrap approach.

Chapter 4 extends the area of application of the estimation techniques to the multiple regression setting. To circumvent the curse of dimensionality, which arises because of the sparseness of data in high dimensional problems, we consider additive functional relationships in nonparametric models, and discuss the extension to semiparametric models.

Chapter 5 is a short visit to another class of smoothers: the penalized regression splines. In general additive models, we study the asymptotic mean average squared error, examine an approximation of the degrees of freedom, and suggest a selection method for the penalization constant.

Chapter 6 considers lack of fit tests based on orthogonal series estimators in regression models with one continuous covariate. We propose several estimators of the series truncation point and illustrate the generality of the tests by several examples.

Chapter 7 compares some other tests for the one-covariate case and extends the results of Chapter 6 to the multiple covariate case.

Chapter 8 studies properties of a parametric bootstrap resampling scheme which is used for testing function fit against a specific parametric alternative. Simulated power and type I error probability of Wald, score and likelihood ratio test statistics are compared.

Chapter 9 proposes a semiparametric bootstrap method which remains valid when the assumed likelihood of the data is incorrect or not completely specified. Since bootstrap parameter estimates, which reflect the null hypothesis, are generated directly, no additional iterative model fitting is required. Both a one-step linear and a one-step quadratic bootstrap are studied.

Chapter 10 takes a closer look at the one-step quadratic bootstrap. In this chapter we use this bootstrap method for bias correction and for the construction of confidence intervals.

To implement the proposed methods, we mainly developed our own functions and procedures using the programming languages S-Plus and Gauss. Programs to fit clustered binary data models are based on Gauss procedures for parametric models, which were kindly made available by G. Molenberghs. S-Plus functions developed by M. Wand are used for the implementation of the results of Chapter 5.

Part I

Smooth Curve Estimation

Chapter 2

Local Polynomial Estimation in Multiparameter Likelihood Models

2.1 Introduction

The purpose of this chapter is to study local polynomial estimators in multiparameter likelihood models. The main motivation is the need for nonparametric alternatives to parametric dose-response models for clustered binary data. Such data arise in developmental toxicity studies, designed to assess the potential adverse effects of drugs and other exposures on developing foetuses of pregnant rodents (dams). A typical experiment includes a control group and some dosed groups with special attention for the low doses. Exposure usually occurs early in gestation, the dams are sacrificed prior to term and the uterine contents examined for defects. The analysis of the resulting binary data (malformation yes/no) must account for the litter effect induced by the clustering of offspring within dams. Different types of models (marginal, conditional, random effects models) are available (for a recent survey, see Pendergast et al., 1996). Next to the malformation probability, these models all include one or more parameters to describe the association between outcomes. More details are given in Sections 2.6.1 and 3.2.2.

These clustered binary data have been analyzed mainly in a parametric way. The selection of the proper functional forms, describing the dependence of the malformation rate and the intra-litter correlation on dose in a specific probability model,

is not always an easy task. A nonparametric estimator can be very useful as a diagnostic tool. For an overview of diagnostic tools in parametric likelihood models, see Davison and Tsai (1992) and in particular for logistic regression, see Landwehr, Pregibon and Shoemaker (1984). In an explorative way, a parametric model can be graphically compared with its nonparametric alternative. In a further stage, a formal test statistic to examine the appropriateness of a certain hypothesized parametric function can be developed. For binomial data (no clustering), kernel estimates of the dose-response curve by locally averaging the sample proportions have been studied by Copas (1983), Staniswalis and Cooper (1988), Müller and Schmitt (1988) and Aragaki and Altman (1998).

In the one-parameter case, several authors have examined strategies to implement nonparametric estimation procedures in likelihood based regression models. Tibshirani and Hastie (1987) introduced the concept of local likelihood estimation. Staniswalis (1989) considered kernel smoothers maximizing a kernel weighted likelihood function. The asymptotic properties of this type of estimator based on more general weights were studied by Chaudhuri and Dewanji (1995). Fan, Heckman and Wand (1995) introduced local polynomial estimators in one-parameter exponential family and quasi-likelihood models. Recently, many other applications have been introduced and studied. For relevant references on this subject, refer to Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996) and Hart (1997).

Local polynomial fitting has become the standard in kernel smoothing. The corresponding smoothers are known to have several advantages in comparison with other linear smoothers, such as the behavior at the boundary (see, e.g., Wand and Jones, 1995; Fan and Gijbels, 1996). The local polynomial maximum likelihood estimator (MLE) is defined in Section 2.2 for the case of a fixed design. This section also introduces some notation and the required regularity assumptions on the distribution of the response variable as well as smoothness conditions on the parameters, conditions on the design, on the kernel function and on the smoothing parameter.

In Section 2.3 it is shown that the local likelihood variance estimator is asymptotically equivalent to the estimator proposed by Ruppert, Wand, Holst and Hössjer (1997). The basic asymptotic properties of the proposed local polynomial estimators are investigated in Section 2.3 (consistency and asymptotic normality). These results are reconsidered for the random design case and extended to more samples in

Section 2.4. The problem of choosing the smoothing parameter is briefly discussed in Section 2.5. Section 2.6 shows the applicability of the method in practice on data sets and contains the results of a small simulation experiment.

Most of the contents of this chapter can also be found in Aerts and Claeskens (1997).

2.2 Local likelihood estimation

Consider a response variable Y, with probability density function (pdf) $f(y; \theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x}))$ involving κ parameters depending on an explanatory variable $\boldsymbol{x} = (x_1, \ldots, x_d)$. In this generalized regression setting, the form of f is supposed to be known and the κ parameters are unknown real-valued functions of the covariate vector $\boldsymbol{x}$. In parametric theory, the functions $\theta_\ell(\boldsymbol{x})$ are modeled globally in terms of a finite number of regression coefficients, e.g. as

$$\theta_\ell(x_1, \ldots, x_d) = \beta_{\ell 0} + \beta_{\ell 1} x_1 + \ldots + \beta_{\ell d} x_d.$$

The maximum likelihood estimators $\hat{\beta}_{\ell j}$ maximize the log-likelihood function

$$\sum_{i=1}^{n} \log f(Y_i; \theta_1(\boldsymbol{x}_i), \theta_2(\boldsymbol{x}_i), \ldots, \theta_\kappa(\boldsymbol{x}_i))$$

as a function of the $\beta_{\ell j}$ $(\ell = 1, \ldots, \kappa; j = 0, \ldots, d)$. Under certain regularity conditions on the pdf $f(y; \theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x}))$, the MLEs $\hat{\beta}_{\ell j}$ are consistent and asymptotically normal (see e.g. Lehmann, 1983).

Instead of assuming a specific functional form that specifies how the explanatory variables $\boldsymbol{x}$ determine the distribution of the dependent variable Y, one can allow the data to describe this relationship nonparametrically, only requiring some weak smoothness assumptions. Without loss of generality, we restrict attention to two parameter probability models ($\kappa = 2$) and one explanatory variable $x \in [b_1, b_2]$ $(d = 1)$.

Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be a sample. Following Sacks and Ylvisaker (1970), we assume a fixed design generated by a known pdf $f_X(x)$ with $\text{supp}(f_X) = [b_1, b_2]$, i.e. for $i = 1, \ldots, n$,

$$x_i = G^{-1}\left(\frac{i-1}{n-1}\right) \quad \text{with} \quad G(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{2.1}$$

This design density f_X gives the experimenter some freedom in "distributing" the design points x_i. The random design case can be treated in a very similar way and is

briefly discussed in Section 2.4. For a kernel function K and a bandwidth parameter h, let $K_h(\cdot) = K(\cdot/h)/h$. For a fixed $x \in [b_1, b_2]$ denote $\boldsymbol{\theta}_r^T(x) = (\theta_{r0}(x), \ldots, \theta_{rp_r}(x))$ the vector containing the higher order derivatives of $\theta_r(x)$ $(r = 1, 2)$, more precisely $\theta_{rj}(x) = \theta_r^{(j)}(x)/j!$, $j = 0, \ldots, p_r$. Local polynomial fitting provides estimates for $\theta_r(x)$ and its derivatives up to order p_r $(r = 1, 2)$. The local polynomial MLE

$$(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T) = (\hat{\beta}_{10}, \ldots, \hat{\beta}_{1p_1}, \hat{\beta}_{20}, \ldots, \hat{\beta}_{2p_2})$$

maximizes the kernel weighted log-likelihood function

$$\mathcal{L}_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i; \sum_{j=0}^{p_1} \beta_{1j}(x_i - x)^j, \sum_{j=0}^{p_2} \beta_{2j}(x_i - x)^j) K_h(x_i - x) \qquad (2.2)$$

with respect to $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T) = (\beta_{10}, \ldots, \beta_{1p_1}, \beta_{20}, \ldots, \beta_{2p_2})$. The data are centered about x and each individual log-likelihood contribution is multiplied by a weight, governed by the kernel K and the bandwidth h. In this way, those observations x_i close to x have a larger impact on the maximization process. Of course, instead of the smoothing weights $w_{ni}(x) = (1/n)K_h(x_i - x)$, other nonnegative weights satisfying $\lim_{n\to\infty} \sum_{i=1}^{n} w_{ni}(x) = 1$ could be chosen; see, e.g., Gasser and Müller (1979) and Eubank (1988).

Taking a one-parameter family and $p_1 = 0$ (local constant) the aforementioned estimator has been studied by Staniswalis (1989) and Chaudhuri and Dewanji (1995). Fan, Heckman and Wand (1995) discussed the local polynomial estimator for $f(y; \theta(x))$ a one-parameter exponential family member.

To examine the asymptotic properties, we will need typical likelihood regularity conditions on the pdf $f(y; \theta_1, \theta_2)$ and smoothness conditions on the parameter functions $\theta_1(x), \theta_2(x)$. Denote

$$q_r(y; v_1, v_2) = \frac{\partial}{\partial u_r} \log f(y; u_1, u_2)|_{(u_1,u_2)=(v_1,v_2)} \quad r = 1, 2$$

$$q_{rs}(y; v_1, v_2) = \frac{\partial^2}{\partial u_r \partial u_s} \log f(y; u_1, u_2)|_{(u_1,u_2)=(v_1,v_2)} \quad r, s = 1, 2$$

$$q_{rst}(y; v_1, v_2) = \frac{\partial^3}{\partial u_r \partial u_s \partial u_t} \log f(y; u_1, u_2)|_{(u_1,u_2)=(v_1,v_2)} \quad r, s, t = 1, 2.$$

(R1) The densities $f(y; \theta_1, \theta_2)$ have a common support. There exists an open subset Θ of the parameter space containing the true parameters $(\theta_1(x), \theta_2(x))$ such that for almost all y the density $f(y; \theta_1, \theta_2)$ admits all third derivatives for all $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta$.

(R2) $E_{\boldsymbol{\theta}}[q_r(Y;\theta_1,\theta_2)] = 0 \quad$ for all $\boldsymbol{\theta} \in \Theta \;\; (r = 1,2)$.

(R3) $I_{rs}(\theta_1,\theta_2) = E_{\boldsymbol{\theta}}[q_r(Y;\theta_1,\theta_2)q_s(Y;\theta_1,\theta_2)] = E_{\boldsymbol{\theta}}[-q_{rs}(Y;\theta_1,\theta_2)]$ for all $\boldsymbol{\theta} \in \Theta$ $(r,s = 1,2)$ and $I_{rs}(\theta_1,\theta_2)$ is Lipschitz continuous and differentiable at $(\theta_1(x)$, $\theta_2(x))$ and the information matrix

$$\boldsymbol{I}(\theta_1(x),\theta_2(x)) = \begin{pmatrix} I_{11}(\theta_1(x),\theta_2(x)) & I_{12}(\theta_1(x),\theta_2(x)) \\ I_{21}(\theta_1(x),\theta_2(x)) & I_{22}(\theta_1(x),\theta_2(x)) \end{pmatrix}$$

is positive definite.

(R4) There exists a function $H(y)$ such that $|q_{rs}(y;\theta_1,\theta_2)| \leq H(y)$ for all $\boldsymbol{\theta} \in \Theta$ $(r,s = 1,2)$ and $E_{\boldsymbol{\theta}}[H^2(Y)]$ is uniformly bounded on Θ.

(R5) There exists a function $J(y)$ such that $|q_{rst}(y;\theta_1,\theta_2)| \leq J(y)$ for all $\boldsymbol{\theta} \in \Theta$ $(r,s,t = 1,2)$ and $E_{\boldsymbol{\theta}}[J^2(Y)]$ is uniformly bounded on Θ.

(R6) There exists a $\delta > 0$ such that $E_{\boldsymbol{\theta}}[|q_r(Y;\theta_1,\theta_2)|^{2+\delta}] < \infty \quad$ for all $\boldsymbol{\theta} \in \Theta \;\; (r = 1,2)$.

(S) $\theta_r(x)$ has a (p_r+1)th $((p_r+2)$nd) derivative for p_r odd (p_r even), $r = 1,2$.

We also need the following standard assumptions on the design density f_X, the kernel K and the bandwidth h:

(G) $f_X(.)$ is differentiable in $[b_1,b_2]$ and $\inf_{x\in[b_1,b_2]} f_X(x) > 0$.

(K) K is a Lipschitz continuous symmetric pdf on $[-1,1]$.

(H) $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

The final asymptotic expressions have a different interpretation depending on x being a boundary point or not. For $h < (b_2 - b_1)/2$, the region of interior points equals $[b_1 + h, b_2 - h]$. Points at the left of $b_1 + h$ (right of $b_2 - h$) are left (resp. right) boundary points. Define

$$\mathcal{R}_x = \begin{cases} [-\alpha,1] & x = b_1 + \alpha h \text{ for some } 0 \leq \alpha < 1 \\ [-1,1] & b_1 + h \leq x \leq b_2 - h \\ [-1,\alpha] & x = b_2 - \alpha h \text{ for some } 0 \leq \alpha < 1 \end{cases} \tag{2.3}$$

and $\quad \nu_\ell(\mathcal{R}_x) = \int_{\mathcal{R}_x} z^\ell K(z)dz.$

2.3 Consistency and asymptotic normality

The first theorem guarantees the existence of at least one solution of the likelihood equations that is consistent.

Theorem 2.1 *Suppose that the regularity conditions (R1)-(R5) on pdf $f(y;\theta_1,\theta_2)$ and conditions (G),(K) and (H) on design, kernel K and bandwidth h hold. If the functions $\theta_1(x)$ and $\theta_2(x)$ satisfy smoothness condition (S), then there exist solutions $(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)$ of the likelihood equations*

$$\frac{\partial}{\partial\beta_{rj}}[\mathcal{L}_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)]=0,\quad j=0,\ldots,p_r;\ r=1,2 \tag{2.4}$$

such that $\hat{\beta}_{rj}$ is consistent for estimating $\theta_{rj}(x)$ for $j=0,\ldots,p_r$; $r=1,2$.

The following lemma will be used frequently. The proof is straightforward.

Lemma 2.1 *Let $L(\cdot)$ and $S(\cdot)$ be bounded and Lipschitz continuous functions. Then, for the fixed design points (2.1),*

$$\frac{1}{n}\sum_{j=1}^{n}L(\frac{x_j-x}{h})S(x_j)=\int_0^1 L(\frac{G^{-1}(y)-x}{h})S(G^{-1}(y))dy+O(\frac{1}{nh}).$$

Proof of Theorem 2.1. Expanding the weighted log-likelihood function (2.2) about the point $(\boldsymbol{\theta}_1(x),\boldsymbol{\theta}_2(x))$, we get

$$\begin{aligned}
\mathcal{L}_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)-\mathcal{L}_n(\boldsymbol{\theta}_1(x),\boldsymbol{\theta}_2(x)) &= \frac{1}{n}\sum_{r=1}^{2}\sum_{k=0}^{p_r}A_{rk}^n(x)(\beta_{rk}-\theta_{rk}(x))\\
&+\frac{1}{2n}\sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{k=0}^{p_r}\sum_{\ell=0}^{p_s}B_{rsk\ell}^n(x)(\beta_{rk}-\theta_{rk}(x))(\beta_{s\ell}-\theta_{s\ell}(x))\\
&+\frac{1}{6n}\sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{t=1}^{2}\sum_{k=0}^{p_r}\sum_{\ell=0}^{p_s}\sum_{m=0}^{p_t}(\beta_{rk}-\theta_{rk}(x))(\beta_{s\ell}-\theta_{s\ell}(x))(\beta_{tm}-\theta_{tm}(x))\\
&\quad\times\sum_{i=1}^{n}K_h(x_i-x)(x_i-x)^{k+\ell+m}\gamma(Y_i)J(Y_i)\\
&= S_{1n}+S_{2n}+S_{3n}
\end{aligned}$$

where

$$A_{rk}^n(x) = \sum_{i=1}^{n}K_h(x_i-x)(x_i-x)^k q_r(Y_i;\bar{\theta}_1(x,x_i),\bar{\theta}_2(x,x_i)) \tag{2.5}$$

$$B_{rsk\ell}^n(x) = \sum_{i=1}^{n}K_h(x_i-x)(x_i-x)^{k+\ell} q_{rs}(Y_i;\bar{\theta}_1(x,x_i),\bar{\theta}_2(x,x_i)) \tag{2.6}$$

and

$$\bar{\theta}_r(x, x_i) \;=\; \sum_{j=0}^{p_r} \theta_{rj}(x)(x_i - x)^j. \tag{2.7}$$

By condition (R5), $|\gamma(y)| \leq 1$.

A main point in studying the asymptotic properties of $A^n_{rk}(x)$ and $B^n_{rskl}(x)$ is the fact that, for $|x_i - x| \leq h$ and $s = 1, 2$,

$$\theta_s(x_i) = \bar{\theta}_s(x, x_i) + \frac{\theta_s^{(p_s+1)}(x)}{(p_s+1)!}(x_i - x)^{p_s+1} + \frac{\theta_s^{(p_s+2)}(x)}{(p_s+2)!}(x_i - x)^{p_s+2} + o(h^{p_s+2}) \tag{2.8}$$

with $\bar{\theta}_s(x, x_i)$ as in (2.7). Only if p_s is even, it is required to specify the third term in the right side of (2.8).

Using (R1)-(R3), we have that $E[\frac{1}{n}A^n_{rk}(x)] =$

$$\frac{1}{n}\sum_{s=1}^{2}\sum_{i=1}^{n} K_h(x_i - x)(x_i - x)^k E[q_{rs}(Y_i; \eta_1(x, x_i), \eta_2(x, x_i)))(\bar{\theta}_s(x, x_i) - \theta_s(x_i)]$$

with $\eta_r(x, x_i)$ between $\bar{\theta}_r(x, x_i)$ and $\theta_r(x_i)$. Using (R4) and $\sum_{i=1}^n K_h(x_i - x)/n = O(1)$, the polynomial expansion (2.8) implies that

$$E[\frac{1}{n}A^n_{rk}(x)] = o(h^{min(p_1,p_2)})$$

and that

$$\mathrm{Var}[\frac{1}{n}A^n_{rk}(x)] \;\leq\; \frac{1}{n^2}\sum_{i=1}^{n} K_h^2(x_i - x)(x_i - x)^{2k} E[q_r^2(Y_i; \bar{\theta}_1(x, x_i), \bar{\theta}_2(x, x_i))].$$

Similar arguments as for the bias show that

$$\mathrm{Var}[\frac{1}{n}A^n_{rk}(x)] = O(\frac{h^{2k}}{nh})$$

and hence, as $n \to \infty$,

$$\frac{1}{n}A^n_{rk}(x) \;\xrightarrow{P}\; 0. \tag{2.9}$$

Using Lemma 2.1, (R5) and the Lipschitz continuity of K,

$$E[\frac{1}{n}B^n_{rsk\ell}(x)] \;=\; -f_X(x)I_{rs}(\theta_1(x), \theta_2(x))\nu_{k+\ell}(\mathcal{R}_x)h^{k+\ell} + o(h^{\min(p_1,p_2)})$$

and

$$\mathrm{Var}[\frac{1}{n}B^n_{rsk\ell}(x)] = O(\frac{h^{2(k+\ell)}}{nh}).$$

This shows that, as $n \to \infty$,

$$\frac{1}{n}B^n_{rsk\ell}(x) \quad \xrightarrow{P} \quad -f_X(x)I_{rs}(\theta_1(x),\theta_2(x))\nu_0(\mathcal{R}_x)\delta_{k0}\delta_{\ell 0} \tag{2.10}$$

(with δ_{ij} the Kronecker delta). From (2.9) and (2.10) it immediately follows that $S_{1n} \xrightarrow{P} 0$ and

$$S_{2n} \xrightarrow{P} -\frac{1}{2}f_X(x)\nu_0(\mathcal{R}_x) \begin{pmatrix} \beta_{10}-\theta_1(x) \\ \beta_{20}-\theta_2(x) \end{pmatrix}^T \boldsymbol{I}(\theta_1(x),\theta_2(x)) \begin{pmatrix} \beta_{10}-\theta_1(x) \\ \beta_{20}-\theta_2(x) \end{pmatrix}. \tag{2.11}$$

By condition (R3) the limit expression (2.11) is strictly negative.
For $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ in a sphere with center at $(\boldsymbol{\theta}_1^T(x), \boldsymbol{\theta}_2^T(x))$ and radius ε, $|S_{3n}| < C\varepsilon^3$ for some constant $C > 0$. Combining these results, the probability tends to 1 that $\mathcal{L}_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) < \mathcal{L}_n(\boldsymbol{\theta}_1(x),\boldsymbol{\theta}_2(x))$ for all $(\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T)$ in a ε-sphere about $(\boldsymbol{\theta}_1^T(x),\boldsymbol{\theta}_2^T(x))$. To complete the proof, proceed as in the proof of Theorem 6.4.1 in Lehmann (1983).

To formulate an asymptotic normality result, we need some further notation. It partly resembles the notation used in Ruppert and Wand (1994) and Fan, Heckman and Wand (1995).

$\boldsymbol{N}_{p_rp_s}(x), \boldsymbol{T}_{p_rp_s}(x)$ and $\boldsymbol{Q}_{p_rp_s}(x)$ $(r,s = 1,2)$ are matrices of dimension $(p_r+1)\times(p_s+1)$, of which the $(k+1,\ell+1)$th entry equals $\nu_{k+\ell}(\mathcal{R}_x)$, $\int_{\mathcal{R}_x} z^{k+\ell}K^2(z)dz$ and $\nu_{k+\ell+1}(\mathcal{R}_x)$ $(k = 0,\ldots,p_r;\ \ell = 0,\ldots,p_s)$. The matrix $\boldsymbol{M}_{tp_s}(z;x)$ is obtained by replacing in $\boldsymbol{N}_{p_sp_s}(x)$ the $(t+1)$th $(t = 0,\ldots,p_s)$ column by $(1\ z\ \ldots\ z^{p_s})^T$, and for $|\boldsymbol{N}_{p_sp_s}(x)| \neq 0$,

$$K_{tp_s}(z) = \frac{|\boldsymbol{M}_{tp_s}(z;x)|}{|\boldsymbol{N}_{p_sp_s}(x)|}K(z).$$

To simplify notation, the dependence of $K_{tp_s}(z)$ on $\mathcal{R}_x$ is omitted. Let

$$\boldsymbol{H}_{p_r} = \mathrm{diag}(1,h,\ldots,h^{p_r}) \quad (r = 1,2)$$

$$\boldsymbol{\Sigma}_x = f_X(x)\boldsymbol{I}(\theta_1(x),\theta_2(x)) \otimes \begin{pmatrix} \boldsymbol{N}_{p_1p_1}(x) & \boldsymbol{N}_{p_1p_2}(x) \\ \boldsymbol{N}_{p_2p_1}(x) & \boldsymbol{N}_{p_2p_2}(x) \end{pmatrix} \tag{2.12}$$

$$\boldsymbol{\Gamma}_x = f_X(x)\boldsymbol{I}(\theta_1(x),\theta_2(x)) \otimes \begin{pmatrix} \boldsymbol{T}_{p_1p_1}(x) & \boldsymbol{T}_{p_1p_2}(x) \\ \boldsymbol{T}_{p_2p_1}(x) & \boldsymbol{T}_{p_2p_2}(x) \end{pmatrix} \tag{2.13}$$

where $\otimes$ is a generalized Kronecker product: for a $(r \times c)$ matrix $\boldsymbol{A} = (a_{ij})$ and a partitioned matrix $\boldsymbol{B}$ with submatrices $\boldsymbol{B}_{ij}$ $(i = 1, \ldots, r,\ j = 1, \ldots, c)$, we define $\boldsymbol{A} \otimes \boldsymbol{B}$ as a partitioned matrix with submatrices $(a_{ij}\boldsymbol{B}_{ij})$, $i = 1, \ldots, r$, $j = 1, \ldots, c$. Note that if all submatrices $\boldsymbol{B}_{ij}$ are identical, then this product simplifies to the ordinary Kronecker product. In definitions (2.12) and (2.13) this will be the case when $p_1 = p_2$. Finally let

$$\boldsymbol{\Lambda}_x = \begin{pmatrix} \frac{d}{dx}(f_X(x)I_{11}(\theta_1(x), \theta_2(x))) & \frac{d}{dx}(f_X(x)I_{12}(\theta_1(x), \theta_2(x))) \\ \frac{d}{dx}(f_X(x)I_{21}(\theta_1(x), \theta_2(x))) & \frac{d}{dx}(f_X(x)I_{22}(\theta_1(x), \theta_2(x))) \end{pmatrix} \otimes \begin{pmatrix} \boldsymbol{Q}_{p_1p_1}(x) & \boldsymbol{Q}_{p_1p_2}(x) \\ \boldsymbol{Q}_{p_2p_1}(x) & \boldsymbol{Q}_{p_2p_2}(x) \end{pmatrix}. \quad (2.14)$$

The asymptotic normality of the local polynomial MLE essentially follows from that of $\boldsymbol{W}^n(x) = (\boldsymbol{W}_1^n(x)^T, \boldsymbol{W}_2^n(x)^T)^T$ where $\boldsymbol{W}_r^n(x)$ $(r = 1, 2)$ is a column vector with components

$$W_{rk}^n(x) = (nh^{2k-1})^{-1/2} \sum_{i=1}^{n} K_h(x_i - x)(x_i - x)^k q_r(Y_i; \bar{\theta}_1(x, x_i), \bar{\theta}_2(x, x_i)). \quad (2.15)$$

Our main result generalizes the main theorem of Fan, Heckman and Wand (1995) by the multiparameter setting in a general full likelihood approach, whereas they consider the nonparametric estimation of a single parameter, the mean function, in a quasi-likelihood setting. Let $\boldsymbol{I}_{p_1+p_2+2}$ be the identity matrix of order $p_1 + p_2 + 2$ and let $\boldsymbol{0}$ be a null matrix of appropriate dimensions.

Theorem 2.2 *Assume that (R6) and all conditions of Theorem 2.1 hold, then for* $n \to \infty$

$$(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Gamma}_x\boldsymbol{\Sigma}_x^{-1})^{-1/2} \Big(\sqrt{nh}\{\boldsymbol{H}_{p_1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\theta}_1(x)), \boldsymbol{H}_{p_2}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\theta}_2(x))\}^T -$$

$$(\boldsymbol{\Sigma}_x^{-1} - h\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1})E[\boldsymbol{W}^n(x)] \Big) \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p_1+p_2+2}).$$

Proof. By a Taylor expansion of $\dfrac{\partial}{\partial \beta_{rk}}\mathcal{L}_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ about $(\boldsymbol{\theta}_1^T(x),\ \boldsymbol{\theta}_2^T(x))$ $(r = 1, 2;$ $k = 1, \ldots, p_r)$, evaluated at the local polynomial MLE $(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$,

$$-W_{rk}^n(x) = \sum_{s=1}^{2} \sum_{\ell=0}^{p_s} C_{rks\ell}^n(x) V_{s\ell}^n(x)$$

with $W^n_{rk}(x)$ as defined in (2.15) and with (using notation introduced in the proof of Theorem 2.1) $V^n_{s\ell}(x) = \sqrt{nh}h^\ell(\hat{\beta}_{s\ell} - \theta_{s\ell}(x))$ and $C^n_{rks\ell}(x) =$

$$\frac{B^n_{rks\ell}(x)}{nh^{k+\ell}} + \frac{1}{2n}\sum_{t=1}^{2}\sum_{m=0}^{p_t}\sum_{i=1}^{n} h^m(\hat{\beta}_{tm} - \theta_{tm}(x))\gamma_1(Y_i)J(Y_i)(\frac{x_i - x}{h})^{k+\ell+m}K_h(x_i - x).$$

By condition (R5), $0 \leq |\gamma_1(y)| \leq 1$. The consistency of $(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$ (Theorem 2.1) and (R5) imply that, as $n \to \infty$,

$$C^n_{rks\ell}(x) - \frac{B^n_{rks\ell}(x)}{nh^{k+\ell}} \xrightarrow{P} 0.$$

Arguments analogous to the one used in the proof of (2.10) show that, as $n \to \infty$,

$$\begin{aligned} &C^n_{rks\ell}(x) + f_X(x)I_{rs}(\theta_1(x), \theta_2(x))\nu_{k+\ell}(\mathcal{R}_x) \\ &+h\frac{d}{dx}(f_X(x)I_{rs}(\theta_1(x), \theta_2(x)))\nu_{k+\ell+1}(\mathcal{R}_x) \xrightarrow{P} 0 \end{aligned}$$

or, in matrix-notation, $\boldsymbol{C}^n(x) + (\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x) \xrightarrow{P} \mathbf{0}$ with $\boldsymbol{C}^n(x)$ the partitioned matrix

$$\begin{pmatrix} \boldsymbol{C}_{11}(x) & \boldsymbol{C}_{12}(x) \\ \boldsymbol{C}_{21}(x) & \boldsymbol{C}_{22}(x) \end{pmatrix}$$

where $(\boldsymbol{C}_{rs}(x))_{k+1,\ell+1} = C^n_{rks\ell}(x)$, $k = 0, \ldots, p_r$, $\ell = 0, \ldots, p_s$ and $r, s = 1, 2$.
Next we turn to $\boldsymbol{W}^n(x) = (\boldsymbol{W}^n_1(x)^T, \boldsymbol{W}^n_2(x)^T)^T$ with components $W^n_{rk}(x)$.
$\boldsymbol{W}^n(x)$ can be written as a sum of independent random vectors

$$\boldsymbol{W}^n(x) = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\begin{pmatrix} \mathbf{Y}^\star_{i1} \\ \mathbf{Y}^\star_{i2} \end{pmatrix}$$

where, for $r = 1, 2$,

$$\mathbf{Y}^\star_{ir} = q_r(Y_i; \bar{\theta}_1(x, x_i), \bar{\theta}_2(x, x_i))\begin{pmatrix} 1 \\ (x_i - x)/h \\ \vdots \\ (x_i - x)^{p_r}/h^{p_r} \end{pmatrix} K(\frac{x_i - x}{h}).$$

From Lemma 2.1 and assumptions (R2) and (R3), it follows that, for $n \to \infty$,

$$Cov(W^n_{rk}(x), W^n_{s\ell}(x)) \to I_{rs}(\theta_1(x), \theta_2(x))f_X(x)\int_{\mathcal{R}_x} K^2(u)u^{k+\ell}du$$

or, in matrix-notation,$Cov(\boldsymbol{W}^n(x)) \to \boldsymbol{\Gamma}_x$.
Using (R6) and (H), Liapunov's condition can be easily verified, and hence,

$$\boldsymbol{W}^n(x) - E[\boldsymbol{W}^n(x)] \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_x).$$

Combining these results, an application of Lemma 6.4.1 of Lehmann (1983) yields that, as $n \to \infty$,

$$\boldsymbol{V}^n(x) - \left(\boldsymbol{\Sigma}_x^{-1} - h\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1}\right) E[\boldsymbol{W}^n(x)] \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Gamma}_x\boldsymbol{\Sigma}_x^{-1})$$

where $\boldsymbol{V}^n(x) = (V_{10}^n(x), \ldots, V_{1p_1}^n(x), V_{20}^n(x), \ldots, V_{2p_2}^n(x))^T$.

Before restating Theorem 2.2 more specifically for each estimator $\hat{\beta}_{rj}$, we briefly discuss the asymptotic bias and covariance expressions in the foregoing general result. After some algebra it follows that, for $0 \le s_1 \le p_1$ and $p_1+1 \le s_2 \le p_1+p_2+1$,

$$\begin{aligned}&((\boldsymbol{\Sigma}_x^{-1} - h\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1})E[\boldsymbol{W}^n(x)])_{s_j+1} \\ &\quad = \sqrt{nh}\left(\sum_{r=1}^{2} h^{p_r+1}(c_{1p_r} + hc_{2p_r}) + o(h^{\min(p_1,p_2)+2})\right)\end{aligned}$$

for some constants c_{1p_r}, c_{2p_r}, $r = 1, 2$. These constants are different for each selected component $s_j + 1$ and depend on several quantities including x, the design point under consideration. For ease of notation, all these dependencies have been suppressed. The constants c_{jp_r} are given below for the special case that $\kappa = 2$. For other values of κ similar calculations will lead to the corresponding results.

From the definition of $\boldsymbol{W}_n$ and by using similar argumentation as before, we obtain that

$$\begin{aligned}E[\boldsymbol{W}^n(x)]_{s_j+1} = \sqrt{nh}\Bigg[\sum_{k=1}^{2} f_X(x) I_{jk}(\theta_1(x), \theta_2(x)) \Big\{h^{p_k+1}\, \nu_{s_j'+p_k} \frac{\theta_k^{(p_k+1)}(x)}{(p_k+1)!} + \\ h^{p_k+2}\nu_{s_j'+p_k+1}\xi_{jk}(x)\Big\}\Bigg] + o(nh^{2\min(p_1,p_2)+5})^{1/2},\end{aligned}$$

where $s_1' = s_1$ and $s_2' = s_2 - p_1 - 1$ and

$$\xi_{ij}(x) = \frac{\theta_j^{(p_j+1)}(x)}{(p_j+1)!} \frac{\frac{d}{dx}\{f_X(x) I_{ij}(\theta_1(x), \theta_2(x))\}}{f_X(x) I_{ij}(\theta_1(x), \theta_2(x))} + \frac{\theta_j^{(p_j+2)}(x)}{(p_j+2)!}\,.$$

To simplify the expressions we now introduce some further notation, let

$$\begin{pmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{pmatrix} = f_X(x)\boldsymbol{\Sigma}_x^{-1}\,; \qquad \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} = f_X^2(x)\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1}$$

so that for $i, j = 1, 2$,

$$\boldsymbol{A}_{ij} = \sum_{k=1}^{2}\sum_{\ell=1}^{2} \boldsymbol{\Sigma}^{ik} \frac{d}{dx}(f_X(x) I_{k\ell}(\theta_1(x), \theta_2(x)))\boldsymbol{Q}_{p_k p_\ell} \boldsymbol{\Sigma}^{\ell j}.$$

By calculating the product

$$(\boldsymbol{\Sigma}_x^{-1} - h\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1})E[\boldsymbol{W}^n(x)]$$

we now find that the vector containing the $p_j + 1$ values of c_{1p_r} corresponding to the j-th parameter $\theta_j(x)$ and its first p_j derivatives is given by

$$\frac{\theta_r^{(p_r+1)}(x)}{(p_r+1)!}\left\{\sum_{k=1}^{2} \boldsymbol{\Sigma}^{jk} I_{kr}(\theta_1(x), \theta_2(x))(\nu_{0+p_r+1}, \ldots, \nu_{p_k+p_r+1})^T\right\} .$$

The following vector contains the $p_j + 1$ values of c_{2p_r}:

$$\sum_{k=1}^{2} \boldsymbol{\Sigma}^{jk} I_{kr}(\theta_1(x), \theta_2(x))\xi_{kr}(x)(\nu_{0+p_r+2}, \ldots, \nu_{p_k+p_r+2})^T$$
$$-\frac{\theta_r^{(p_r+1)}(x)}{(p_r+1)!}\left\{\sum_{k=1}^{2} \frac{\boldsymbol{A}_{jk}}{f_X(x)} I_{kr}(\theta_1(x), \theta_2(x))(\nu_{0+p_r+1}, \ldots, \nu_{p_k+p_r+1})^T\right\} .$$

In case x is an interior point of $\mathrm{supp}(f_X)$, some of the constants in (2.16) can be zero: due to the symmetry of the kernel K, all odd moments ν_j are zero. This implies that $[\boldsymbol{N}_{p_k p_k}]_{ij} = 0$ for i even and j odd and for i odd and j even. By using Lemmas 3 and 4 of Fan, Heckman and Wand (1995), a similar property holds for $\boldsymbol{\Sigma}_x^{-1}$. If $p_r - s_r'$ is even, then it follows that $c_{1p_r} = 0$. In addition, if p_1 and p_2 are both odd or both even, then $c_{1p_1} = c_{1p_2} = 0$.

The main conclusion concerning the asymptotic bias expression (2.16) is that the leading term is essentially of an order determined by the polynomial of the smallest degree. Hence, we recommend using polynomials of the same degree $p_1 = p_2 = p$, preferably odd. Indeed, for odd degree polynomial fits, boundary bias and interior bias are of the same order of magnitude. This automatic incorporation of boundary treatment is one of several nice properties to which local polynomial fitting owes its popularity (see, e.g., Fan and Gijbels, 1996).

A more explicit formula for the asymptotic covariance can be derived in case $p_1 = p_2 = p$, namely, $\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Gamma}_x\boldsymbol{\Sigma}_x^{-1}$ equals

$$\{f_X(x)\boldsymbol{I}(\theta_1(x), \theta_2(x))\}^{-1} \otimes \boldsymbol{N}_{pp}(x)^{-1}\boldsymbol{T}_{pp}(x)\boldsymbol{N}_{pp}(x)^{-1}.$$

The marginal asymptotic normality of $\hat{\beta}_{rj}$ when $p_1 = p_2 = p$ can now be restated explicitly as follows.

Corollary 2.1 *Under the conditions of Theorem 2.2, we have for x an interior point of $supp(f_X)$: if $p-j$ odd $(j = 0, \ldots, p)$ and $r = 1, 2$, then for $n \to \infty$*

$$\sqrt{nh^{2j+1}} \left(f_X^{-1}(x) \left(\boldsymbol{I}^{-1}(\theta_1(x), \theta_2(x)) \right)_{r,r} \int_{\mathcal{R}_x} K_{jp}^2(z) dz \right)^{-1/2} \times$$

$$\left[\hat{\beta}_{rj} - \theta_{rj}(x) - h^{p-j+1} \frac{\theta_r^{(p+1)}(x)}{(p+1)!} \int_{\mathcal{R}_x} z^{p+1} K_{jp}(z) dz (1 + O(h)) \right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1); \quad (2.16)$$

if $p-j$ even $(j = 0, \ldots, p)$ and $r = 1, 2$, then for $n \to \infty$

$$\sqrt{nh^{2j+1}} \left(f_X^{-1}(x) \left(\boldsymbol{I}^{-1}(\theta_1(x), \theta_2(x)) \right)_{r,r} \int_{\mathcal{R}_x} K_{jp}^2(z) dz \right)^{-1/2} \times$$

$$\left[\hat{\beta}_{rj} - \theta_{rj}(x) - \varrho h^{p-j+2} (1 + O(h)) \right] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $(\boldsymbol{A})_{r,r}$ is the (r,r)th entry of the matrix $\boldsymbol{A}$ and the explicit form of the constant ϱ depends in a rather complicated way on $\theta_r^{(p+1)}(x)$, $\theta_r^{(p+2)}(x)$, $\boldsymbol{I}(\theta_1(x), \theta_2(x))$, $\int_{\mathcal{R}_x} z^{p+2} K_{jp}(z) dz$, $\int_{\mathcal{R}_x} z^{p+1} K_{j-1,p}(z) dz$ and $f_X(x)$, see equation (2.17).
For x a boundary point, the result (2.16) always holds, regardless whether $p-j$ is odd or even.

Proof. First, note that for $p_1 = p_2 = p$, expressions (2.12),(2.13) and (2.14) simplify to

$$\boldsymbol{\Sigma}_x = f_X(x) \boldsymbol{I}(\theta_1(x), \theta_2(x)) \otimes \boldsymbol{N}_{pp}(x), \quad \boldsymbol{\Gamma}_x = f_X(x) \boldsymbol{I}(\theta_1(x), \theta_2(x)) \otimes \boldsymbol{T}_{pp}(x)$$

and

$$\boldsymbol{\Lambda}_x = \begin{pmatrix} \frac{d}{dx}(f_X(x) I_{11}(\theta_1(x), \theta_2(x))) & \frac{d}{dx}(f_X(x) I_{12}(\theta_1(x), \theta_2(x))) \\ \frac{d}{dx}(f_X(x) I_{21}(\theta_1(x), \theta_2(x))) & \frac{d}{dx}(f_X(x) I_{22}(\theta_1(x), \theta_2(x))) \end{pmatrix} \otimes \boldsymbol{Q}_{pp}(x).$$

with $\otimes$ the ordinary Kronecker product. The result then follows from some well-known properties of the Kronecker product and using Lemma 3 and Lemma 4 of Fan, Heckman and Wand (1995).

In order to give the exact definition of ϱ in the asymptotic bias expression for the case $p-j$ even, we define $\boldsymbol{B}$ as the matrix

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \boldsymbol{I}^{-1}(\theta_1(x),\theta_2(x)) \times \begin{pmatrix} \frac{d}{dx}(f_X(x)I_{11}(\theta_1(x),\theta_2(x))) & \frac{d}{dx}(f_X(x)I_{12}(\theta_1(x),\theta_2(x))) \\ \frac{d}{dx}(f_X(x)I_{21}(\theta_1(x),\theta_2(x))) & \frac{d}{dx}(f_X(x)I_{22}(\theta_1(x),\theta_2(x))) \end{pmatrix} \times \boldsymbol{I}^{-1}(\theta_1(x),\theta_2(x)),$$

and

$$\delta_r(x) = \sum_{k=1}^{2}\sum_{\ell=1}^{2}[\boldsymbol{I}^{-1}(\theta_1(x),\theta_2(x))]_{r\ell}I_{lk}(\theta_1(x),\theta_2(x))\xi_{\ell k}(x).$$

The definition of ϱ is now given by,

$$\varrho = \delta_r(x)\int_{\mathcal{R}_x} z^{p+2}K_{jp}(z)dz - \frac{1}{f_X(x)}\sum_{k=1}^{2}\sum_{\ell=1}^{2}\frac{\theta_k^{(p+1)}(x)}{(p+1)!}B_{r\ell}I_{\ell k}\int_{\mathcal{R}_x} z^{p+1}K_{j-1,p}(z)dz. \tag{2.17}$$

Remark 2.1. Consider independent and normally distributed responses $Y_i \sim \mathcal{N}(\mu(x_i),\ \sigma^2(x_i))$, $i = 1,\ldots,n$. The proposed local likelihood method admits the simultaneous estimation of mean and variance function. Plotting both smoothed curves on a scatter plot can suggest or confirm a specific heteroscedasticity pattern. Traditionally $\mu(x)$ and $\sigma^2(x)$ are estimated separately, see e.g. Ruppert, Wand, Holst and Hössjer (1997). A specification of Corollary 2.1 shows that both methods are asymptotically equivalent. Indeed, for normally distributed responses and $\theta_1(x) = \mu(x)$, $\theta_2(x) = \sigma^2(x)$, it is easy to see that

$$\boldsymbol{I}(\mu(x),\sigma^2(x)) = \begin{pmatrix} \frac{1}{\sigma^2(x)} & 0 \\ 0 & \frac{1}{2\sigma^4(x)} \end{pmatrix}$$

such that in this case $\left(\boldsymbol{I}^{-1}(\theta_1(x),\theta_2(x))\right)_{1,1} = \sigma^2(x)$ and $\left(\boldsymbol{I}^{-1}(\theta_1(x),\theta_2(x))\right)_{2,2} = 2\sigma^4(x)$. This yields exactly the same asymptotic normality results as in Ruppert and Wand (1994) for the estimation of $\mu(x)$ and as in Ruppert, Wand, Holst and Hössjer (1997) for the estimation of $\sigma^2(x)$.

Remark 2.2. Consider the one-parameter case. For $p = 0$ (local constant), Staniswalis (1989) states an asymptotic normality result similar to Corollary 2.1. For a response Y having an exponential family pdf $f(y; \theta(x))$ (generalized linear models), Corollary 2.1 reduces to Theorem 1 of Fan, Heckman and Wand (1995).

We conclude this section with a second corollary. The estimation of $\theta_1(x)$ and $\theta_2(x)$ (corresponding to $j = 0$ in Corollary 2.1) is of primary interest. Let $\hat{\theta}_1(x) = \hat{\beta}_{10}$ and $\hat{\theta}_2(x) = \hat{\beta}_{20}$ denote their respective estimators. Although the role of link functions here is less crucial than in parametric models (because the fitting is localized), link functions still can be very useful in this nonparametric setting. First of all, a properly defined link function guarantees that the final estimators $\hat{\theta}_1(x)$ and $\hat{\theta}_2(x)$ will have the correct range of admissible values. Moreover, the choice of the link function affects the computational aspects of the estimation procedure. Let $\theta_r(x) = g_r(\xi_r(x))$ $(r = 1, 2)$ where ξ_r is the real parameter of interest and $g_r : \mathbb{R} \to \mathbb{R}$ some link function. A proper estimator for $\xi_r(x)$ is given by $\hat{\xi}_r(x) = g_r^{-1}(\hat{\theta}_r(x))$. Corollary 2.1 and the δ-method immediately imply the following result.

Corollary 2.2 *Assume the conditions of Theorem 2.2. If the link function g_r is differentiable and g_r' is continuous at $\xi_r(x)$ with $g_r'(\xi_r(x)) \neq 0$ $(r = 1, 2)$, then $\hat{\xi}_r(x)$ has the same limiting distribution as $\hat{\theta}_r(x)$ but with asymptotic bias and variance divided by $g_r'(\xi_r(x))$ and $g_r'(\xi_r(x))^2$ respectively.*

2.4 Extensions

Next to parameters which describe the success probability and the within cluster correlation, most models for clustered binary data e.g. the beta-binomial model (see Section 2.6.1) also depend on n_i, the number of correlated Bernoulli trials. We refer to this parameter as the "cluster size". Because not every cluster has the same size, Theorems 2.1 and 2.2 and their corollaries are not directly applicable. The complete sample must be split into several (independent) subsamples according to different values of n_i. Lehmann (1983, section 6.6) showed how consistency and normality results for each subsample (with a fixed known value of n_i) can be combined and carried over to the full sample.

We now formulate this extension to two or more samples in general. Suppose that we have a independent samples $(x_{\alpha 1}, Y_{\alpha 1}), \ldots, (x_{\alpha N_\alpha}, Y_{\alpha N_\alpha})$, $\alpha = 1, \ldots, a$. Denote

$f_\alpha(y;\theta_1(x_{\alpha i}),\theta_2(x_{\alpha i}))$ the pdf of $Y_{\alpha i}$. Using the information in all a samples, we want to estimate $\theta_1(x)$ and $\theta_2(x)$ for a fixed design point x in $\mathrm{supp}(f_X)$. The total weighted log-likelihood, combining the contribution of each separate sample, is given by, with total sample size $N=\sum_{\alpha=1}^{a} N_\alpha$,

$$\mathcal{L}_N(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=\frac{1}{N}\sum_{\alpha=1}^{a}\sum_{i=1}^{N_\alpha}\log f_\alpha(Y_{\alpha i};\sum_{j=0}^{p_1}\beta_{1j}(x_{\alpha i}-x)^j,\sum_{j=0}^{p_2}\beta_{2j}(x_{\alpha i}-x)^j)K_h(x_{\alpha i}-x) \tag{2.18}$$

and the maximizer $(\hat{\boldsymbol{\beta}}_1^T,\hat{\boldsymbol{\beta}}_2^T)^T$ of $\mathcal{L}_N(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ is the local polynomial MLE for $(\boldsymbol{\theta}_1^T(x),\boldsymbol{\theta}_2^T(x))^T$.

To discuss the asymptotic properties of $(\hat{\boldsymbol{\beta}}_1^T,\hat{\boldsymbol{\beta}}_2^T)$ we consider, as in Lehmann (1983), sequences of sample sizes N_α which tend to infinity all at the same rate. The number of samples a remains fixed. More precisely, for a total sample size N tending to infinity,

$$\frac{N_\alpha}{N}\to\lambda_\alpha,\quad \alpha=1,\ldots,a \tag{2.19}$$

where $\sum_{\alpha=1}^{a}\lambda_\alpha=1$ with all $\lambda_\alpha>0$. Let $\boldsymbol{I}^{(\alpha)}(\theta_1(x),\theta_2(x))$ denote the information matrix corresponding to $f_\alpha(y;\theta_1(x),\theta_2(x))$ and let $\boldsymbol{\Sigma}_x^{(\alpha)},\boldsymbol{\Gamma}_x^{(\alpha)},\boldsymbol{\Lambda}_x^{(\alpha)}$ be defined as in (2.12)-(2.14) with $\boldsymbol{I}(\theta_1(x),\theta_2(x))$ replaced by $\boldsymbol{I}^{(\alpha)}$ $(\theta_1(x),\theta_2(x))$. By the same arguments as those used in the proof of Theorem 6.6.1 in Lehmann (1983), we get the following.

Theorem 2.3 *Assume that the conditions of Theorem 2 hold for each of the pdf $f_\alpha(y;\theta_1,\theta_2)$ and $N_\alpha\to\infty$ $(\alpha=1,\ldots,a)$. If N tends to infinity according to (2.19), then the results of Theorem 2.1 and 2.2 for the consistency and asymptotic normality of $(\hat{\boldsymbol{\beta}}_1^T,\hat{\boldsymbol{\beta}}_2^T)$ remain valid with*

$$\boldsymbol{\Sigma}_x=\sum_{\alpha=1}^{a}\lambda_\alpha\boldsymbol{\Sigma}_x^{(\alpha)},\quad \boldsymbol{\Gamma}_x=\sum_{\alpha=1}^{a}\lambda_\alpha\boldsymbol{\Gamma}_x^{(\alpha)},\quad \boldsymbol{\Lambda}_x=\sum_{\alpha=1}^{a}\lambda_\alpha\boldsymbol{\Lambda}_x^{(\alpha)}.$$

In case $(X_1,Y_1),\ldots,(X_n,Y_n)$ is a random sample from (X,Y) with $f(y;\theta_1(x),\theta_2(x))$ the conditional density of Y given $X=x$, it can easily be verified that, under the same set of conditions (the Lipschitz condition in (R3) and (K) can be omitted), all results remain valid with the same asymptotic bias and variance expressions and with $f_X(x)$ interpreted as the marginal density of the covariate X.

2.5 Bandwidth choice

The choice of the bandwidth parameter h is important in nonparametric curve estimation in general. When experimenting with different bandwidth choices and exploring the resulting fitted curves by eye, one often will be able to select a possibly good value for h. But it is clear that finding an "optimal bandwidth" is not an easy task. A typical starting point is to balance the (asymptotic) bias and variance of the proposed estimators. Also in our multiparameter case, this is a possible approach. In terms of the estimators $\hat{\theta}_1(x)$ and $\hat{\theta}_2(x)$, it follows from Corollary 2.1 that for, say, $p_1 = p_2 = p$ odd,

$$\begin{aligned}\mathrm{AMSE}(\hat{\theta}_1(x), \hat{\theta}_2(x)) &= \lim_{n\to\infty} E||(\hat{\theta}_1(x), \hat{\theta}_2(x)) - (\theta_1(x), \theta_2(x))||^2 \\ &= \frac{tr\left(\boldsymbol{I}^{-1}(\theta_1(x), \theta_2(x))\right)}{nhf_X(x)} \int_{\mathcal{R}_x} K_{0p}^2(z)dz + h^{2p+2}\Big(\int_{\mathcal{R}_x} z^{p+1} K_{0p}(z)dz\Big)^2 \sum_{\ell=1}^{2} \left(\frac{\theta_\ell^{(p+1)}(x)}{(p+1)!}\right)^2.\end{aligned}$$

Minimizing $\mathrm{AMSE}(\hat{\theta}_1(x), \hat{\theta}_2(x))$ (asymptotic mean squared error) as a function of h balances bias squared and variance of both estimators, leading to

$$h_{1,opt}(x) = C(K, f_X) C(\theta_1(x), \theta_2(x)) n^{-1/(2p+3)}$$

where the constants $C(K, f_X)$ and $C(\theta_1(x), \theta_2(x))$ are given by respectively

$$\left(\frac{(p+1)!p! \int_{\mathcal{R}_x} K_{0p}^2(z)dz}{f_X(x)\Big(\int_{\mathcal{R}_x} z^{p+1} K_{0p}(z)dz\Big)^2}\right)^{1/(2p+3)} \quad \text{and} \quad \left(\frac{tr(\boldsymbol{I}^{-1}(\theta_1(x), \theta_2(x)))}{2[(\theta_1^{(p+1)}(x))^2 + (\theta_2^{(p+1)}(x))^2]}\right)^{1/(2p+3)}.$$

As in locally weighted least squares regression $h_{1,opt}(x)$ is of the order $O(n^{-1/(2p+3)})$. The first constant $C(K, f_X)$ only depends on kernel K and design density $f_X(x)$ but the second contains several unknown quantities.

Another more natural approach in likelihood estimation is to consider the bandwidth which maximizes $E[\log f(Y; \hat{\theta}_1(x), \hat{\theta}_2(x))]$. Using assumptions (R1)-(R3), this expected loglikelihood can be approximated by

$$\sum_{r,s=1}^{2} I_{rs}(\theta_1(x), \theta_2(x)) \left[Cov(\hat{\theta}_r(x), \hat{\theta}_s(x)) + (E[\hat{\theta}_r(x)] - \theta_r(x))(E[\hat{\theta}_s(x)] - \theta_s(x))\right].$$

Turning to the asymptotic bias and covariance as in the AMSE-criterion and after some rewriting, we get

$$h_{2,opt}(x) = C(K, f_X) \tilde{C}(\theta_1(x), \theta_2(x)) n^{-1/(2p+3)} \tag{2.20}$$

with $C(K, f_X)$ as before but with a different unknown constant

$$\tilde{C}(\theta_1(x), \theta_2(x)) \;=\; \left((\boldsymbol{\theta}^{(p+1)}(x))^T \boldsymbol{I}(\theta_1(x), \theta_2(x)) \boldsymbol{\theta}^{(p+1)}(x) \right)^{-1/(2p+3)}$$

where $\boldsymbol{\theta}^{(p+1)}(x) = (\theta_1^{(p+1)}(x)^T \; ; \; \theta_2^{(p+1)}(x)^T)^T$. Equation (2.20) yields an optimal bandwidth of the same order $O(n^{-1/(2p+3)})$, but different weights are put on the components of the parameter vector and it also incorporates the (asymptotic) covariance between both estimators $\hat{\theta}_1(x)$ and $\hat{\theta}_2(x)$. Both optimal bandwidths, $h_{1,opt}(x)$ and $h_{2,opt}(x)$, are local bandwidths. Taking a global loss measure such as the asymptotic mean integrated squared error, we get an optimal global fixed bandwidth.

Plug-in procedures could be implemented to get a data-driven bandwidth, but this approach gets very complicated. An alternative approach is to define

$$\hat{h}_{CV} \;=\; \text{argmax}_{h>0} \sum_{i=1}^{n} \log f(Y_i; \hat{\theta}_1^{[i]}(x_i), \hat{\theta}_2^{[i]}(x_i)) \tag{2.21}$$

where $\hat{\theta}_1^{[i]}(x_i)$ and $\hat{\theta}_2^{[i]}(x_i)$ are the estimators based on the sample without the ith observation (x_i, Y_i). This cross-validation method is straightforward and fully data-driven but generally it is not recommended, even in much simpler settings, for its large sample variation.

For one-parameter (classical) regression models, several data-driven bandwidth selectors have been studied; extensions of these methods to the likelihood models that we consider is not straightforward. Since for our applications the particular choice of the bandwidth is not critical, in the data examples, bandwidth choice is based on cross-validation as described in (2.21).

2.6 Data examples and some simulations

In this section we first will give some details on the parametric likelihood implementation of a typical probability model for clustered binary data, the beta-binomial model, next we apply the local polynomial estimators to some data sets introduced in Chapter 1 and we present the results of a small simulation study.

2.6.1 The beta-binomial likelihood

Turning to our main application, we selected the beta-binomial model as a well-known example of a probability model for these type of data, (some other models

appear in Section 3.2.2). Consider an experiment involving N litters (pregnant dams), the ith of which contains n_i foetuses. Suppose Y_{ij} indicates whether the jth foetus in cluster i is abnormal. Then the total number of malformed foetuses in cluster i is $Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij}$. The covariate of interest is d_i. This is defined to be the dosing that was given to cluster i (or any other cluster-specific covariate of interest). Further we assume an exchangeable model, in the sense that each foetus within a litter has the same malformation probability, and the association between any pair of foetuses within a litter is equal.

The beta-binomial approach assumes a random malformation probability P_i in cluster i to come from a beta distribution with mean π_i. Given P_i, the outcomes within the ith cluster follow a binomial distribution (see, e.g., Kleinman, 1973). Let $\rho_i = \text{Corr}(Y_{ij}, Y_{ik})$ (using exchangeability) denote the intra-litter correlation. Then, this leads to the following pdf

$$f(y_{i\bullet}; \pi_i, \rho_i) = \binom{n_i}{y_{i\bullet}} \frac{B(\pi_i(\rho_i^{-1} - 1) + y_{i\bullet}, (1 - \pi_i)(\rho_i^{-1} - 1) + (n_i - y_{i\bullet}))}{B(\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1))} \tag{2.22}$$

where $B(.,.)$ denotes the beta function. It can easily be shown that the contribution of the ith cluster ($i = 1, \ldots, N$) to the log-likelihood $\ln f(y_{i\bullet}; \pi_i, \rho_i)$ is given by

$$\ln \binom{n_i}{y_{i\bullet}} + \sum_{r=0}^{y_{i\bullet}-1} \ln\left(\pi_i + \frac{r\rho_i}{1-\rho_i}\right) + \sum_{r=0}^{n_i - y_{i\bullet}-1} \ln\left(1 - \pi_i + \frac{r\rho_i}{1-\rho_i}\right) - \sum_{r=0}^{n_i-1} \ln\left(1 + \frac{r\rho_i}{1-\rho_i}\right). \tag{2.23}$$

Note that this expression reduces to the familiar binomial log-likelihood when putting $\rho_i = 0$.

Parametrically, we can model the marginal parameters π_i and ρ_i as a function of a covariate matrix X_i and a parameter vector $\boldsymbol{\beta}$. Since Y_{ij} is binary, the logistic link function for π_i is a natural choice. A convenient transformation of ρ_i is Fisher's z-transform. This leads to the following model

$$\begin{pmatrix} \ln(\frac{\pi_i}{1-\pi_i}) \\ \ln(\frac{1+\rho_i}{1-\rho_i}) \end{pmatrix} = \boldsymbol{X}_i\boldsymbol{\beta} \tag{2.24}$$

with $\boldsymbol{X}_i$ a design matrix based on dose d_i associated with cluster i and $\boldsymbol{\beta}$ the coefficient-vector.

Local polynomial estimates are obtained by defining $\theta_1(x) = \text{logit}\{\pi(x)\}$, $\theta_2(x) = \ln\{(1 + \rho(x))/(1 - \rho(x))\}$, and by using the beta-binomial log-likelihood (2.23) (in

terms of $\theta_1(x)$ and $\theta_2(x)$) in the definition of the kernel weighted log-likelihood function (2.2). Since for the beta-binomial model explicit expressions for the maximizer of the function (2.2) are not available, a Newton-Raphson algorithm is used to obtain the estimates in the data examples below.

2.6.2 The low-iron rat teratology data

As a first illustrative example we consider data from a laboratory study on rats investigating the effects of dietary regiments on fetal development (Shepard, Mackler and Finch, 1980). Moore and Tsiatis (1991) analyzed these data using the hemoglobin levels of the mother rats as a covariate. Here the local polynomial MLE will be applied.

For a grid of hemoglobin levels and with $p_1 = p_2 = 1$ (i.e. local linear), we estimated the proportion dead $\pi(x)$ and the correlation $\rho(x)$. We chose the logit and Fisher z-transform links. After rescaling the hemoglobin levels to the unit interval, the cross-validation (CV) procedure (2.21) resulted in a bandwidth $\hat{h}_{CV} = 0.319$ (see Figure 2.1a for a plot of the CV-measure as a function of $\log(h)$). Figure 2.1b shows the data together with the fitted $\pi(x)$-curve; the same curve on logit scale is shown in Figure 2.1c. This last picture suggests a possible quadratic type of curvature. Figure 2.1d shows the fitted $\rho(x)$-curve. Apparently intra-litter correlation decreases linearly as hemoglobin level increases. Inspired by these suggestions, we fitted the data in a parametric way using in (2.24) the design matrix

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & x_i & x_i^2 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_i \end{pmatrix}.$$

The quadratic effect of hemoglobin on the proportion dead and the linear effect on the correlation appeared to be statistically significant (P-value of likelihood ratio test: 0.029). This illustrates a typical application of a nonparametric estimator as a tool to suggest specific functional relationships.

2.6.3 The twins data

The same estimation procedure can be applied to the twins data. We here show local linear/linear ($p_1 = p_2 = 1$) and local linear/constant ($p_1 = 1; p_2 = 0$) estimates

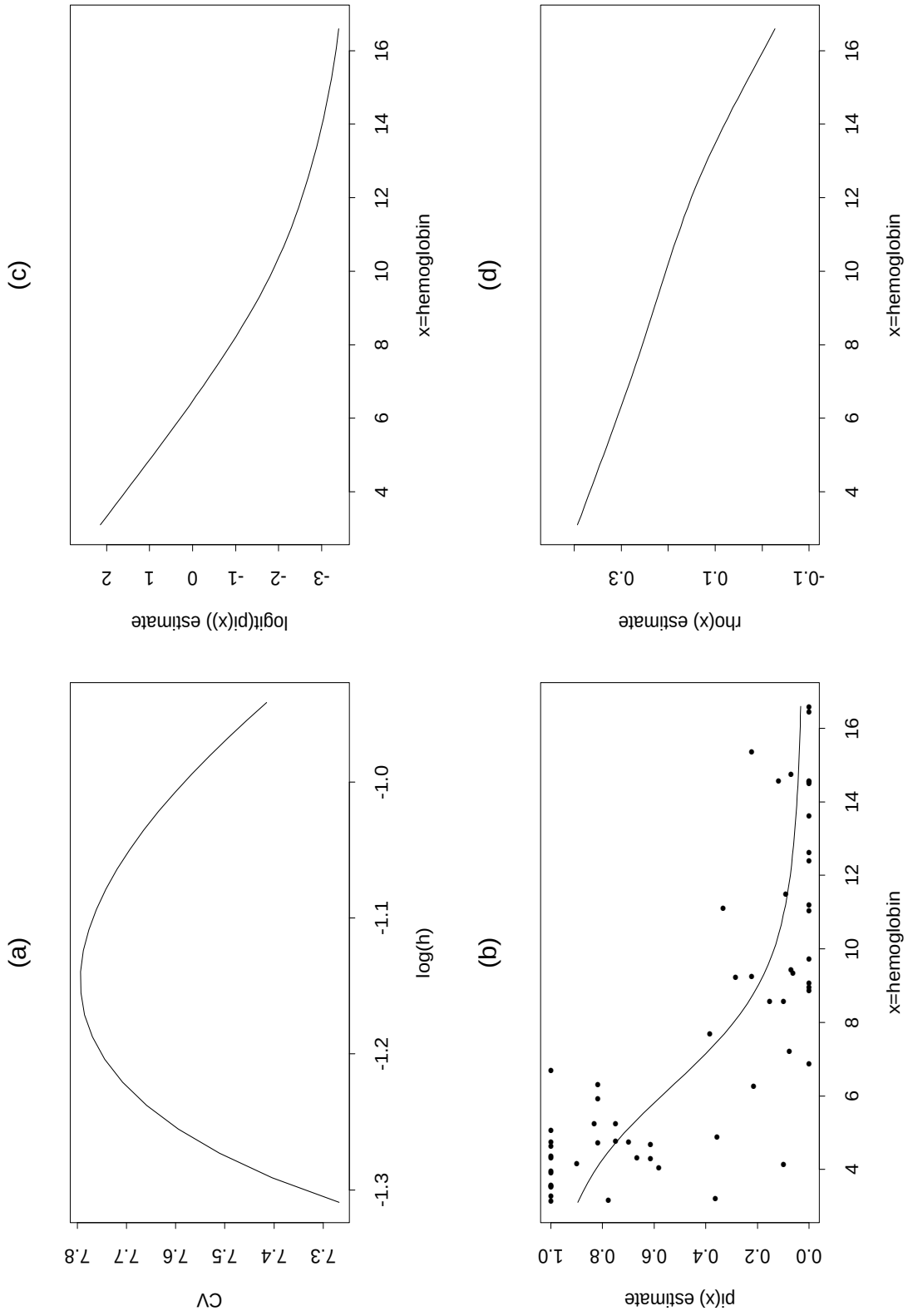


Figure 2.1: Local linear/linear beta-binomial estimates for the low-iron rat teratology data.

of the proportion of mortality and morbidity within 28 days after birth and of the within-twin correlation.

The estimates were obtained using the same bandwidth $h = 0.2$ for all curves. The local linear ($p_1 = 1$) estimated proportions nearly coincide. As expected, differences will occur at the boundaries, if we compare local constant ($p_2 = 0$) and local linear ($p_2 = 1$) fits. From the graphs resulting from both estimation procedures it is observed that there indeed is some correlation (Figure 2.2d) and that the probability of mortality and morbidity decreases for increasing gestational age (Figure 2.2b).

2.6.4 The Wisconsin diabetes study

For the study on younger onset diabetes patients, we get the following estimates for the probability of macular edema and for the correlation between the outcomes of the same person (Figure 2.3). Most often the correlation structure is assumed to be constant. In Section 6.6 we will test this assumption using an omnibus lack of fit test. Figure 2.3 already indicates that this assumption might be violated here. The estimates shown are the local linear/linear ($p_1 = p_2 = 1$) and local quadratic/linear ($p_1 = 2; p_2 = 1$) estimates, based on the beta-binomial likelihood model, using bandwidth h=0.3 (on zero-one scale). For these data the estimated curves are in close agreement.

2.6.5 The study of herbicides on mice

For the data arising from this study we fitted local linear/linear estimates of the proportion of death foetuses per litter and of the within-litter correlation. The estimated curves are shown in Figure 2.4. Note that although there are only seven dose levels, smoothing techniques can be applied, see Staniswalis and Cooper (1988) where kernel estimators are used in a similar setting. Figure 2.4 shows an approximate linear estimated curve for the Fisher's z-transform of the within-cluster correlation. For this data the correlation between the littermates increases rapidly with the dose level, and reaches an estimated value of approximately 0.72 for the highest dose level of 60 mg/kg/day.

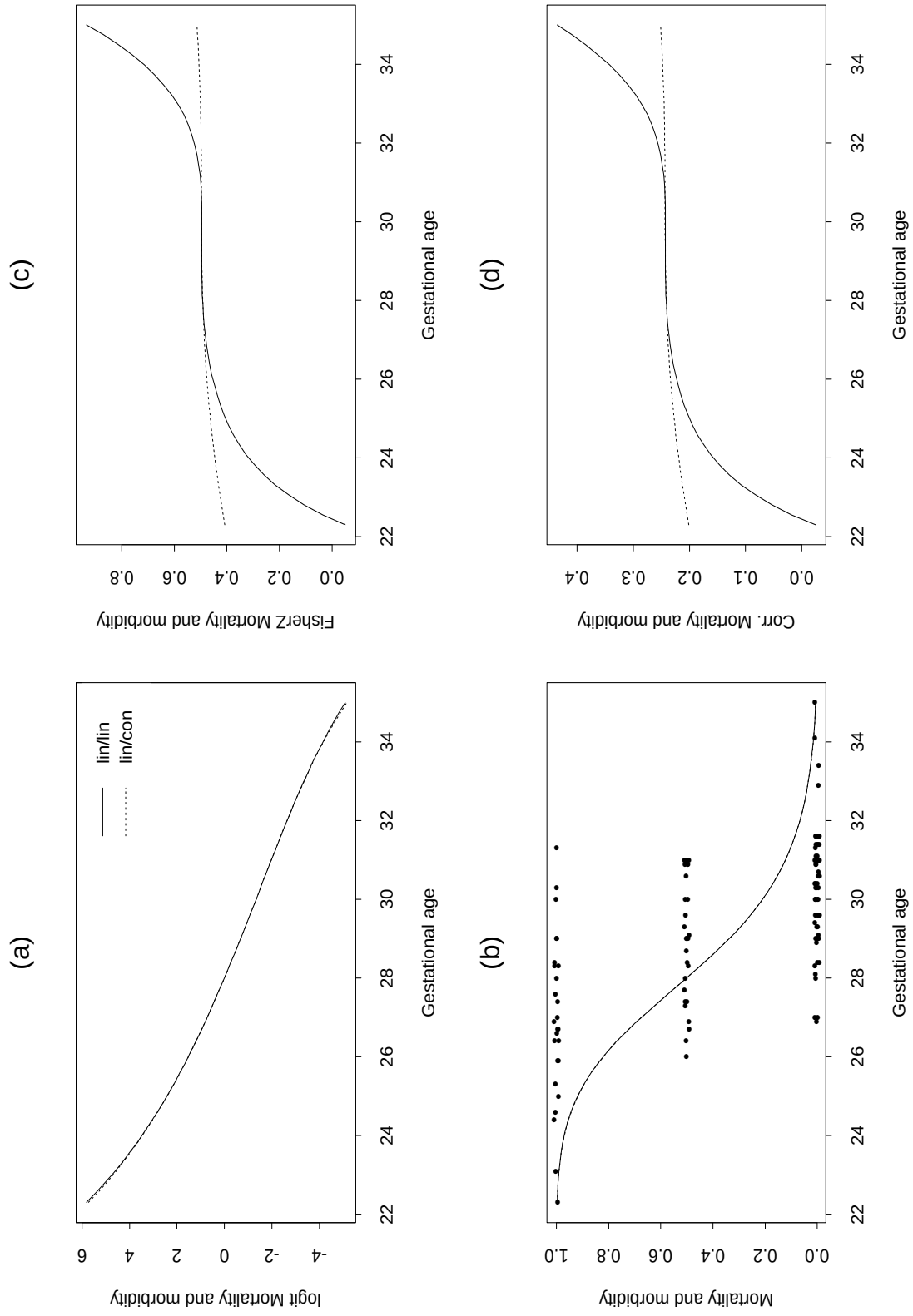


Figure 2.2: Local linear/linear (solid line) and local linear/constant (dashed line) beta-binomial estimates for the twins data.

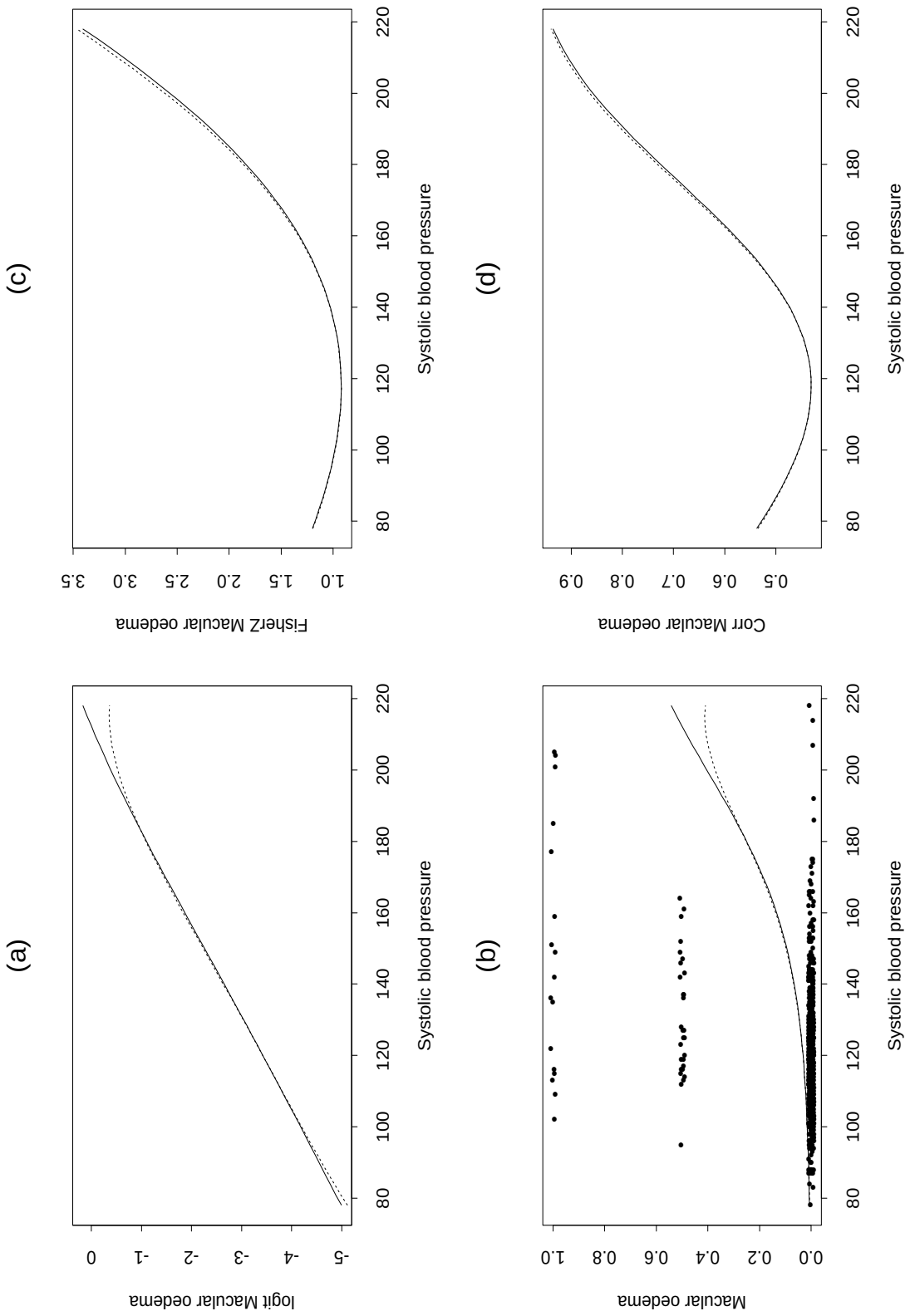


Figure 2.3: Local linear/linear (solid line) and local quadratic/linear (dashed line) beta-binomial estimates for the Wisconsin diabetes data.

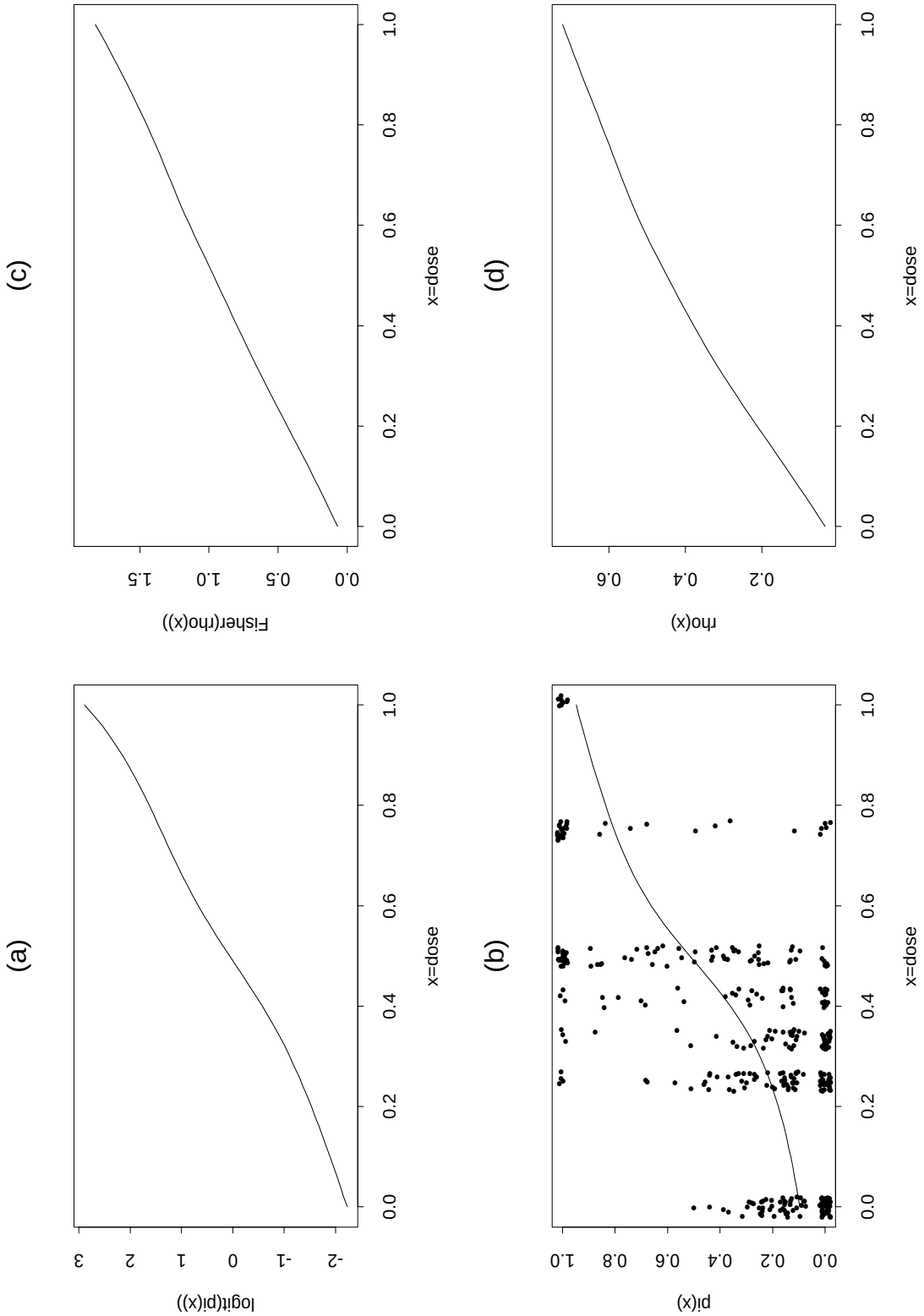


Figure 2.4: Local linear/linear beta-binomial estimates for the study of herbicides on mice ($h = 0.2$).

2.6.6 Simulation of a toxicological experiment

To show the finite sample behavior of the local polynomial MLE in the setting of a developmental toxicity study, we performed a limited simulation study. The design was generated according to definition (2.1). We included 40 dose levels, so $d_i = G^{-1}((i-1)/39)$, $i = 1, \ldots, 40$ where G is a slightly modified beta $B(0.5, 1)$ distribution function. More precisely, we chose $f_D(d) = \left(2(\sqrt{1+\epsilon} - \sqrt{\epsilon})\sqrt{d+\epsilon}\right)^{-1}$ such that, for $i = 1, \ldots, 40$,

$$d_i = \left(\frac{i-1}{39}(\sqrt{1+\epsilon} - \sqrt{\epsilon}) + \sqrt{\epsilon}\right)^2 - \epsilon.$$

Any $\epsilon > 0$ results in a design density satisfying assumption (G) ($\epsilon = 0$ coincides with the $B(0.5, 1)$ distribution). We took $\epsilon = 10^{-10}$. This design consists of one control group ($d_1 = 0$) and 39 active groups. This skewed beta-type design with high concentration near the zero dose has been selected because toxicologist have special interest in the estimates at low doses. An equal number of 5 clusters is assigned to each dose level. The number n_i of foetuses per cluster is assumed to follow a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper et al (1986), which is considered representative of that encountered in actual experiments. The smoothed frequencies can be found in Molenberghs, Declerck and Aerts (1998) where least squares cross-validation has been used to choose the bandwidth. Here, the same set of n_i's is used in each simulation run.

Using the built-in GAUSS routine RNDU, data are generated from the beta-binomial probability model with parameters n_i , $\pi_i = \pi(d_i)$ and $\rho_i = \rho(d_i)$, the latter two defined as

$$\begin{aligned} \ln\left(\frac{\pi(d)}{1-\pi(d)}\right) &= -2.8 + 5.2(d+0.05)^{0.4}\ln(d+1), \\ \ln\left(\frac{1+\rho(d)}{1-\rho(d)}\right) &= 0.5 + 0.4d + 0.2(0.3-d)^2. \end{aligned}$$

Because of the computational complexity, we used for all data sets the same bandwidth parameter $h = 0.19$, a choice based on a limited cross-validation search on a few samples. We estimated the parameters $\pi(d)$ and $\rho(d)$ in a grid of equidistant points $d_k^\star = k/40$, $k = 0, \ldots, 40$, starting in the midpoint $d_{20}^\star = 0.5$ and proceeding alternately towards the left and right boundary. Starting values for $\hat{\beta}_{10}$, $\hat{\beta}_{11}$, $\hat{\beta}_{20}$ and $\hat{\beta}_{21}$ in the Newton-Raphson procedure were only needed in the midpoint; for

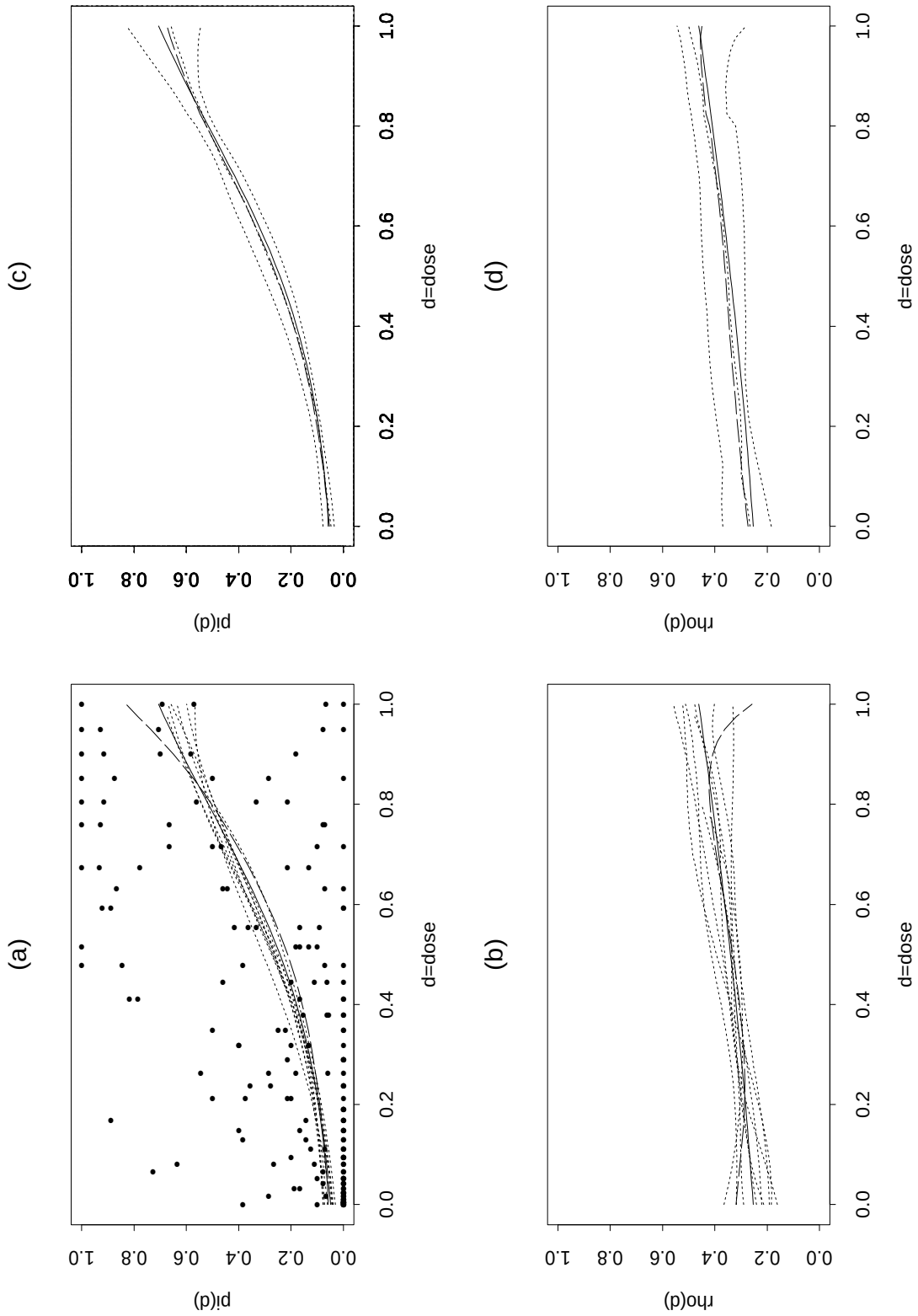


Figure 2.5: Results of a simulation study

the other points the previously estimated betas were automatically used as starting points.

The results of 500 simulation runs are shown in Figure 2.5 containing the true curve (solid line), a curve connecting the means (of the 500 estimates) at the points $d_k^\star$ and three other curves connecting the 5 %, 50 % and 95 % percentiles of the estimates at the $d_k^\star$'s, for $\pi(d)$ (Figure 2.5c) and $\rho(d)$ (Figure 2.5d). To give an idea of individual realizations of $\hat{\pi}(d)$ and $\hat{\rho}(d)$, Figure 2.5a and 2.5b show for 10 arbitrary chosen samples the corresponding fitted curves together with data of one of these 10 samples. The true curves are shown in the solid line type.

We also computed the local linear estimator $\tilde{\pi}(d)$ for $\pi(d)$ which ignores the correlation structure ($\rho(d) = 0$). It is well-known that, regardless the value of $\rho(d)$, $\tilde{\pi}(d)$ is a consistent estimator for $\pi(d)$. Some additional simulations showed that in most cases the bias of $\tilde{\pi}(d)$ is very comparable with that of the beta-binomial estimator $\hat{\pi}(d)$ and, as expected, the phenomenon of "overdispersion" or "extra binomial variation" is also observable in this nonparametric setting. The actual variance of $\tilde{\pi}(d)$ is substantially larger than explained by its nominal value. Finally we also observed that, in case of independent binomial trials ($\rho(d) = 0$), both estimators $\hat{\pi}(d)$ and $\tilde{\pi}(d)$ performed very similar.

2.7 Discussion

Our illustrative examples in this chapter all contain cluster correlated binary outcomes. Of course, this is just one of many domains where the method is applicable. For example, Davison and Ramesh (1998) used the estimators of Chapter 2 to study trends in sample extremes. One of their examples is based on the extreme value distribution. For $X_1, \ldots, X_m$ a set of independent identically distributed variables from any of a wide class of distributions, the distribution of the maximum $Y = \max(X_1, \ldots, X_m)$ (for large m) is approximately

$$H(y; \eta, \tau, \kappa) = \exp\left\{-\left(1 + \kappa\frac{y-\eta}{\tau}\right)^{-1/\kappa}\right\}, \quad -\infty < \kappa,\ \eta < \infty,\ \tau > 0, \qquad (2.25)$$

where the range of the response y is such that $1 + \kappa(y-\eta)/\tau > 0$, and η, τ and κ respectively control the distribution's location, scale and shape. If the data $y_1, \ldots, y_n$ are the maxima for successive time points (e.g. years), one might assume that location, scale and shape vary according to smooth functions $\eta(t), \tau(t)$ and $\kappa(t)$. Local

polynomial estimators of these functions are now directly obtained by maximizing the kernel weighted 3-parameter log likelihood function, using distribution (2.25).

Although the methods in this chapter already cover quite some statistical models, there is still room for further extensions. Sometimes, one is not willing to specify the full likelihood of the data, since this would imply putting too stringent restrictions on the model. In parametric modeling, several interesting alternatives exist, pseudolikelihood (Arnold and Straus, 1991, Besag, 1975) or generalized estimating equations (GEE, Liang and Zeger, 1986, Zeger and Liang, 1986), to mention just two of them.

In Chapter 3 we will take a closer look at the local pseudolikelihood estimating equations under model misspecification.

Carroll, Ruppert and Welsh (1998) considered the general estimating equations setting. For independent observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ they define the local polynomial estimators as the solution to the following local estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}\left(\boldsymbol{Y}_i; \sum_{j=0}^{p} \beta_{1j}(X_i - x)^j, \ldots, \sum_{j=0}^{p} \beta_{\kappa j}(X_i - x)^j\right)(X_i - x)^k K\left(\frac{x_i - x}{h}\right) = 0,$$

$k = 0, \ldots, p$. In this equation, $\boldsymbol{\psi}$ is a κ dimensional function. This setting corresponds to replace in equation (2.4) the vector of first partial derivatives of the local log likelihood by the vector equations $\boldsymbol{\psi}$. Although Carroll, Ruppert and Welsh (1998) only calculated the expressions for bias and variance, it is expected that consistency and asymptotic normality results can be obtained similarly as in Theorems 2.1 and 2.2, under the appropriate regularity conditions.

Several extensions of the local polynomial estimating equations in a general estimating equations framework are presented in Chapter 4.

Chapter 3

Bootstrapping Local Polynomial Estimators in Likelihood-Based Models

3.1 Introduction

In this chapter we study statistical properties of local polynomial estimators in a multi-parameter "semilikelihood" context. Although a local likelihood technique already allows flexible nonparametric estimation of the systematic component, i.e. the functional dependence of the parameters of interest on explanatory variables (see Chapter 2), there is also a need for more flexibility with respect to the random component, the probability density function of the response variable. Especially for more complex type of data, there are several possible sensible choices of response pdf's on which to base a local likelihood. Likelihood misspecification and robustness, which have been studied extensively for parametric estimation (see White, 1994), are also important issues in the local setting. So, we no longer assume the response pdf, and hence the local likelihood, to be correctly specified. Moreover, for numerically intractable likelihood functions, we introduce a local version of the pseudolikelihood approach of Arnold and Strauss (1991), see also Besag (1975) and Cressie (1991). Semilikelihood represents pseudolikelihood with possible misspecification.

We show strong consistency of the local polynomial estimators in this context of model misspecification. Joint asymptotic normality of the estimators is easily

obtained, but the resulting distribution function is rather complicated, still containing several unknown components. Estimators for the asymptotic bias are studied via estimation of derivatives of the estimators. As an alternative to the asymptotic normal distribution, we propose a bootstrap approach based on a one-step bootstrap estimator. The method avoids any additional iterative model fitting in the bootstrap simulations and is robust against misspecification of the pdf.

The outline of this chapter is as follows. Section 3.2 introduces the local polynomial estimators in the context of misspecified likelihood-based models. Theorems showing strong consistency and asymptotic normality of the estimators are in Section 3.3, which also provides the estimation of the derivatives. The one-step bootstrap estimator is studied in Section 3.4. Some applications are briefly discussed in Section 3.5.1 and are illustrated by a data example. All regularity conditions and some technical lemmas can be found in Section 3.7.

The results of this chapter can also be found in Claeskens and Aerts (1998).

3.2 Local semilikelihood estimation

In this section, we first define the semilikelihood equations in a fully parametric context. As an illustration, we briefly describe the Molenberghs-Ryan (1999) statistical model for clustered binary data, which will be used for the example in Section 3.5.1.

3.2.1 Semilikelihood estimation

In the pseudolikelihood approach of Arnold and Strauss (1991), the probability density function of the data is replaced by a product of conditional pdf's which do not necessarily represent a joint pdf anymore. The motivation behind this estimation technique is to avoid the calculation of a complicated normalizing constant, which frequently arises in exponential family models (Arnold, Castillo and Sarabia, 1992).

Let $(\boldsymbol{x}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{Y}_n)$ be n independent observations where $\boldsymbol{x}_i$ is the covariate, $\boldsymbol{Y}_i$ the response vector and $g(\boldsymbol{y}, \boldsymbol{x})$ the true pdf of $\boldsymbol{Y}$ given x. The assumed pdf is denoted by $f(\boldsymbol{y}; \theta_1(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x}))$, which might not contain the true structure $g(\boldsymbol{y}, \boldsymbol{x})$. In case $\theta_1, \ldots, \theta_\kappa$ are modeled parametrically as a function of $\boldsymbol{x}$, this situation corresponds to the general misspecification setting as described in White (1994) and earlier references therein.

Here, in what we call semilikelihood estimation, we use the construction of pseudolikelihood in combination with the fact that the true data generating process g is not necessarily correctly described by the assumed model f.

Let A represent the set of all $2^m - 1$ vectors $\mathbf{a}$ of length m=dim($\boldsymbol{Y}$), consisting solely of zeros and ones, with each vector having at least one non-zero entry, and $\{\gamma_{\mathbf{a}} \mid \mathbf{a} \in A\}$ is a set of $2^m - 1$ known real numbers, not all zero. Denote by $\boldsymbol{Y}_i^{(\mathbf{a})}$ the subvector of $\boldsymbol{Y}_i$ corresponding to the non-zero components of $\mathbf{a}$ with associated assumed pdf $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x}))$. The logarithm of the semilikelihood is defined as

$$\sum_{\mathbf{a}\in A} \gamma_{\mathbf{a}} \sum_{i=1}^{n} \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \theta_1(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x})). \tag{3.1}$$

Examples are given below.

3.2.2 Models for clustered binary data

For clustered binary data several types of probability models (marginal, conditional, random effects models) are available – it is often not clear which probability model generates the data (for a recent survey, see Pendergast et al., 1996).

As an example we revisit the "Low-iron rat teratology data". In Chapter 2 we estimated the proportion of malformed foetuses (π) and the correlation (ρ) between foetuses having the same mother, both as a function of the mother's hemoglobin level, assuming a beta-binomial model.

The well-established *beta-binomial model* (Section 2.6.1) fits into the semilikelihood approach (3.1), as any other full likelihood model, by taking (for a cluster of size m) $\gamma_{\mathbf{a}} = 1$ for $\mathbf{a} = \mathbf{1}_m = (1, \ldots, 1)$ and $\gamma_{\mathbf{a}} = 0$ otherwise.

As a second clustered binary data model, we consider the *conditional model* of Molenberghs and Ryan (1999). They propose a likelihood model for multivariate clustered binary outcomes, based on the multivariate exponential family model as proposed by Cox (1972). The model benefits from the elegance and simplicity of exponential family theory. It allows for flexible response relationships and combines a likelihood basis with numerical stability. Let $\boldsymbol{y}$ be the observed vector, indicating whether a fetus is malformed or not, and let $y_\bullet$ be the total number of malformed foetuses in that litter. The pdf is given by

$$f(\boldsymbol{y}; \theta_1, \theta_2) = \exp\{y_\bullet\theta_1 - y_\bullet(m - y_\bullet)\theta_2 - A(\theta_1, \theta_2)\}$$

where $A(\cdot)$ is a (rather complicated) normalizing constant. A problem however, particularly with large clusters, is the evaluation of this constant $A(\cdot)$. Geys, Molenberghs and Ryan (1997, 1999) propose the use of a pseudolikelihood for correlated outcomes. The principal idea is to replace the joint pdf by a product of m univariate conditional pdf's. The advantage of this particular type of model is that the normalizing constant cancels. This so-called *full conditional pseudolikelihood model* corresponds to the choice $\gamma_{\mathbf{a}} = m$ if $\mathbf{a} = \mathbf{1}_m$, $\gamma_{\mathbf{a}} = -1$ if $\mathbf{a}$ consists of ones everywhere, except for the ℓth entry ($\ell = 1, \ldots, m$) and $\gamma_{\mathbf{a}} = 0$ otherwise (see also Section 8.5). Pseudolikelihood estimation is widely applicable, see, e.g., Besag (1975), Arnold and Strauss (1991), Cressie (1991).

Note that all three models, beta-binomial, Molenberghs-Ryan and its pseudolikelihood version, reduce to the logistic regression model when there is no clustering ($\rho = 0$ in the beta-binomial model and $\theta_2 = 0$ in the other two models).

3.2.3 Local estimators

The pseudolikelihood models of Arnold and Strauss (1991) have been analyzed only in a parametric way, that is, one specifies each of the κ unknown parameters as real-valued functions of the covariate $\boldsymbol{x}$. Here, we do not specify any functional relationship, resulting in a model $f(\boldsymbol{y}, \theta_1(\boldsymbol{x}), \ldots, \theta_\kappa(\boldsymbol{x}))$, where the functions $\theta_1(\cdot), \ldots, \theta_\kappa(\cdot)$ are unknown. The local polynomial estimation technique is used to estimate each parameter $\theta_r(\boldsymbol{x})$ locally by a polynomial of degree p_r. Although all results can be formulated for multivariate explanatory variables and for the random design case, we chose to present theorems and their proofs for the fixed design (2.1) generated by a known univariate pdf $f_X(x)$ with $\text{supp}(f_X) = [b_1, b_2]$. The *true* parameters $(\theta_1(x), \ldots, \theta_\kappa(x))$ maximize the following expression:

$$\sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} E\left[\log f^{(\mathbf{a})}(\boldsymbol{Y}^{(\mathbf{a})}; \theta_1(x), \ldots, \theta_\kappa(x))\right] \tag{3.2}$$

where $A_0 = \{\mathbf{a} \in A | \gamma_{\mathbf{a}} \neq 0\}$. This is a generalization of the Kullback-Leibler information criterion to the semilikelihood context. Here and in the sequel E and P denote the expectation and probability with respect to the true g. Among other results, Gourieroux, Monfort and Trognon (1984) show that if the pdf f belongs to a quadratic exponential family, parameters appearing in first and second moments, can be consistently estimated even if the "true" pdf does not belong to this family.

In practical situations where the pdf of the data is hardly ever known, we are doing "the best one can do".

We now define the local polynomial estimators as the maximizers of the kernel weighted log semilikelihood function. For a fixed $x \in [b_1, b_2]$, denote $\boldsymbol{\theta}_r^T(x) = (\theta_{r0}(x), \ldots, \theta_{rp_r}(x))$ the vector containing $\theta_r(x)$ and higher order derivatives (see Section 2.2). Local polynomial fitting provides estimates for

$$\boldsymbol{\theta}(x) = (\boldsymbol{\theta}_1^T(x), \ldots, \boldsymbol{\theta}_\kappa^T(x))^T.$$

The local polynomial maximum semilikelihood estimator (MSLE)

$$\widehat{\boldsymbol{\theta}}_n(x,h) \equiv (\widehat{\boldsymbol{\theta}}_1^T(x), \ldots, \widehat{\boldsymbol{\theta}}_\kappa^T(x))^T = \left(\hat{\theta}_{10}(x), \ldots, \widehat{\theta}_{1p_1}(x), \ldots, \widehat{\theta}_{\kappa 0}(x), \ldots, \widehat{\theta}_{\kappa p_\kappa}(x)\right)^T$$

maximizes the kernel weighted log semilikelihood function, which is defined as

$$\mathcal{L}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\kappa) = \frac{1}{n} \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} \sum_{i=1}^{n} \log f^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \boldsymbol{\beta}_1^T \boldsymbol{X}_{i1}, \ldots, \boldsymbol{\beta}_\kappa^T \boldsymbol{X}_{iv}) K_h(x_i - x) \quad (3.3)$$

with respect to $(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_\kappa^T) = (\beta_{10}, \ldots, \beta_{1p_1}, \ldots, \beta_{\kappa 0}, \ldots, \beta_{\kappa p_\kappa})$, where $\boldsymbol{X}_{ij}$ is the column vector $(1, (x_i - x), \ldots, (x_i - x)^{p_j})^T$.

This definition is an extension to the semilikelihood setting of the local likelihood estimators as defined by equation (2.2).

In Figure 3.1, local linear MSL estimators are shown for the low-iron rat teratology data, using all models introduced in Section 3.2.2, including the logistic regression model (put $\rho = 0$ in the beta-binomial model). Figure 3.1a shows the proportion of death foetuses as a function of the hemoglobin level of the mother. The beta-binomial estimate is shown in the solid line type, the logistic regression estimate is represented by the dotted line. Figure 3.1b-c are the estimates for the Molenberghs-Ryan model (solid line) and the full conditional semilikelihood model (dotted line).

3.3 Asymptotic results

This section focuses on some strong consistency results, derivative estimation and the construction of bias estimators. These latter estimators will be incorporated in the bootstrap procedures defined in Section 3.4. Theorems are stated in general but, for simplicity, all proofs are given for the case of two parameters ($\kappa = 2$).

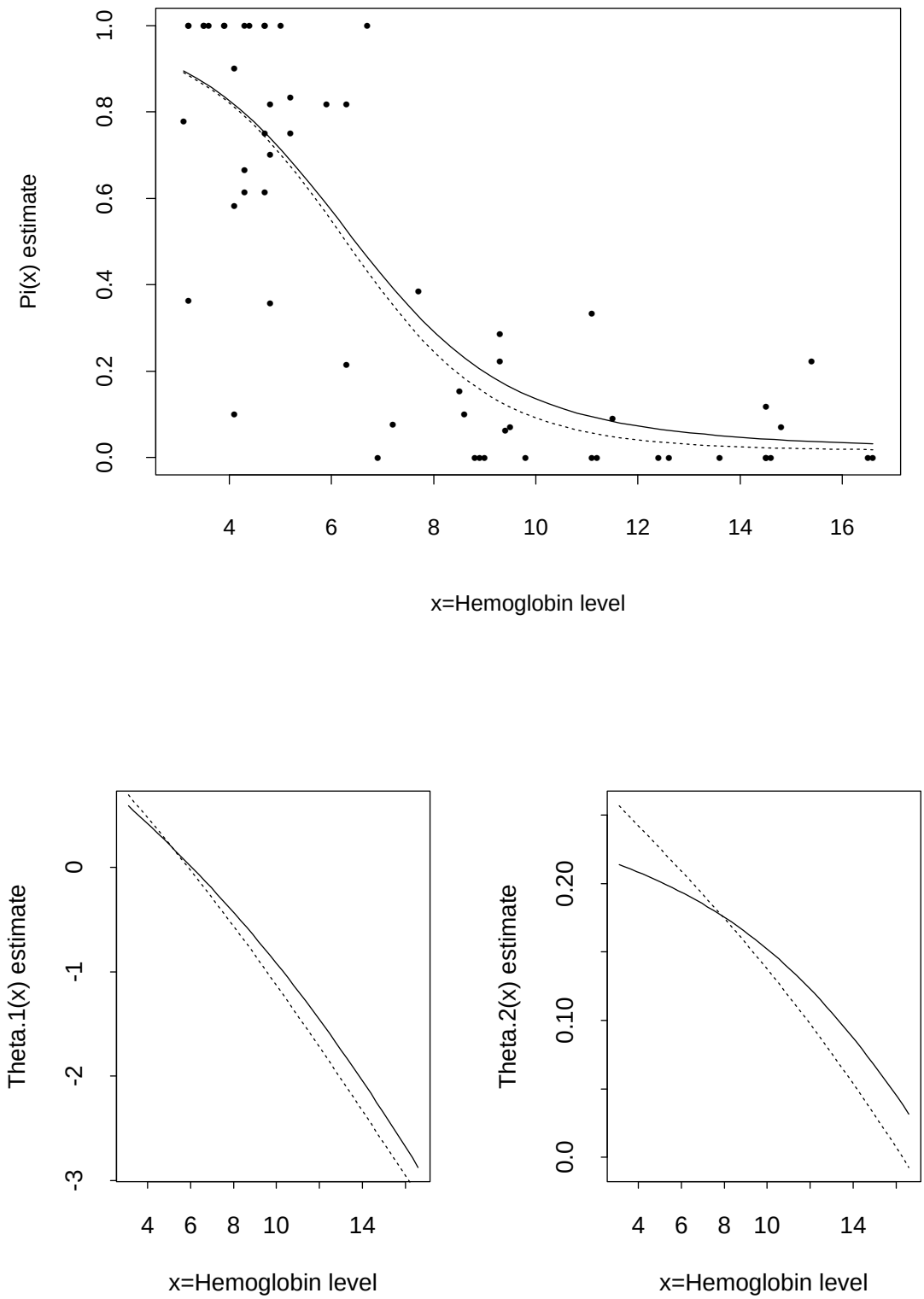


Figure 3.1: Local linear semilikelihood estimates for the low-iron rat teratology data.

3.3.1 Regularity conditions

To examine the asymptotic properties of these estimators, we need to reformulate the regularity conditions for the pseudolikelihood pdf's. All regularity conditions and proofs are stated for the two-parameter case ($\kappa = 2$) and for fixed design.

The basic assumptions (G), (K) and (S) on the design density, the kernel and the smoothness of the curves $\theta_1(\cdot), \ldots, \theta_\kappa(\cdot)$ remain the same. Assumption (H) is replaced by (H').

(H') $h \to 0$, $nh^2 \to \infty$ and $nh/\log n \to \infty$ as $n \to \infty$.

Next, following assumptions are needed on the semilikelihood.

(C1) The densities $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)$ are distinct for different values of the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The support of $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. There exists an open subset Θ of the parameter space containing the true parameters $(\theta_1(x), \theta_2(x))$ such that for almost all $\boldsymbol{y}$ and for all $\mathbf{a}$ in $A_0 = \{\mathbf{a} \in A | \gamma_{\mathbf{a}} \neq 0\}$, the density $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)$ admits all third derivatives for all $\boldsymbol{\theta}$ in Θ.

(C2) There exists a $\delta > 0$, such that for each $\mathbf{a}, \mathbf{a}'$ in A_0, there exist functions H_i, i=1,2,3 such that $|q_r^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)| \leq H_1(\boldsymbol{y}^{(\mathbf{a})})$, $|q_{rs}^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)| \leq H_2(\boldsymbol{y}^{(\mathbf{a})})$, and $|q_r^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)\ q_s^{(\mathbf{a}')}(\boldsymbol{y}^{(\mathbf{a}')}; \theta_1, \theta_2)| \leq H_3(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{y}^{(\mathbf{a}')})$, for all $\boldsymbol{\theta} \in \Theta$ $(r, s = 1, 2)$ and $E[H_i^{2+\delta}(\boldsymbol{Y}^{(\mathbf{a})})]$, $(i = 1, 2)$ and $E[H_3^{2+\delta}(\boldsymbol{Y}^{(\mathbf{a})}, \boldsymbol{Y}^{(\mathbf{a}')})]$ are uniformly bounded on Θ.

(C3) There exists a function $J(\boldsymbol{y}^{(\mathbf{a})})$ such that $|q_{rst}^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1, \theta_2)| \leq J(\boldsymbol{y}^{(\mathbf{a})})$ for all $\boldsymbol{\theta}$ in Θ $(r, s, t = 1, 2)$ and $J(\boldsymbol{Y}^{(\mathbf{a})})$ is almost surely uniformly bounded on Θ.

(C4) For each $r, s = 1, 2$ $E\left[|q_{rs}^{(\mathbf{a})}(\boldsymbol{Y}^{(\mathbf{a})}; \boldsymbol{\theta})|\right] < \infty$ and the 2×2 matrix $\boldsymbol{J}(\theta_1(x), \theta_2(x))$ (see equation (3.7)) is positive definite. $J_{rs}(\theta_1, \theta_2)$ is Lipschitz continuous and differentiable at $(\theta_1(x), \theta_2(x))$.

3.3.2 Strong consistency and asymptotic normality

The next theorem guarantees the existence of at least one solution of the semilikelihood equations which is strongly consistent. In the proof of this theorem, Lemmas 3.3 and 3.4 of Section 3.7 are used.

Theorem 3.1 *Suppose that the regularity conditions (C1)-(C4) on* $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})};\theta_1,\ldots,\theta_\kappa)$ *(for all* $\mathbf{a}$ *in* A_0*) and conditions (G), (K) and (H') hold, and in addition that* $n^{(\delta-\eta)/(2+\delta-\eta)}h/\log n \to \infty$ *for some* $0<\eta<\min(1,\delta)$*. If the functions* $\theta_1(x),\ldots,\theta_\kappa(x)$ *satisfy smoothness condition (S), then there exists a solution* $\widehat{\boldsymbol{\theta}}_n(x,h)$ *of the semilikelihood equations* $\partial/\partial\boldsymbol{\beta}_r[\mathcal{L}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa)]=0, r=1,\ldots,\kappa$ *such that* $\widehat{\boldsymbol{\theta}}_n(x,h)$ *is strongly consistent for estimating* $\boldsymbol{\theta}(x)$*.*

Proof. By a Taylor expansion of $\frac{\partial}{\partial\beta_{rk}}\mathcal{L}_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ about $(\boldsymbol{\theta}_1^T(x),\boldsymbol{\theta}_2^T(x))$ $(r=1,2;$ $k=1,\ldots,p_r)$, we get that

$$\frac{\partial}{\partial\beta_{rk}}\mathcal{L}_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)=A_{rk}^n(x)+\sum_{s=1}^{2}\sum_{\ell=0}^{p_s}B_{rsk\ell}^n(x)(\beta_{s\ell}-\theta_{s\ell}(x))$$
$$+\ \frac{1}{2}\sum_{s=1}^{2}\sum_{t=1}^{2}\sum_{\ell=0}^{p_s}\sum_{m=0}^{p_t}(\beta_{s\ell}-\theta_{s\ell}(x))(\beta_{tm}-\theta_{tm}(x))\alpha^*H_{k\ell m}^n(x).$$

In this equation,

$$\begin{aligned}
A_{rk}^n(x) &= \frac{1}{n}\sum_{i=1}^{n}K_h(x_i-x)(x_i-x)^k\tilde{A}_r^i(x),\\
B_{rsk\ell}^n(x) &= \frac{1}{n}\sum_{i=1}^{n}K_h(x_i-x)(x_i-x)^{k+\ell}\tilde{B}_{rs}^i(x),\\
\tilde{A}_r^i(x) &= \sum_{\mathbf{a}\in A_0}\gamma_{\mathbf{a}}q_r^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})};\bar{\theta}_1(x,x_i),\bar{\theta}_2(x,x_i)),\\
\tilde{B}_{rs}^i(x) &= \sum_{\mathbf{a}\in A_0}\gamma_{\mathbf{a}}q_{rs}^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})};\bar{\theta}_1(x,x_i),\bar{\theta}_2(x,x_i))\\
q_r^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})};v_1,v_2) &= \frac{\partial}{\partial u_r}\log f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})};u_1,u_2)|_{(u_1,u_2)=(v_1,v_2)}\ \ r=1,2\\
q_{rs}^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})};v_1,v_2) &= \frac{\partial^2}{\partial u_r\partial u_s}\log f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})};u_1,u_2)|_{(u_1,u_2)=(v_1,v_2)}\ \ r,s=1,2
\end{aligned}\tag{3.4}$$

and, using (C3),

$$H_{k\ell m}^n(x)=\frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{a}\in A_0}\gamma_{\mathbf{a}}K_h(x_i-x)(x_i-x)^{k+\ell+m}J(\boldsymbol{Y}_i^{(\mathbf{a})})$$

with $|\alpha^*|\le 1$.

The choice of η, (H') and (C2) are sufficient to apply Lemma 3.4 with $V_i=\tilde{A}_r^i(x)$, resp. $V_i=\tilde{B}_{rs}^i(x)$. We now immediately obtain that $A_{rk}^n(x)-E\left[A_{rk}^n(x)\right]\to 0$ a.s. and,

respectively, $B^n_{rsk\ell}(x) - E\left[B^n_{rsk\ell}(x)\right] \to 0$ a.s. Lemma 3.3 gives the limit expressions of $E[A^n_{rk}(x)]$ and $E[B^n_{rsk\ell}(x)]$, see the proof of Theorem 2.1. Similarly, by using (C3) one can show that $H^n_{k\ell m}(x)$ is bounded, almost surely. An application of an extension to multidimensional parameters of e.g. the proof of Theorem 4.2.2 in Serfling (1980) p.145-148, completes the proof.

For bandwidth sequences of the form $h = cn^{-b}$ where $0 < b < 1$ and c is a constant, the bandwidth conditions are satisfied if (H') holds and $\delta > 2b/(1-b)$.

To formulate an asymptotic normality result, we need some further notation, which partly extends the notation used in Chapter 2. To avoid notational complexity, the following definitions are restricted to the case $\kappa = 2$ (but can easily be extended).

Let $\boldsymbol{H}_r = \mathrm{diag}(1, h, \ldots, h^{p_r})$ $(r = 1, 2)$, $\boldsymbol{H} = \mathrm{diag}(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_\kappa)$ and

$$\boldsymbol{\Sigma}_x = f_X(x)\boldsymbol{J}(\theta_1(x), \theta_2(x)) \otimes \begin{pmatrix} \boldsymbol{N}_{p_1p_1}(x) & \boldsymbol{N}_{p_1p_2}(x) \\ \boldsymbol{N}_{p_2p_1}(x) & \boldsymbol{N}_{p_2p_2}(x) \end{pmatrix} \tag{3.5}$$

$$\boldsymbol{\Gamma}_x = f_X(x)\boldsymbol{K}(\theta_1(x), \theta_2(x)) \otimes \begin{pmatrix} \boldsymbol{T}_{p_1p_1}(x) & \boldsymbol{T}_{p_1p_2}(x) \\ \boldsymbol{T}_{p_2p_1}(x) & \boldsymbol{T}_{p_2p_2}(x) \end{pmatrix} \tag{3.6}$$

where $\otimes$ is the generalized Kronecker product, $\boldsymbol{J}(\theta_1, \theta_2)$ is the matrix with components

$$J_{rs}(\theta_1, \theta_2) = \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} E[-q^{(\mathbf{a})}_{rs}(\boldsymbol{Y}^{(\mathbf{a})}; \theta_1, \theta_2)] \quad (r, s = 1, 2) \tag{3.7}$$

and $\boldsymbol{K}(\theta_1, \theta_2)$ is the matrix with components

$$K_{rs}(\theta_1, \theta_2) = \sum_{\mathbf{a} \in A_0} \sum_{\mathbf{a}' \in A_0} \gamma_{\mathbf{a}} \gamma_{\mathbf{a}'} E[q^{(\mathbf{a})}_r(\boldsymbol{Y}^{(\mathbf{a})}; \theta_1, \theta_2) q^{(\mathbf{a}')}_s(\boldsymbol{Y}^{(\mathbf{a}')}; \theta_1, \theta_2)] \quad (r, s = 1, 2).$$

Finally let

$$\boldsymbol{\Lambda}_x = \begin{pmatrix} \frac{d}{dx}\{f_X(x)J_{11}(\theta_1(x), \theta_2(x))\} & \frac{d}{dx}\{f_X(x)J_{12}(\theta_1(x), \theta_2(x))\} \\ \frac{d}{dx}\{f_X(x)J_{21}(\theta_1(x), \theta_2(x))\} & \frac{d}{dx}\{f_X(x)J_{22}(\theta_1(x), \theta_2(x))\} \end{pmatrix} \otimes \begin{pmatrix} \boldsymbol{Q}_{p_1p_1}(x) & \boldsymbol{Q}_{p_1p_2}(x) \\ \boldsymbol{Q}_{p_2p_1}(x) & \boldsymbol{Q}_{p_2p_2}(x) \end{pmatrix}.$$

Similar to the results in correctly specified likelihood models, the asymptotic normality of the local polynomial MSLE follows from that of $\boldsymbol{W}^n(x) = (\boldsymbol{W}_1^n(x)^T, \boldsymbol{W}_2^n(x)^T)^T$ where $\boldsymbol{W}_r^n(x)$ $(r = 1, 2)$ is the column vector

$$(nh)^{-1/2} \sum_{i=1}^{n} K(\frac{x_i - x}{h}) \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} q_r^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \bar{\theta}_1(x, x_i), \bar{\theta}_2(x, x_i)) \boldsymbol{H}_r^{-1} \boldsymbol{X}_{ir} \tag{3.8}$$

with $\bar{\theta}_r(x, x_i)$ as in (2.7).

Theorem 3.2 *Assume that all conditions of Theorem 3.1 hold. Then for* $n \to \infty$

$$\sup_{t \in I\!R^{\mathcal{P}}} \left| P\left(\boldsymbol{V}(\boldsymbol{\theta}(x))^{-1/2} \sqrt{nh} \left(\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_n(x, h) - \boldsymbol{\theta}(x)) - \boldsymbol{B}_n(\boldsymbol{\theta}(x)) \right) \leq t \right) - \Phi_{\mathcal{P}}(t) \right| = o(1)$$

where

$$\begin{aligned} \boldsymbol{B}_n(\boldsymbol{\theta}(x)) &= (nh)^{-1/2} (\boldsymbol{\Sigma}_x^{-1} - h \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Lambda}_x \boldsymbol{\Sigma}_x^{-1}) E[\boldsymbol{W}^n(x)] \\ \boldsymbol{V}(\boldsymbol{\theta}(x)) &= \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Gamma}_x \boldsymbol{\Sigma}_x^{-1}, \end{aligned}$$

and $\Phi_{\mathcal{P}}$ *is the* $\mathcal{P} = \sum_{i=1}^{\kappa} (p_i + 1)$ *dimensional multivariate standard normal distribution function.*

The proof of Theorem 3.2 is a straightforward generalization of the proof of Theorem 2.2. The same remarks concerning the asymptotic bias expression are valid here, leading to the conclusion that the polynomials should preferably be taken of the same degree, since the smallest degree determines the order of the asymptotic bias. Boundary bias and interior bias are of the same order of magnitude for odd degree polynomial fits. If all polynomials are of degree p, the "optimal" bandwidth for estimation of the curves, balancing bias squared and variance, is of the order $O(n^{-1/(4[p/2]+5)})$ where $[a]$ denote the largest integer $\leq a$. The marginal asymptotic normality of $\hat{\beta}_{rj}$ when all polynomials are of the same degree, can be found in Corollary 2.1 by replacing the Fisher information matrix $\boldsymbol{I}(\theta_1(x), \theta_2(x))$ by the product $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}(\theta_1(x), \theta_2(x))$. There it is seen that the asymptotic bias and variance expressions depend in a complicated way on several unknown quantities. In Section 3.4 we obtain bootstrap estimators for the joint distribution of the components of the vector $\hat{\boldsymbol{\theta}}(x, h)$. In the next sections we focus on the estimation of the bias of $\hat{\boldsymbol{\theta}}(x, h)$. A possible approach is based on the estimation of derivatives. We restrict attention to the case that all polynomials have degree p and to interior points x.

3.3.3 Estimation of derivatives

One of the advantages of local polynomial estimation is that the method automatically provides estimators for the unknown curves and all derivatives up to degree p. In the marginal bias expression however, $2([p/2]+1)$st derivatives appear. The question we address now is whether the local polynomial estimator itself can be differentiated. Since the local MSLEs are defined implicitly, this question is not straightforward to answer. Let g be a second bandwidth parameter. In the next theorem we show that, under some conditions, the jth derivatives of the MSLE $\hat{\theta}_{r0}(x,g)$ exist, and the method of proof suggests a technique to obtain these estimators. Of course, the results of this theorem are mostly of interest for $j > p$. We use the notation $\boldsymbol{\theta}^{(j)}(x) = (d^j/dx^j)(\theta_1(x),\ldots,\theta_\kappa(x))^T$ and $\hat{\boldsymbol{\theta}}_k^{(j)}(x,g) = (d^j/dx^j)(\widehat{\theta}_{1k}(x,g),\ldots,\widehat{\theta}_{\kappa k}(x,g))^T, 0 \le k \le p, j \ge 0$.

Some additional assumptions are necessary for the estimation of the derivatives of the estimators. Condition (D2) is a simple extension of Iverson and Randles' (1989) condition to this context (see their Theorem 2.9). First, we define $\tilde{\boldsymbol{q}}^{(j)}(\boldsymbol{Y},\beta_1,\beta_2)$ as being the vector containing all jth partial derivatives of $\sum_{\mathbf{a}\in A_0}\gamma_{\mathbf{a}}\log f^{(\mathbf{a})}(\boldsymbol{Y}^{(\mathbf{a})},\beta_1,\beta_2)$ with respect to (β_1,β_2).

(D1) K and $f_X(\cdot)$ are m times differentiable, $K^{(k)}(-1) = K^{(k)}(1) = 0$, $0 \le k \le m$, $K^{(m)}$ is Lipschitz continuous, $g \to 0$ and $ng^{2m+1}/\log n \to \infty$.

(D2) For each value x_0, define $\boldsymbol{X}_0$ as the column vector $(1,(x_0-x),\ldots,(x_0-x)^p)^T$. There exists a neighborhood $\mathcal{N}(x)$ of x such that for all x_0 in $\mathcal{N}(x)$, for all $t = 1,\ldots,m+2$,

(i) the expectation $E[\tilde{q}^{(t)}(\boldsymbol{Y},\boldsymbol{\theta}_1^T\boldsymbol{X}_0,\boldsymbol{\theta}_2^T\boldsymbol{X}_0)]$ is a bounded and Lipschitz continuous function of x_0, uniformly on Θ and for $t \ge 2$, $(m-t+2)$ times differentiable,

and, uniformly in x_0 in $\mathcal{N}(x)$,

(ii) for each component $\tilde{q}^{(j)}$ of $\tilde{\boldsymbol{q}}^{(j)}$,

$$\lim_{d\to 0} E[\sup_{\mathbf{s}\in\mathcal{B}(d)} |\tilde{q}^{(t)}(\boldsymbol{Y};(\boldsymbol{\theta}_1+\boldsymbol{s})^T\boldsymbol{X}_0,(\boldsymbol{\theta}_2+\boldsymbol{s})^T\boldsymbol{X}_0) - \tilde{q}^{(t)}(\boldsymbol{Y};\boldsymbol{\theta}_1^T\boldsymbol{X}_0,\boldsymbol{\theta}_2^T\boldsymbol{X}_0)|] = 0$$

where $\mathcal{B}(d)$ is a $\mathcal{P}$ dimensional sphere centered at the origin with radius $d > 0$,

(iii) there exists a $\delta > 0$ and there exists a $d > 0$ such that for each component of $\tilde{\boldsymbol{q}}^{(j)}$,

$$E[\{\sup_{\mathbf{s}\in\mathcal{B}(d)} |\tilde{q}^{(t)}(\boldsymbol{Y}; (\boldsymbol{\theta}_1 + \boldsymbol{s})^T \boldsymbol{X}_0, (\boldsymbol{\theta}_2 + \boldsymbol{s})^T \boldsymbol{X}_0) - \tilde{q}^{(t)}(\boldsymbol{Y}; \boldsymbol{\theta}_1^T \boldsymbol{X}_0, \boldsymbol{\theta}_2^T \boldsymbol{X}_0)|\}^{2+\delta}] < \infty.$$

Theorem 3.3 *If all conditions of Theorem 3.1 hold for bandwidth g and conditions (D1)-(D2) hold with $m = j$, then the vector $\hat{\boldsymbol{\theta}}_0^{(j)}(x, g)$ is strongly consistent for estimating $\boldsymbol{\theta}^{(j)}(x)$.*

Proof. Define, for $r = 1, 2$, $\tilde{\boldsymbol{\theta}}_r(x) = (u_{r0}, \hat{\theta}_{r1}(x, g), \ldots, \hat{\theta}_{rp}(x, g))^T$ and

$$u_r^n(u_{10}, u_{20}, x) = n^{-1} \sum_{i=1}^{n} \sum_{\mathbf{a}\in A_0} \gamma_{\mathbf{a}} K_g(x_i - x) q_r^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \tilde{\boldsymbol{\theta}}_1(x)\boldsymbol{X}_{1i}, \tilde{\boldsymbol{\theta}}_2(x)\boldsymbol{X}_{2i}).$$

From Theorem 3.1 we know that there exists a strongly consistent solution $(\hat{\theta}_{10}(x, g), \hat{\theta}_{20}(x, g))$ for the set of equations $u_r^n(u_{10}, u_{20}, x) = 0, r = 1, 2$. An application of Lemma 3.5 and Lemma 3.3 gives that

$$(\partial u_r^n / \partial u_{s0})(\hat{\theta}_{10}(x, g), \hat{\theta}_{20}(x, g), x) \to -J_{rs}(\theta_1(x), \theta_2(x)) f_X(x) \nu_0(\mathcal{R}_x)$$

almost surely as $n \to \infty$, for $r, s = 1, 2$. Hence, for n sufficiently large, the determinant of the matrix of these partial derivatives will be non-zero. Next, we apply the implicit function theorem to obtain that both $\hat{\theta}_{10}(x, g)$ and $\hat{\theta}_{20}(x, g)$ are C^1 functions and that $u_r^n(\hat{\theta}_{10}(\cdot, g), \hat{\theta}_{20}(\cdot, g), \cdot) = 0$ in some neighborhood of x. Therefore, also $v_r^n(\hat{\theta}_{10}(x, g), \hat{\theta}_{20}(x, g), (d/dx)\hat{\theta}_{10}(x, g), (d/dx)\hat{\theta}_{20}(x, g), x)$, the total derivative of $u_r^n(\cdot)$, equals zero. Since

$$v_r^n(u_{10}, u_{20}, v_{10}, v_{20}, x) = \sum_{j=1}^{2} \frac{\partial u_r^n}{\partial u_{j0}}(u_{10}, u_{20}, x) v_{j0} + \frac{\partial u_r^n}{\partial x}(u_{10}, u_{20}, x), \tag{3.9}$$

we have that $\partial v_r^n(\cdot)/\partial v_{10} = \partial u_r^n(\cdot)/\partial u_{10}$. Similar arguments as above yield the existence of continuous functions $(d^2/dx^2)\hat{\theta}_{r0}(x)$, $(r = 1, 2)$. Repeating this another $(j-2)$ times gives the existence of continuous jth derivatives of the estimators $\hat{\theta}_{j0}(x)$. The consistency of the estimators is proven by similar arguments as in Foutz (1977). First, we show that the Jacobian matrix of the resulting set of $2(j + 1)$ equations converges almost surely to some finite limit, uniformly in a neighborhood about x.

This is obtained by applying Lemmas 3.3 and 3.5, assuming conditions (D1) and (D2). After some calculations, the results of Lemmas 3.3 and 3.4 guarantee that the true vector of derivatives solves this set of equations, as n tends to infinity. An application of the inverse function theorem now leads to the desired result.

From the proof it is clear how to construct derivative estimators in practice. For example, equation (3.9) can be solved for v_{10}, v_{20} to obtain the estimator $\hat{\boldsymbol{\theta}}_0^{(1)}(x,g)$, and so on. In the special case of Gaussian responses and $p = 0$, the resulting estimators are identical to those studied by Schuster and Yakowitz (1979).

Especially for p large, this technique can be computationally unattractive because of the calculation of higher order partial derivatives of $\log f$. An elegant alternative then is to exploit the properties of the local polynomial MSLE. Since we already know that $\widehat{\theta}_{rj}(x)$ is a strongly consistent estimator for $\theta_r^{(j)}(x)/j!$ for $j = p-1, p$, the following corollary can be used to construct derivative estimators.

Corollary 3.1 *Assume all conditions of Theorem 3.1 for bandwidth g and conditions (D1)-(D2) with $m = 2$. Let the bandwidth g be such that $ng^{4[p/2]+5}/\log n \to \infty$. If p is odd, then $(p-1)!\hat{\boldsymbol{\theta}}_{p-1}^{(2)}(x,g)$ is strongly consistent for estimating $\boldsymbol{\theta}^{(p+1)}(x)$. In case p is even, $p!\hat{\boldsymbol{\theta}}_p^{(1)}(x,g)$ is strongly consistent for estimating $\boldsymbol{\theta}^{(p+1)}(x)$ and $p!\hat{\boldsymbol{\theta}}_p^{(2)}(x,g)$ is strongly consistent for estimating $\boldsymbol{\theta}^{(p+2)}(x)$.*

Proof. For p odd, define $\tilde{\boldsymbol{\theta}}_r(x) = (\hat{\theta}_{r0}(x,g), \ldots, \hat{\theta}_{r,p-2}(x,g), u_{r,p-1}, \hat{\theta}_{rp}(x,g))^T$, and

$$u_r^n(u_{1,p-1}, u_{2,p-1}, x) = n^{-1} \sum_{i=1}^{n} \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} K_h(x_i - x) q_r^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \tilde{\boldsymbol{\theta}}_1(x)\boldsymbol{X}_{1i}, \tilde{\boldsymbol{\theta}}_2(x)\boldsymbol{X}_{2i}).$$

For p even, we take, $\tilde{\boldsymbol{\theta}}_r(x) = (\hat{\theta}_{r0}(x,g), \ldots, \hat{\theta}_{r,p-1}(x,g), u_{rp})^T$, and

$$u_r^n(u_{1,p}, u_{2,p}, x) = n^{-1} \sum_{i=1}^{n} \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} K_h(x_i - x) q_r^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \tilde{\boldsymbol{\theta}}_1(x)\boldsymbol{X}_{1i}, \tilde{\boldsymbol{\theta}}_2(x)\boldsymbol{X}_{2i}).$$

Now, we proceed as in the proof of Theorem 3.3. The different treatment of p odd and p even is due to the symmetry of the kernel, which causes, for interior points, odd moments of K to be zero.

In Theorem 3.3 and Corollary 3.1, the condition on bandwidth g, used to estimate the derivatives, excludes the optimal bandwidth h for estimation of the curves themselves. Therefore it is necessary to use two bandwidth sequences, a well-known feature of derivative estimation in nonparametric regression (see, e.g., Gasser and Müller, 1984, Mack and Müller, 1989).

3.3.4 Estimating the bias

A first method is based on the explicit expression for the leading term of the asymptotic bias of the vector $\hat{\boldsymbol{\theta}}(x,h)$. This leading term depends on bandwidth h, kernel K and the unknown derivatives of $\boldsymbol{\theta}^{(p+1)}(x)$ (see Chapter 2). Based on the derivative estimators introduced in the previous sections, a typical direct plug-in approach suggests a possible bias estimator. For p odd, this leads to the following estimator,

$$\hat{\boldsymbol{B}}_n(x,g,h) = \frac{h^{p+1}}{p(p+1)}\hat{\boldsymbol{\theta}}^{(2)}_{p-1}(x,g) \otimes \mathcal{M}_p(x) \otimes \mathbf{e}_1,$$

where $\mathbf{e}_1 = (1,0)^T$, $\mathcal{M}_p(x)$ is defined as the column vector with $(j+1)$st element $\int_{\mathcal{R}_x} z^{p+1} K_{2j,p}(z)dz$, $j = 0,\ldots,[p/2]$ and $\otimes$ is the Kronecker product. The next corollary states the bias estimator's consistency. The proof follows immediately from Corollary 3.1.

Theorem 3.4 *Assume all conditions of Theorem 3.1 on bandwidth g, (D1)-(D2) with $m = 2$ and $ng^{2p+3}/\log n \to \infty$. If $h = O(n^{-1/(2p+3)})$, then*

$$\sqrt{nh}\left(\hat{\boldsymbol{B}}_n(x,g,h) - \boldsymbol{B}_n(\boldsymbol{\theta}(x))\right) \to 0, \quad a.s.\ P.$$

Another bias estimator is based on expression (3.8). The bias of the estimator finds its origin in the approximation of $\theta_k(x_i)$ by a finite number of terms (i.e. the degree p of the polynomial) in a Taylor series about x. This is the intuitive idea behind the numerator in the bias estimator $\check{\boldsymbol{B}}_n(x,g,h)$. The first term is defined in terms of the local polynomial approximations, while the second term contains estimators at the data points x_i. Define

$$\check{\boldsymbol{B}}_n(x,g,h) = \left\{-\boldsymbol{A}_n\left(\widehat{\boldsymbol{\theta}}_1(x,g),\ldots,\widehat{\boldsymbol{\theta}}_v(x,g)\right)\right\}^{-1} \boldsymbol{U}_n\left(\widehat{\boldsymbol{\theta}}_1(x,g),\ldots,\widehat{\boldsymbol{\theta}}_v(x,g)\right)$$

where, for $r,s = 1\ldots,\kappa, \boldsymbol{U}_n(\cdot) = (\boldsymbol{U}_n(\cdot)_1,\ldots,\boldsymbol{U}_n(\cdot)_\kappa)^T$

$$\begin{aligned}
\boldsymbol{U}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa)_r &= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\boldsymbol{\eta}_i(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa)_r \\
&\quad - \sum_{\mathbf{a}\in A_0}\gamma_{\mathbf{a}} q_r^{(\mathbf{a})}\left(\boldsymbol{Y}_i^{(\mathbf{a})};\widehat{\theta}_{10}(x_i,g),\ldots,\widehat{\theta}_{v0}(x_i,g)\right)\boldsymbol{X}_{ir}K_h(x_i-x)\Bigg), \\
\boldsymbol{A}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa) &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}_i(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa),
\end{aligned}$$

and, for $\boldsymbol{\eta}_i(\cdot) = (\boldsymbol{\eta}_i(\cdot)_1, \ldots, \boldsymbol{\eta}_i(\cdot)_\kappa)^T$, and $\boldsymbol{\alpha}_i(\cdot)$ the matrix with submatrices $\boldsymbol{\alpha}_i(\cdot)_{rs}$,

$$\boldsymbol{\eta}_i(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\kappa)_r = \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} q_r^{(\mathbf{a})} \left(\boldsymbol{Y}_i^{(\mathbf{a})}; \boldsymbol{\beta}_1^T \boldsymbol{X}_{i1}, \ldots, \boldsymbol{\beta}_\kappa^T \boldsymbol{X}_{i\nu} \right) \boldsymbol{X}_{ir} K_h(x_i - x), \quad (3.10)$$

$$\boldsymbol{\alpha}_i(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\kappa)_{rs} = \sum_{\mathbf{a} \in A_0} \gamma_{\mathbf{a}} q_{rs}^{(\mathbf{a})} \left(\boldsymbol{Y}_i^{(\mathbf{a})}; \boldsymbol{\beta}_1^T \boldsymbol{X}_{i1}, \ldots, \boldsymbol{\beta}_\kappa^T \boldsymbol{X}_{i\nu} \right) \boldsymbol{X}_{ir} \boldsymbol{X}_{is}^T K_h(x_i - x) \quad (3.11)$$

Under similar assumptions, Theorem 3.4 remains valid for $\check{\boldsymbol{B}}_n(x, g, h)$.

3.3.5 Estimating the variance

Without going into too many details we here give a motivation of how a strongly consistent variance estimator might be obtained. It basically uses the "sandwich" construction (see also Carroll, Ruppert and Welsh, 1998).

Under the previous set of conditions we obtain from the proof of Theorem 3.1, and by using similar arguments as in the proof of Theorem 2.2, that

$$\frac{B_{rskl}^n}{nh^{k+l}} - (\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x)_{kl} \to 0, \quad \text{a.s.}$$

If products of the first and second order partial derivatives of the log (semi-) likelihood are uniformly bounded, via similar calculations, we should obtain that the difference of

$$\frac{1}{n^2} \sum_{i=1}^n K_h^2(x_i - x) \left(\frac{x_i - x}{h} \right)^{k+l} \sum_{\mathbf{a} \in A_0} \sum_{\mathbf{a}' \in A_0} \gamma_{\mathbf{a}} \gamma_{\mathbf{a}'} q_r^{(\mathbf{a})} (\boldsymbol{Y}_i^{(\mathbf{a})}, \sum_{j=0}^{p_1} \widehat{\theta}_{1j}(x)(x_i - x)^j,$$

$$\sum_{j=0}^{p_2} \widehat{\theta}_{2j}(x)(x_i - x)^j) q_s^{(\mathbf{a}')} (\boldsymbol{Y}_i^{(\mathbf{a}')}, \sum_{j=0}^{p_1} \widehat{\theta}_{1j}(x)(x_i - x)^j, \sum_{j=0}^{p_2} \widehat{\theta}_{2j}(x)(x_i - x)^j)$$

and $\boldsymbol{\Gamma}_x/(nh)$ converges to zero a.s. The form of the asymptotic variance in Theorem 3.2 now immediately suggests an estimator for the variance.

We will not discuss these variance estimators further, since our goal is to use bootstrap techniques to obtain an estimator for the full unknown joint distribution of the local polynomial MSLE.

3.4 The linear one-step bootstrap

For regression models of the form $Y_i = \theta(x_i) + \varepsilon_i$ with ε_i independent error terms, bootstrap methods based on residuals have been studied extensively by, e.g., Härdle

and Bowman (1988) and Härdle and Mammen (1993). There, bootstrap errors ε_i^* are drawn (with replacement) from the set of estimated residuals or from a two-point distribution (wild bootstrap). Response values Y_i^* are constructed as the sum of an initial estimate for $\theta(x_i)$ plus ε_i^*, thus mimicking the specific location structure of the regression model.

It is not clear how this approach can be generalized to multiparameter likelihood-based models. Indeed, a first question which arises is which kind of residuals should be used. Next, there is no simple additive error structure which can be reconstructed to get the Y_i^* variables. Of course, one could use a parametric bootstrap procedure but this would never be able to deal with possible likelihood misspecification.

So there is a need for a method avoiding the introduction of residuals. Therefore we propose to resample the score values $\boldsymbol{\eta}_i$ together with the matrix of differentiated scores $\boldsymbol{\alpha}_i$ (as defined in (3.10) and (3.11)). In this way, no explicit bootstrap data $\boldsymbol{Y}_i^*$ have to be generated, the method immediately provides bootstrap estimates. The linear one-step bootstrap estimator is defined as

$$\widehat{\boldsymbol{\theta}}_n^*(x,h) = \widehat{\boldsymbol{\theta}}_n(x,h) + \boldsymbol{H}^{-1}\hat{\boldsymbol{B}}_n(x,g,h) - \boldsymbol{A}_n^*\left(\widehat{\boldsymbol{\theta}}_n(x,h)\right)^{-1}\boldsymbol{U}_n^*\left(\widehat{\boldsymbol{\theta}}_n(x,h)\right), \quad (3.12)$$

where $\hat{\boldsymbol{B}}_n(x,g,h)$ is one of the bias estimators discussed in Section 3.3.4 and,

$$\begin{aligned}\boldsymbol{A}_n^*(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa) &= \frac{1}{n}\sum_{i=1}^n\{\boldsymbol{\alpha}_i(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa)\}^*,\\ \boldsymbol{U}_n^*(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa) &= \frac{1}{n}\sum_{i=1}^n\{\boldsymbol{\eta}_i(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_\kappa)\}^*.\end{aligned}$$

The pairs $(\boldsymbol{\alpha}_i,\boldsymbol{\eta}_i)^*$, $(i = 1,\ldots,n)$ are sampled with replacement from the set $\{(\boldsymbol{\alpha}_1,\boldsymbol{\eta}_1)\ldots,(\boldsymbol{\alpha}_n,\boldsymbol{\eta}_n)\}$.

Note that this definition of the linear one-step bootstrap estimator is based on a linear approximation of the semilikelihood equations. A similar one-step bootstrap estimator used for hypothesis testing in parametric models will be discussed in Chapter 8. The concept of one-step estimators is well-known and applied in several statistical estimation procedures, see, e.g., Shao and Tu (1995, chapter 8). This approach is also related to a bootstrap resampling scheme based on resampling the vector of first derivatives of the log likelihood, which is proposed by Kauermann and Tutz (1996), hereby extending the results of Hu and Zidek (1995). In the context of local likelihood estimation in generalized linear models, Kauermann and Tutz (1996)

define the bootstrap estimator based on a first step in a Fisher scoring algorithm. A similar method based on one-step Fisher scoring in the *bootstrap world* can be applied in our situation, leading to the following definition

$$\widehat{\boldsymbol{\theta}}_n^*(x,h) = \widehat{\boldsymbol{\theta}}_n(x,h) + \boldsymbol{H}^{-1}\hat{\boldsymbol{B}}_n(x,g,h) - \boldsymbol{A}_n\left(\widehat{\boldsymbol{\theta}}_n(x,h)\right)^{-1}\boldsymbol{U}_n^*\left(\widehat{\boldsymbol{\theta}}_n(x,h)\right), \quad (3.13)$$

where now the $\boldsymbol{\eta}_i^*$, $i = 1, \ldots n$ are sampled with replacement from the set $\{\boldsymbol{\eta}_1 \ldots, \boldsymbol{\eta}_n\}$.

In both definitions it is necessary to include the bias estimator $\hat{\boldsymbol{B}}_n(x,g,h)$. A similar construction has been used before by, e.g., Härdle and Bowman (1988). We now show that the joint distribution of the local polynomial MSLE is consistently estimated by the distribution of the linear one-step bootstrap estimator (3.12). The proof for bootstrap estimator (3.13) is similar. First we need the following definition.

Definition 3.1 *A function $\mathcal{F}(\cdot;\boldsymbol{\eta})$ is locally uniformly bounded about $\boldsymbol{\eta}_0$ if there exists a neighborhood $\mathcal{N}(\boldsymbol{\eta}_0)$ of $\boldsymbol{\eta}_0$ and a constant M, such that for each $\boldsymbol{\eta}$ in $\mathcal{N}(\boldsymbol{\eta}_0)$ and for all $\boldsymbol{y}$, $|\mathcal{F}(\boldsymbol{y};\boldsymbol{\eta})| \leq M$.*

In what follows, $P^*, \mathcal{D}^*, E^*$ and Var* stand for probability, convergence in distribution, expectation and variance, conditionally on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ and $\mathbf{0}$ is a zero matrix of the appropriate dimension.

Lemma 3.1 *Assume the conditions of Theorem 3.1 and the local uniform boundedness of the second derivatives of the log semilikelihood. Then, as $n \to \infty$,*

$$\boldsymbol{H}^{-1}\boldsymbol{A}_n^*\left(\widehat{\boldsymbol{\theta}}_n(x,h)\right)\boldsymbol{H}^{-1} - (\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x) \xrightarrow{P^*} \mathbf{0} \quad a.s.\ P.$$

Proof. By definition of the bootstrap resampling scheme, we have that $E^*[\boldsymbol{A}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h)] = \boldsymbol{A}_n(\widehat{\boldsymbol{\theta}}_n(x,h))$. Conditions (C2) and (B3) permit the application of Lemmas 3.3 and 3.5 from which it follows that

$$\boldsymbol{H}^{-1}E^*[\boldsymbol{A}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h))]\boldsymbol{H}^{-1} - (\boldsymbol{\Sigma}_x + h\boldsymbol{\Lambda}_x) \to 0 \quad \text{a.s. } P.$$

The local uniform boundedness of the second derivatives of the log semilikelihood is sufficient to show that

$$\text{Var}^*[\boldsymbol{H}^{-1}\boldsymbol{A}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h))\boldsymbol{H}^{-1}] = O((nh)^{-1}) \quad \text{a.s. } P.$$

Lemma 3.2 *Under the conditions of Theorem 3.1, as* $n \to \infty$,

$$nh\mathrm{Var}^*\left(\boldsymbol{H}^{-1}\boldsymbol{U}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h))\right) \to \boldsymbol{\Gamma}_x \quad a.s.\ P.$$

Proof. The definition of the MSLE (3.3) and definition (3.12) imply that $E^*[\boldsymbol{U}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h)] = 0$. Therefore, $\mathrm{Var}^*[\boldsymbol{H}^{-1}\boldsymbol{U}_n^*(\widehat{\boldsymbol{\theta}}(x,h)] =$

$$\frac{1}{n^2}\sum_{i=1}^{n} E^*\left[\boldsymbol{H}^{-1}\eta_i\left(\widehat{\boldsymbol{\theta}}_1(x,h),\widehat{\boldsymbol{\theta}}_2(x,h)\right)\eta_i\left(\widehat{\boldsymbol{\theta}}_1(x,h),\widehat{\boldsymbol{\theta}}_2(x,h)\right)^T\boldsymbol{H}^{-1}\right].$$

Due to conditions (B3) and (C2), the result follows immediately by application of Lemma 3.3 and 3.5.

The results of Lemma 3.1 and 3.2 can be combined to find that

$$\boldsymbol{V}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h)) = nh\boldsymbol{A}_n^{-1}(\widehat{\boldsymbol{\theta}}_n(x,h))\boldsymbol{H}\mathrm{Var}^*(\boldsymbol{U}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h))\boldsymbol{H}\boldsymbol{A}_n^{-1}(\widehat{\boldsymbol{\theta}}_n(x,h))$$

is a strongly consistent estimator for $\boldsymbol{V}(\boldsymbol{\theta}(x))$.

The next theorem states that our bootstrap method is consistent for estimating the distribution of $\widehat{\boldsymbol{\theta}}(x,h)$.

Theorem 3.5 *Assume all conditions of Theorem 3.4, and the local uniform boundedness of the first and second derivatives of the log semilikelihood. Then, almost surely* P,

(i) $\sup_{t\in\mathbb{R}^p}\left|P^*\left(\sqrt{nh}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}^*(x,h)-\widehat{\boldsymbol{\theta}}(x,h))\le t\right)-P\left(\sqrt{nh}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}(x,h)-\boldsymbol{\theta}(x))\le t\right)\right| = o(1).$

(ii) $\sup_{t\in\mathbb{R}^p}\left|P^*\left(\left(\boldsymbol{V}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h))\right)^{-1/2}\sqrt{nh}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}^*(x,h)-\widehat{\boldsymbol{\theta}}(x,h))\le t\right)-P\left((\boldsymbol{V}(\boldsymbol{\theta}(x)))^{-1/2}\sqrt{nh}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}(x,h)-\boldsymbol{\theta}(x))\le t\right)\right| = o(1).$

Proof. We first show that $\sqrt{nh}\boldsymbol{H}^{-1}\boldsymbol{U}_n^*(\widehat{\boldsymbol{\theta}}_n(x,h)) \xrightarrow{\mathcal{D}^*} \mathcal{N}(0,\boldsymbol{\Gamma}_x)$, almost surely P.

Define for $r = 1,2; i = 1,\ldots,n$

$$Z_{ni}^* \;=\; (n/h)^{-1/2}u^T\boldsymbol{H}^{-1}\left\{\boldsymbol{\eta}_i(\widehat{\boldsymbol{\theta}}_1(x,h),\widehat{\boldsymbol{\theta}}_2(x,h))\right\}^*,$$

where u is a $\mathcal{P}$ dimensional vector with $\|u\| = 1$. Since the Z_{ni}^* are conditionally independent, Liapunov's condition is sufficient for the asymptotic normality result

to hold. Now, for n large enough, the local uniform boundedness of first derivatives of the log semilikelihood, gives that, by Lemma 3.2

$$\left(\sum_{i=1}^{n} E^*[|Z_{ni}^*|^{2+\delta}]\right)\left(\sum_{i=1}^{n} E^*[|Z_{ni}^*|^2]\right)^{-(2+\delta)/2} = O\left((nh)^{-\delta/2}\right) \quad \text{a.s.} P.$$

The asymptotic normality result now follows from the Cramér-Wold theorem. The proof can easily be completed by application of Theorems 3.2 and 3.4 and Lemma 3.1.

3.5 Applications

The main application that we consider is the construction of simultaneous confidence regions (on a finite grid). The technique is applied on three datasets.

3.5.1 Simultaneous confidence regions

Theorem 3.5 can be used to construct pointwise confidence regions for $\boldsymbol{\theta}(x)$, but simulation of the bootstrap distribution also allows for the construction of simultaneous confidence regions. For a finite number of grid points $\{z_1, \ldots, z_N\}$, confidence regions for the parameter vector $\boldsymbol{\theta}(x)$ can be constructed without much more computational effort. If there are B repetitions of the bootstrap experiment, define $\widetilde{\boldsymbol{\theta}}_n^*(x,h)^{(t)} = (\widehat{\boldsymbol{\theta}}_n^*(x,h)^{(t)} - \widehat{\boldsymbol{\theta}}_n(x,h))$, where $\widehat{\boldsymbol{\theta}}_n^*(x,h)^{(t)}$ is the estimator resulting from the tth bootstrap run ($t = 1, \ldots, B$). For each grid point z_i and for each parameter $\theta_{j0}(\cdot)$, the estimators $\left\{\widetilde{\theta}_{n,j0}^*(z_i,h)^{(t)}; t = 1, \ldots, B\right\}$ are ordered and their corresponding ranks are denoted by $\{r_{ij}^{(t)}; t = 1, \ldots, B\}$. The t th ordered value is $\widetilde{\theta}_{n,j0}^*(z_i,h)^{[t]}$. Next, define t_k as the k-th order statistic of the set

$$\left\{\max\left(\max_{1\le j\le\kappa}\max_{1\le i\le N}\left(r_{ij}^{(t)}\right); B+1-\min_{1\le j\le\kappa}\min_{1\le i\le N}\left(r_{ij}^{(t)}\right)\right); t = 1, \ldots, B\right\}.$$

Then, by construction, the intervals

$$\left\{\left[\widehat{\theta}_{j0}(z_i,h) - \widetilde{\theta}_{j0}^{*[B+1-t_k]}(z_i,h); \widehat{\theta}_{j0}(z_i,h) - \widetilde{\theta}_{j0}^{*[t_k]}(z_i,h)\right]; i = 1, \ldots, N; j = 1, \ldots, \kappa\right\}$$

have a global confidence level of at least $100(k/B)\%$. Note that also the derivatives of the curves could be included in the same way. This would however lead to very

wide confidence regions because of the conservative nature of this method. For one-parameter generalized linear models, a similar method for constructing confidence intervals has been proposed by Kauermann and Tutz (1996), see also Besag, Green, Higdon and Mengersen (1995).

3.5.2 The low-iron rat teratology data

This rank-based approach is applied to the low-iron rat teratology data to construct a global 80% confidence band for the proportion of death foetuses and the intra-litter correlation as functions of the mother's hemoglobin level. The result for the beta-binomial model is shown in Figure 3.2. Global confidence intervals (dotted line) and pointwise confidence intervals (short dashed line) are based on the one-step linear bootstrap for the proportion of death foetuses and the intra-litter correlation as a function of the hemoglobin level of the mother. The very wide band at the right boundary is probably due to the sparseness of the data in that region. The beta-binomial local linear estimates are shown by the dotted line and the bias corrected estimate by the solid line.

In the proposed estimation technique two unknown bandwidths are involved. For practical applications the bandwidth choice should be data-driven. All curves in Figure 3.1 are based on the same bandwidth $h = 0.319$ (for covariate values transformed on the 0–1 scale). For bias estimation in the bootstrap procedure we used $g = 0.486 = h(\log n)^{0.3}$, where n=58 is the number of data values. This choice for g satisfies the bandwidth condition in (D1). The value for h is based on a cross-validation criterion for the beta-binomial model (see Section 2.6.2). As for every other nonparametric estimator, the selection of the smoothing parameters is not obvious. Optimal bandwidth choice in the context of local semilikelihood estimation seems to be a challenging problem.

3.5.3 The twins data

For the twins data set, we get the following confidence intervals for logit of the probability of mortality and morbidity within 28 days after birth.

To obtain these results, we took $h = 0.184$ and $g = h(\log n)^{0.3} = 0.228$. Simultaneous confidence intervals are represented by the dots and pointwise confidence intervals by the dashed lines. The finite grid of x values is obtained by projecting

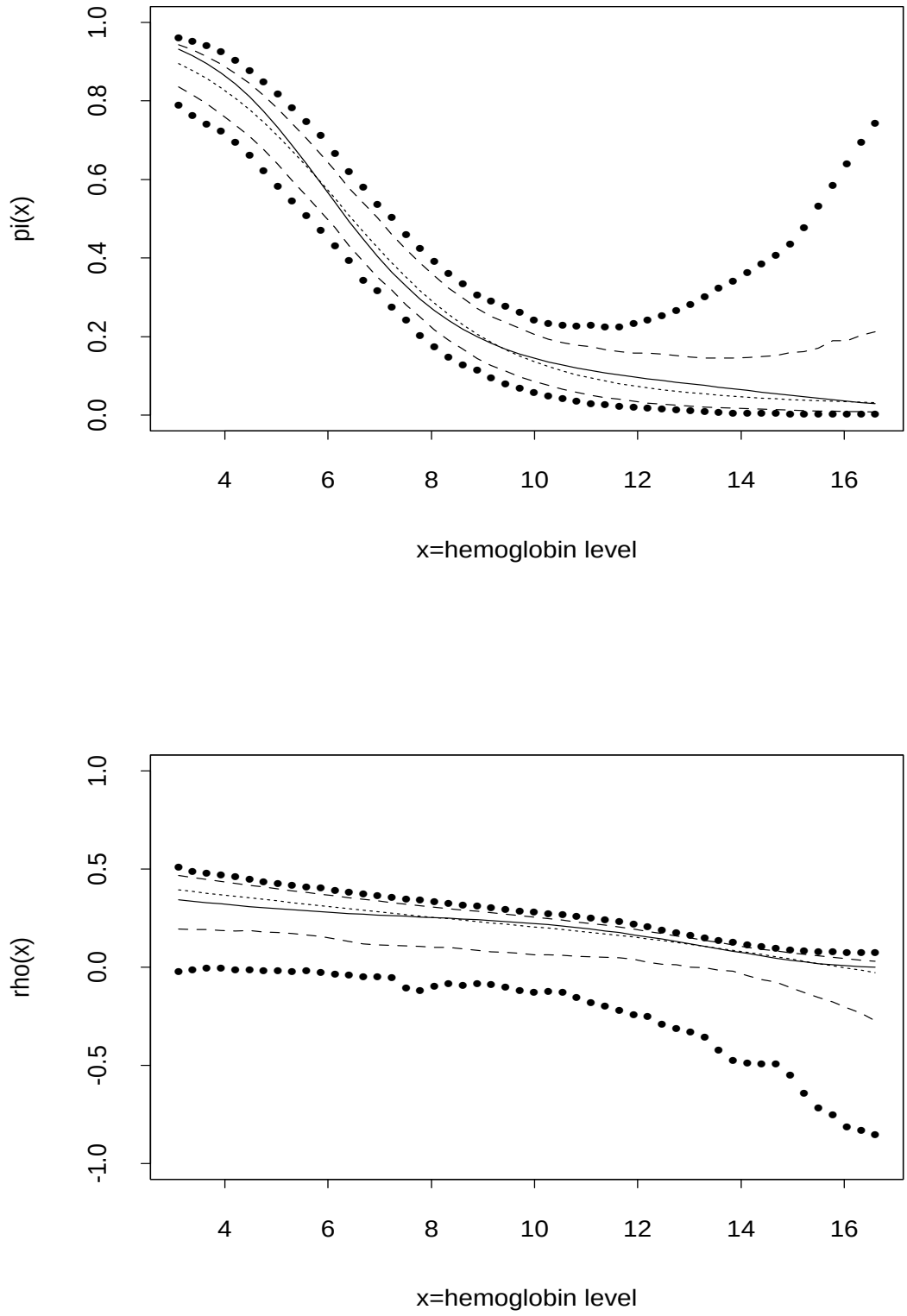


Figure 3.2: Low-iron rat teratology data. Simultaneous and pointwise 80% confidence intervals based on the one-step linear bootstrap.

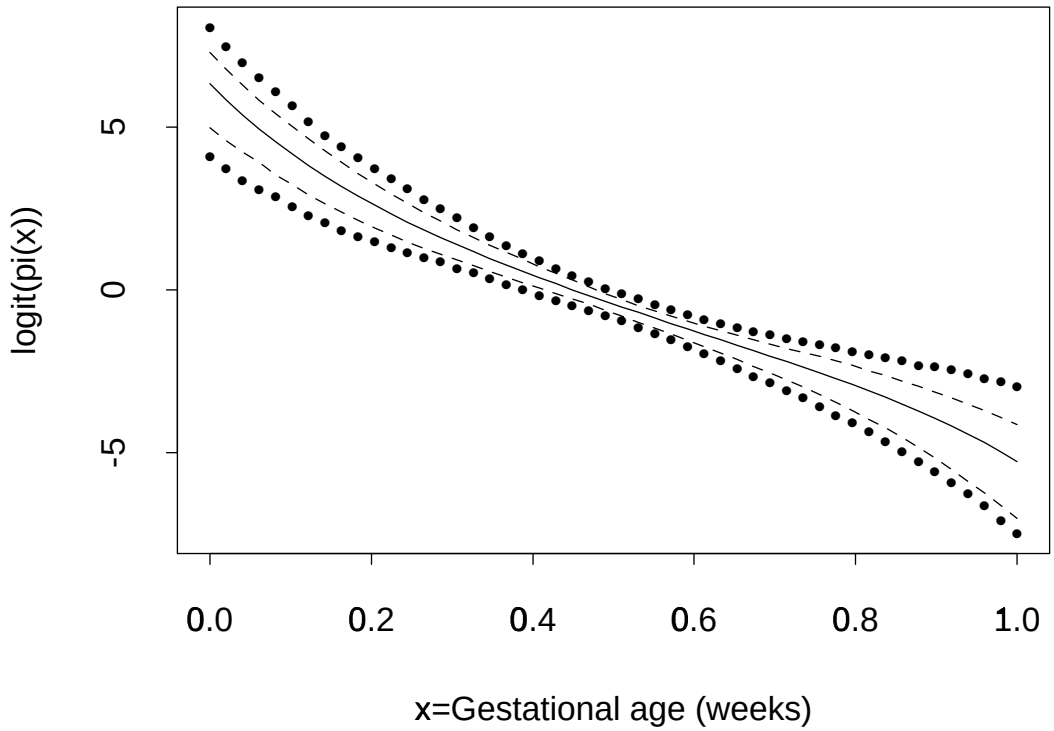


Figure 3.3: Twins data. Simultaneous and pointwise 80% confidence intervals based on the one-step linear bootstrap.

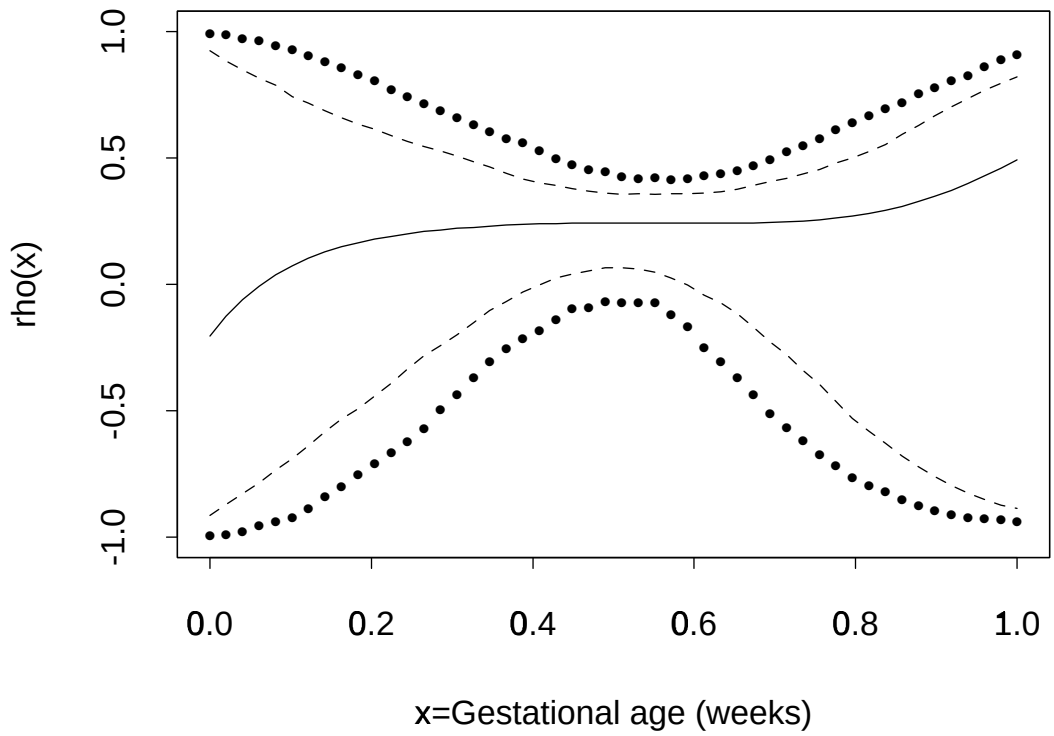


Figure 3.4: Twins data. Simultaneous and pointwise 80% confidence intervals for the within-twin correlation, based on the one-step linear bootstrap.

the dots on the x-axis. The method gives really nice results in this case, since the intervals are rather small.

A bit disappointing though are the confidence intervals for the correlation parameter. The sand-glass looking curves seem to indicate that there is very few information available about the correlation parameter. This is not really unexpected. Since all clusters are of size two (twins), the only possible outcomes are (0,0), (0,1), (1,0) and (1,1), where a one denotes the occurrence of the event of interest. Within a cluster, there are only two outcomes available to estimate the correlation. With clusters of larger size, the estimation is easier since then much more information is available.

The very wide bands are not only a consequence of the estimation method. We

also fitted a parametric linear/constant beta-binomial model. Based on the asymptotic normality, an 80% confidence interval for ρ reaches from 0.017 to 0.463, which is also rather wide. For higher confidence probabilities, also these parametric confidence intervals include negative values. For an accurate estimate of the correlation parameter, it seems that much more data are necessary (this data set contains 113 twins).

3.5.4 The Wisconsin diabetes study

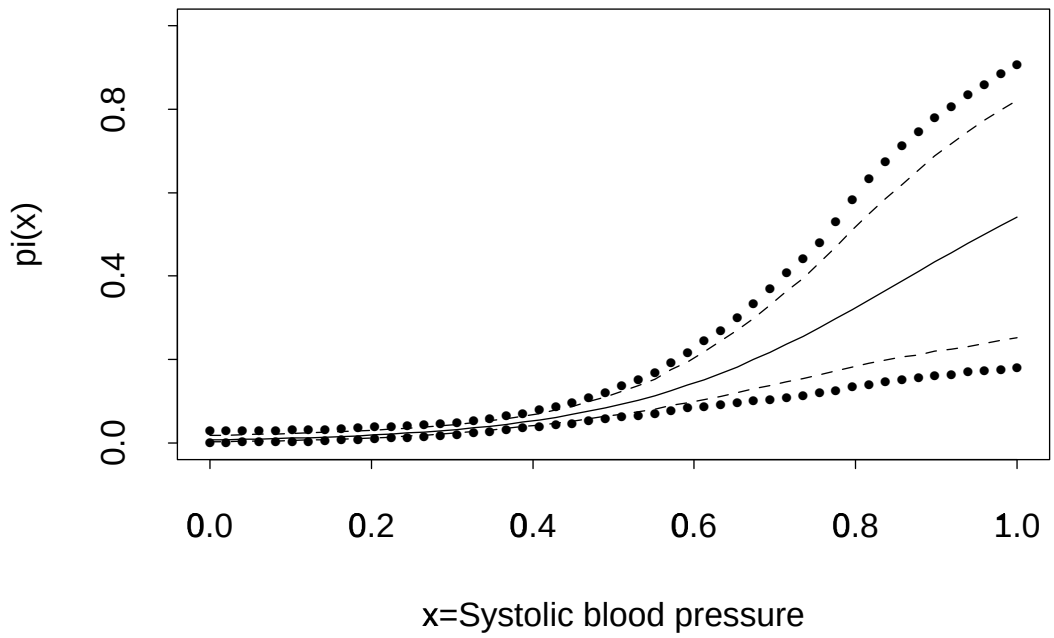
The simultaneous and pointwise 90% confidence intervals for the probability of macular edema in younger onset diabetic persons, are given in Figure 3.5. This data set contains 720 observations on the two eyes. So, also here all clusters are of size two, but there are more than six times as many clusters as compared with the twins data. Notice the effect of the sparseness of the data for high levels of systolic blood pressure. The boundary effect on the probability parameter for small values of systolic blood pressure is a bit masked by showing the graph on probability scale.

That weird behavior of the confidence intervals for the correlation parameter does not show up anymore, see Figure 3.5. In this estimation method we took $h = 0.3$ and $g = 0.3(\log n)^{0.3} = 0.4111$ (both on zero-one scale of the covariate) and 1000 bootstrap simulations.

3.6 Discussion

In the above bootstrap method, everywhere a one-step linear approximation method is being applied. Another interesting topic is the investigation of the quadratic one-step bootstrap estimator. For testing hypothesis in parametric models, we show in Chapters 9 and 10 that this leads to a substantial improvement of the proposed semiparametric bootstrap procedure. This approximation is easily programmed for exponential family models, but the implementation gets quite complicated for, e.g., the beta-binomial distribution.

Other ideas for estimating derivatives exist. Fan, Farmen and Gijbels (1998) use a local polynomial estimator of degree $(p + k)$ to estimate the $(p + k)$th derivative. For this estimation method, a pilot bandwidth is necessary. Next, these estimators are used to obtain approximations of the bias.



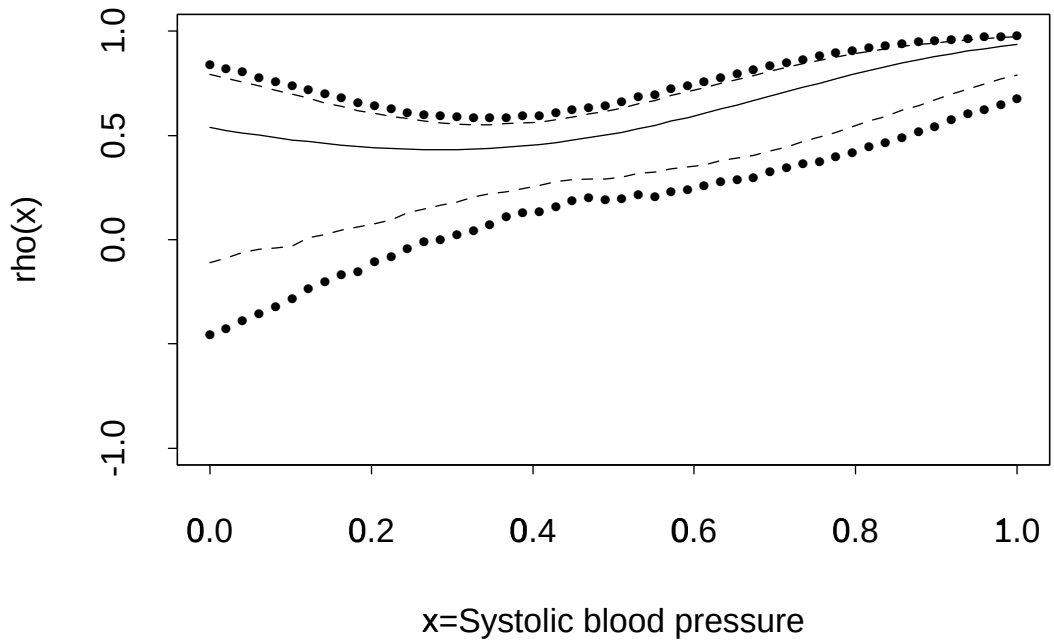


Figure 3.5: Wisconsin diabetes data. Simultaneous and pointwise 90% confidence intervals based on the one-step linear bootstrap.

From these estimated bias and variance expressions *and* by assuming a Gaussian distribution, Fan, Farmen and Gijbels (1998) also obtain pointwise confidence intervals.

3.7 Technical lemmas

Three technical lemmas, which are used frequently in the proofs throughout this chapter are formulated and proven.

Lemma 3.3 *Assume (G), (K), (H') and that the kernel K is $m \geq 0$ times differentiable with $K^{(k)}(-1) = K^{(k)}(1) = 0$ for $0 \leq k \leq m$ and $K^{(m)}$ Lipschitz continuous. Let $S(\cdot)$ be a bounded and Lipschitz continuous function. If the stated derivatives of $S(\cdot)f_X(\cdot)$ exist, then, for the fixed design points (2.1) and for integers $0 \leq k \leq m$, $\ell \geq 0$,*

$$\begin{aligned} &\frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)\left(\frac{x_i - x}{h}\right)^{\ell} S(x_i) \\ &= \quad b(k,\ell)\frac{d^{a(k,\ell)-1}}{dx^{a(k,\ell)-1}}(S(x)f_X(x))\nu_{c(k,\ell)} + o(1) + O(\frac{1}{nh^{a(k,\ell)+1}}) \end{aligned}$$

where

$$a(k,\ell) = \begin{cases} k - \ell + 1 & \ell \leq k \\ 2 & \ell > k, \ell - k \ \ odd \\ 1 & \ell > k, \ell - k \ \ even \end{cases}$$

and $b(k,\ell)$ and $c(k,\ell)$ are some integer constants.

Proof. The proof is based on Lemma 2.1.

Lemma 3.4 *Assume the same conditions on $f_X(x)$, K and h as in Lemma 3.3. Consider $(x_1, V_1), \ldots, (x_n, V_n)$ where $V_1, \ldots, V_n$ are independent mean zero random variables with $E|V_i|^s < M$ for all i in $\mathcal{I}_n = \{i : |x_i - x| \leq h\}$, for some $s > 2$. If, with $a(k,\ell)$ defined as in Lemma 3.3, the bandwidth h satisfies the conditions $nh^{2a(k,\ell)-2\ell-1}/\log n \to \infty$, $nh^{a(k,\ell)+1} \to \infty$, and for some $0 < \eta < 1$, $n^{(s-1)/(s-\eta)}h^{a(k,\ell)-\ell} \to \infty$ and $n^{2/(s-\eta)-1}\log n/h \to 0$, then for integers $0 \leq k \leq m$, $\ell \geq 0$,*

$$\frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)(x_i - x)^{\ell} V_i = O(n^{\frac{1-s}{s-\eta}}h^{\ell-a(k,\ell)}) + O(n^{-\frac{1}{2}}h^{\ell-a(k,\ell)+\frac{1}{2}}\log^{\frac{1}{2}} n)$$

almost surely.

Proof. Define $\tilde{V}_i = V_i I\{|V_i| \leq i^{1/r}\}$ where $r = s - \eta$ and $I\{\cdot\}$ is the indicator function, then

$$\begin{aligned}
\mathcal{V}_n(x) &= \frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)(x_i - x)^\ell V_i \\
&= \left(\mathcal{V}_n(x) - \tilde{\mathcal{V}}_n(x)\right) + \left(\tilde{\mathcal{V}}_n(x) - E(\tilde{\mathcal{V}}_n(x))\right) + \left(E(\tilde{\mathcal{V}}_n(x)) - E(\mathcal{V}_n(x))\right) \\
&= I_n + I\!I_n + I\!I\!I_n,
\end{aligned}$$

where $\tilde{\mathcal{V}}_n(x)$ is defined as $\mathcal{V}_n(x)$ but with V_i replaced by $\tilde{V}_i$. If $E|V_i|^s < M$ for all i in $\mathcal{I}_n$, typical arguments based on the Borel Cantelli lemma (see, e.g., the proof of Lemma 5.2 in Müller and Stadtmüller, 1987) imply that $I_n = O(n^{-1}h^{\ell - a(k,\ell)})$ almost surely. An application of Bernstein's inequality (see, e.g., Serfling, 1980, p. 95), Lemma 3.3 with $S(\cdot) \equiv 1$ and the Borel Cantelli lemma yield for the second term $I\!I_n = O(n^{-\frac{1}{2}}h^{\ell - a(k,\ell) + \frac{1}{2}} \log^{\frac{1}{2}} n)$ almost surely. The third term is bounded by $I\!I\!I_n \leq Cn^{-1}h^{\ell - a(k,\ell)} \sum_{i \in \mathcal{I}_n} E(|V_i|^s) i^{(1-s)/(s-\eta)} = O(n^{(1-s)/(s-\eta)} h^{\ell - a(k,\ell)})$.

The next lemma is an extension of results found in Iverson and Randles (1989).

Lemma 3.5 *Assume all conditions of Lemma 3.4 but now let $V_1(\boldsymbol{\theta}), \ldots, V_n(\boldsymbol{\theta})$ be independent mean zero random variables with $E|V_i(\boldsymbol{\theta})|^s < M_1$ for all i in $\mathcal{I}_n$, for some $s > 2$. Suppose that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ a.s.; and denoting by $\mathcal{B}(d)$ a $\mathcal{P}$ dimensional sphere around the origin with radius d, assume that*

(i) $\lim_{d \to 0} E[\sup_{\mathbf{s} \in \mathcal{B}(d)} |V_i(\boldsymbol{\theta} + \boldsymbol{s}) - V_i(\boldsymbol{\theta})|] = 0$, *uniformly in i in $\mathcal{I}_n$.*

(ii) for some $d > 0$, $E[(\sup_{\mathbf{s} \in \mathcal{B}(d)} |V_i(\boldsymbol{\theta} + \boldsymbol{s}) - V_i(\boldsymbol{\theta})|)^s] < M_2$, for all i in $\mathcal{I}_n$,

then for integers $0 \leq k \leq m$, $\ell \geq 0$, $\frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)(x_i - x)^\ell V_i(\hat{\boldsymbol{\theta}}_n) =$

$$O(n^{\frac{1-s}{s-\eta}} h^{\ell - a(k,\ell)}) + O(n^{-\frac{1}{2}} h^{\ell - a(k,\ell) + \frac{1}{2}} \log^{\frac{1}{2}} n) + o(1)$$

almost surely.

Proof. Define

$$\begin{aligned}
\mathcal{V}_n(\boldsymbol{\theta}) &= \frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)(x_i - x)^\ell V_i(\boldsymbol{\theta}) \\
Q_n(\boldsymbol{s}) &= \frac{1}{nh^{a(k,\ell)}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right)(x_i - x)^\ell \left(V_i(\boldsymbol{\theta} + \boldsymbol{s}) - V_i(\boldsymbol{\theta})\right).
\end{aligned}$$

Then, $\mathcal{V}_n(\widehat{\boldsymbol{\theta}}_n) = Q_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \mathcal{V}_n(\boldsymbol{\theta})$. We show that $Q_n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to 0$, a.s. Let ε and γ be arbitrary positive numbers. By the strong convergence of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}$, there is an integer N_1 such that for all $n \geq N_1$, $P(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \notin \mathcal{B}(d)$, for all n but finitely many $) < \gamma/2$. For all $n \geq N_1$, $P(|Q_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})| > \varepsilon$, for all n but finitely many) $\leq P(\sup_{\mathbf{s}\in\mathcal{B}(d)} |Q_n(\boldsymbol{s})| > \varepsilon$, for all n but finitely many)$+\gamma/2$. Assumption (ii) and the assumptions on the bandwidth allow to apply Lemma 3.4 on $\sup_{\mathbf{s}\in\mathcal{B}(d)} |V_i(\boldsymbol{\theta} + \boldsymbol{s}) - V_i(\boldsymbol{\theta})|$ centered about its mean. By (i) there is an N_2 sufficiently large such that for all $n \geq \max\{N_1, N_2\}$, $P(|Q_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})| > \varepsilon$, all n but finitely many $) \leq \gamma$.

Chapter 4

Local Polynomial Estimation in Multiparameter Additive Models

In this chapter we study several extensions of the local polynomial estimation method. To be more specific, each of the curves $\theta_k(\cdot)$ $(k = 1\ldots,\kappa)$ can be estimated using a smoothing parameter h_k, which might be different for each selected component. For the multiple covariate case, we focus attention to local polynomial estimation in additive models.

4.1 Introduction

The local polynomial estimators, as defined in Chapter 2, use the same smoothing parameter for each of the parameters $\theta_k(\cdot)$, $(k = 1,\ldots,\kappa)$. In statistical models where the role of the parameters can be quite different, e.g., probability of success and correlation in a beta-binomial model, mean and variance in a Gaussian regression model, or location, scale and shape parameters in an extreme value distribution (Davison and Ramesh, 1998), there is no reason to assume all curves to possess exactly the same amount of smoothness. This naturally asks for allowing the possibility to estimate each curve with its own optimal smoothing parameter.

In this chapter we study local polynomial estimators in a random design where all polynomials are of the same degree p, as motivated by discussions in Chapters 2 and 3. Throughout this chapter, results are formulated in the context of local

estimating equations (Carroll, Ruppert and Welsh, 1998). These contain the local (semi)likelihood models (see Chapter 3) as an important example. Other examples of estimating equations are the generalized estimating equations, GEE, of Liang and Zeger (1986), Zeger and Liang (1986) and the equations used in M-estimation, leading to robust estimators (e.g., Huber, 1981).

Preferably, each curve should be estimated in the most optimal way, that is, the estimator should have the same asymptotic properties as if all other components were known, and the estimation scheme is essentially one-dimensional.

Sections 4.2 and 4.3 each provide a specific set of estimating equations for the multiparameter model. An important result in Section 4.2 is that the most natural set of estimating equations for the multiparameter model lacks this optimality property when different bandwidths are used. This problem is solved in Section 4.3 by defining a set of estimating equations which are similar to those defined for estimating components of additive models.

A study of the local polynomial estimators in general additive one-parameter models is presented in Section 4.4, and a generalization to additive multiparameter models is the topic of Section 4.5. From these sections it is clear that, in additive models, backfitting estimators do not satisfy the optimality property mentioned above. We close this chapter with a short discussion on semiparametric models.

Results of this chapter are also presented in Claeskens and Aerts (1999).

4.2 Multiparameter single covariate models I

In this section we consider models with more than one parameter, where each parameter $\theta_k(\cdot)$ is estimated locally by a polynomial of degree p_k. First we assume these functions to depend on a single covariate X, so we observe i.i.d. data $(\boldsymbol{Y}_1, X_1), \ldots, (\boldsymbol{Y}_n, X_n)$; in Section 4.5 we obtain results on the multiple covariate case. To each parameter θ_k $(k = 1, \ldots, \kappa)$ corresponds an estimating equation ψ_k, e.g. the partial derivative of a log likelihood function with respect to this parameter. We study the estimation of the curves in a fixed, but arbitrary interior point x.

As a simple extension of the set of estimating equations for local semilikelihood estimation, we might obtain the estimators as the solutions to the following set of $\sum_{j=1}^{\kappa}(p_j+1)$ estimating equations, where for the estimating function ψ_k a smoothing parameter h_k is used to determine the amount of smoothing as defined by the kernel

weights. For $k = 1, \ldots, \kappa$,

$$\begin{aligned}
&\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \ldots, \sum_{j=0}^{p_\kappa} \beta_{\kappa j}(x)(X_i - x)^j\} K_{h_k}(X_i - x) = 0 \\
&\qquad \vdots \\
&\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \sum_{j=0}^{p_1} \beta_{1j}(x)(X_i - x)^j, \ldots, \sum_{j=0}^{p_\kappa} \beta_{\kappa j}(x)(X_i - x)^j\} (X_i - x)^{p_k} K_{h_k}(X_i - x) = 0
\end{aligned} \tag{4.1}$$

We now list some assumptions.

(B1) There exists an open subset Θ of the parameter space containing the true parameters $\boldsymbol{\theta}(x) = (\theta_1(x), \theta_2(x))$ such that $\psi_k(\boldsymbol{y}; \theta_1, \theta_2)$ admits all second partial derivatives for all $(\theta_1, \theta_2) \in \Theta$, for almost all y and for all $k = 1, \ldots, \kappa$.

(B2) For all $k = 1, \ldots, \kappa$, $E[\psi_k(\boldsymbol{y}; \theta_1(x), \theta_2(x))] = 0$.

(B3) There exist functions $H_i(\cdot), i = 1, 2, 3$ such that $|\psi_r(\boldsymbol{y}; \theta_1, \theta_2)\psi_s(\boldsymbol{y}; \theta_1, \theta_2)| \leq H_1(\boldsymbol{y})$, $|\frac{\partial \psi_r}{\partial \theta_s}(\boldsymbol{y}; \theta_1, \theta_2)| \leq H_2(\boldsymbol{y})$ and $|\frac{\partial^2 \psi_r}{\partial \theta_s \partial \theta_t}(\boldsymbol{y}; \theta_1, \theta_2)| \leq H_3(\boldsymbol{y})$ for all $(\theta_1, \theta_2) \in \Theta$, for all $r, s, t = 1, 2$ and for almost all $\boldsymbol{y}$. The functions H_i are such that $E[H_i^2(\boldsymbol{Y})]$ is uniformly bounded on Θ.

(B4) The matrix elements $J_{rs}(\boldsymbol{\theta}(x)) = E[\frac{\partial \psi_r}{\partial \theta_s}(\boldsymbol{y}; \theta_1(x), \theta_2(x))]$ are Lipschitz continuous and differentiable at $\boldsymbol{\theta}(x)$ and the resulting matrix $\boldsymbol{J}(\boldsymbol{\theta}(x))$ is positive definite.

Assuming the necessary conditions for existence of the estimators to hold, we focus attention to asymptotic bias and variance expressions of the estimators. In parametric estimating equations models, conditions implying existence, consistency and asymptotic normality of the estimators are formulated by Yuan and Jennrich (1998), and for local semilikelihood equations, see Chapter 3.

Since odd degree polynomials have several advantages in comparison with polynomials of even degree (e.g. properties concerning boundary bias) we only present the following results for p_1 and p_2 odd, and both equal to p. Results of other cases can easily be obtained by calculations similar to those of Corollary 2.1.

Theorem 4.1 *Assume conditions (S), (G), (K), (H) and (B1)–(B4) to hold, where the density f_X is the probability density function of the covariate X. For $p_1 = p_2 = p$,*

the conditional bias of $\hat{\beta}_{kj}(x)$, $(j = 0, \ldots, p_1)$ *is for* $p - j$ *odd approximated by*

$$\begin{aligned} & E[\hat{\beta}_{kj}(x) - \beta_{kj}(x)|\tilde{\boldsymbol{X}}] \\ &= \left(\sum_{r=1}^{2} \sum_{\ell=1}^{2} h_\ell^{p+1} \sqrt{\frac{h_\ell}{h_k}} \frac{\theta_r^{(p+1)}(x)}{(p+1)!} J_{\ell r}(\boldsymbol{\theta}(x))[J^{-1}(\boldsymbol{\theta}(x))]_{k\ell} \right) \\ &\quad \times \int_{\mathcal{R}_x} z^{p+1} K_{jp}(z) dz + O(\sum_{r=1}^{2} h_r^{p+2} \sqrt{\frac{h_r}{h_k}}), \end{aligned} \tag{4.2}$$

where $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_n)$.

For each choice of k and j, this expression depends on both parameters $\theta_1(x)$, $\theta_2(x)$, and on both bandwidths h_1 and h_2. From (4.2) it is observed that, to avoid problems, one might want to take both bandwidths of the same order, such that the ratios h_r/h_k are $O(1)$ for all r and k. In this case, for $h_k = c_k n^\gamma$ (for some γ depending on p and j), the selection of optimal constants c_1 and c_2 is difficult because both curves $\theta_1(\cdot)$ and $\theta_2(\cdot)$ appear in bias expression (4.2).

In the next theorem we obtain the conditional variance of the estimators. Define, for an interior point x, $\boldsymbol{\Sigma}_x = f_X(x)\boldsymbol{J}(\boldsymbol{\theta}(x)) \otimes \boldsymbol{N}_p(x)$, where $\boldsymbol{N}_p(x)$ is the $(p+1) \times (p+1)$ matrix with (k,l)th entry equal to $\nu_{k+l-2}(\mathcal{R}_x)$.

$$\boldsymbol{\Lambda}_x = \begin{pmatrix} h_1 \frac{d}{dx}\{f_X(x) J_{11}(\boldsymbol{\theta}(x))\} & h_1 \frac{d}{dx}\{f_Z(x) J_{12}(\boldsymbol{\theta}(x))\} \\ h_2 \frac{d}{dx}\{f_Z(x) J_{21}(\boldsymbol{\theta}(x))\} & h_2 \frac{d}{dx}\{f_X(x) J_{22}(\boldsymbol{\theta}(x))\} \end{pmatrix},$$

and

$$\boldsymbol{\Gamma}_x = f_X(x) \begin{pmatrix} K_{11}(\boldsymbol{\theta}(x)) \boldsymbol{T}_p^{11}(x) & K_{12}(\boldsymbol{\theta}(x)) \boldsymbol{T}_p^{12}(x) \\ K_{21}(\boldsymbol{\theta}(x)) \boldsymbol{T}_p^{21}(x) & K_{22}(\boldsymbol{\theta}(x)) \boldsymbol{T}_p^{22}(x) \end{pmatrix},$$

where $K_{rs}(\theta_1(x), \theta_2(x)) = E[\psi_r\{\boldsymbol{y}; \theta_1(x), \theta_2(x)\}\psi_s\{\boldsymbol{y}; \theta_1(x), \theta_2(x)\}]$ and

$$[\boldsymbol{T}_p^{rs}(x)]_{k\ell} = [\boldsymbol{T}_p^{sr}(x)]_{\ell k} = \int_{\mathcal{R}_x} K(\frac{h_s}{h_r} z) K(z) z^{k+\ell} dz \left(\frac{h_s}{h_r}\right)^{k+1/2}.$$

Note that when $h_1 = h_2$,

$$[\boldsymbol{T}_p^{rs}(x)]_{k\ell} = [\boldsymbol{T}_p(x)]_{k\ell} = \int_{\mathcal{R}_x} K^2(z) z^{k+\ell} dz,$$

for all r, s.

Theorem 4.2 *Under the assumptions of Theorem 4.1, for $p_1 = p_2 = p$, the conditional variance of $\hat{\beta}_{kj}(x)$ is approximated by*

$$\begin{aligned} & Var(\hat{\beta}_{kj}(x)|\tilde{\boldsymbol{X}}) \\ = \; & \frac{f_X^{-1}(x)}{nh_k} \sum_{r=1}^{2} \sum_{s=1}^{2} \left([\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{kr} K_{rs}(\boldsymbol{\theta}(x)) [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))^T]_{sk} \right. \\ & \left. \times \, [\boldsymbol{N}_p^{-1}(x) \boldsymbol{T}_p^{rs}(x) \boldsymbol{N}_p^{-1}(x)]_{jj} \right) + o(\frac{1}{nh_k}). \end{aligned}$$

The complicatedness of this formula is for most part due to the bandwidths which are allowed to be different. In the case of equal bandwidths, this reduces to

$$\frac{f_X^{-1}(x)}{nh_k} [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x)) \boldsymbol{K}(\boldsymbol{\theta}(x)) \boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))^T]_{kk} \int_{\mathcal{R}_x} K_{jp}^2(z) dz + o(\frac{1}{nh_k}). \tag{4.3}$$

The results in the above theorems in fact discourage the use of different bandwidths for different components.

Proof of Theorems 4.1 and 4.2. The proof goes along the same lines as the proof of Corollary 2.1. Define

$$\begin{aligned} (W_n)_{kr} \; = \; & \frac{1}{nh_k} \sum_{i=1}^{n} K \left(\frac{x_i - x}{h_k} \right) \left(\frac{x_i - x}{h_k} \right)^r \\ & \times \psi_k \{ \boldsymbol{Y}_i; \sum_{j=0}^{p} \beta_{1j}(x)(x_i - x)^j, \sum_{j=0}^{p} \beta_{2j}(x)(x_i - x)^j \}, \end{aligned}$$

$$\boldsymbol{W}_n = ((W_n)_{10}, \ldots, (W_n)_{1p}, (W_n)_{20}, \ldots, (W_n)_{2p})^T .$$

The bias expression is obtained by working out the matrix products in

$$E \left[\left(\boldsymbol{\Sigma}_x^{-1} - \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Lambda}_x \boldsymbol{\Sigma}_x^{-1} \right) \boldsymbol{W}_n \right]_{kj} .$$

To simplify the results, use that

$$\frac{|\boldsymbol{M}_{rp}(z;x)|}{|\boldsymbol{N}_p(x)|} = \sum_{j=1}^{p+1} [N_p^{-1}(x)]_{r+1,j} z^{j+1}$$

such that

$$\int_{\mathcal{R}_x} K_{rp}(z) z^{p+1} dz = \sum_{j=1}^{p+1} [N_p^{-1}(x)]_{r+1,j} \nu_{p+j}(\mathcal{R}_x).$$

To obtain the marginal conditional variance of the estimator $\hat{\beta}_{kj}$, we have to select the corresponding component of the matrix $\mathbf{\Sigma}_x^{-1}\mathbf{\Gamma}_x\mathbf{\Sigma}_x^{-1}$. By definition of the kernel $K_{jp}(\cdot)$ we obtain that

$$[\boldsymbol{N}_p^{-1}(x)\boldsymbol{T}_p^{k\ell}(x)\boldsymbol{N}_p^{-1}(x)]_{rs} = \sqrt{\frac{h_\ell}{h_k}}\int_{\mathcal{R}_x} K_{r-1,p}(z)K_{s-1,p}(\frac{h_\ell}{h_k}z)dz.$$

This motivates the simplifiction for $h_1 = h_2$. From the above, the result follows immediately.

A disadvantage of these formulae is that, in general, the asymptotic conditional bias of $\hat{\beta}_{rj}(x)$ is determined by all functions $\theta_k(\cdot)$ $(k = 1, \ldots, \kappa)$. The occurrence of the functions $\theta_k(\cdot)$, $k \neq r$, can be explained by the set of local estimating equations (4.1). By construction, in the set of equations for estimating $\theta_k(x)$ and its derivatives up to order p_k, there is not only a "local polynomial" for this component, but also for each other component. In the next section we propose an alternative set of estimating equations which will yield local polynomial estimators with more appealing asymptotic properties.

Remark 4.1. For some specific sets of estimating functions ψ_k, simplifications can occur. If the "Fisher" information matrix $\boldsymbol{J}(\theta_1(x), \theta_2(x))$ is a diagonal matrix, i.e., all off-diagonal elements are zero, the above bias expression simplifies to:

$$E[\hat{\beta}_{kj}(x) - \beta_{kj}(x)|\tilde{\boldsymbol{X}}] = h_k^{p+1}\frac{\theta_k^{(p+1)}(x)}{(p+1)!}\int_{\mathcal{R}_x} z^{p+1}K_{jp}(z)dz + O(h_k^{p+2}),$$

and the variance to (4.3). Except for condition (H), no other restrictions on the bandwidths h_k need to be imposed. One such example is the set of log-likelihood equations for a Gaussian model.

Remark 4.2. Results also simplify in multi-stage equations where, for example, estimating function ψ_k is a function of $\theta_1, \ldots, \theta_k$, but not of the other θ_ℓ's $(\ell > k)$. This implies that $\boldsymbol{J}(\boldsymbol{\theta}(x))$ is a lower triangular matrix, i.e., $J_{rs}(\boldsymbol{\theta}(x)) = 0$ for all $r < s$. Because of this structure, only terms containing quotients h_s/h_r with $r < s$ will occur in the leading terms of the conditional bias expressions. More explicitly, for a two parameter model,

$$\begin{aligned} &E[\hat{\beta}_{1j}(x) - \beta_{1j}(x)|\tilde{\boldsymbol{X}}] \\ &= \left(h_1^{p+1}\frac{\theta_1^{(p+1)}(x)}{(p+1)!} + h_2^{p+1}\sqrt{\frac{h_2}{h_1}}\left\{\frac{\theta_1^{(p+1)}(x)}{(p+1)!}J_{21}(\boldsymbol{\theta}(x)) + \frac{\theta_2^{(p+1)}(x)}{(p+1)!}J_{22}(\boldsymbol{\theta}(x))\right\}\right. \end{aligned}$$

$$\times[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{12}\Big)\int_{\mathcal{R}_x} z^{p+1}K_{jp}(z)dz + O(h_1^{p+2} + h_2^{p+2}\sqrt{\frac{h_2}{h_1}}),$$

$$\begin{aligned}&E[\hat{\beta}_{2j}(x) - \beta_{2j}(x)|\tilde{\boldsymbol{X}}]\\ &= h_2^{p+1}\left(\frac{\theta_2^{(p+1)}(x)}{(p+1)!} + \frac{\theta_1^{(p+1)}(x)}{(p+1)!}J_{21}(\boldsymbol{\theta}(x))[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{22}\right)\int_{\mathcal{R}_x} z^{p+1}K_{jp}(z)dz + O(h_2^{p+2}).\end{aligned}$$

The function $\theta_1(\cdot)$ appears in both estimating functions ψ_1 and ψ_2 but $\theta_2(\cdot)$ only in ψ_2, a fact which is clearly reflected in the conditional bias expression. Also here, it will be safe to take $h_2/h_1 = O(1)$.

For the conditional variance we obtain,

$$\begin{aligned}&\mathrm{Var}(\hat{\beta}_{1j}(x)|\tilde{\boldsymbol{X}})\\ &= \frac{f_X^{-1}(x)}{nh_1}[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{11}\left\{K_{11}(\boldsymbol{\theta}(x))[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))^T]_{11}[\boldsymbol{N}_p^{-1}(x)\boldsymbol{T}_p^{11}(x)\boldsymbol{N}_p^{-1}(x)]_{jj}\right.\\ &\quad \left. + K_{21}(\boldsymbol{\theta}(x))[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{12}[\boldsymbol{N}_p^{-1}(x)\boldsymbol{T}_p^{21}(x)\boldsymbol{N}_p^{-1}(x)]_{jj}\right\} + o(\frac{1}{nh_1}),\end{aligned}$$

$$\begin{aligned}&\mathrm{Var}(\hat{\beta}_{2j}(x)|\tilde{\boldsymbol{X}})\\ &= \frac{f_X^{-1}(x)}{nh_2}[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{22}\left\{K_{21}(\boldsymbol{\theta}(x))[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))^T]_{12}[\boldsymbol{N}_p^{-1}(x)\boldsymbol{T}_p^{21}(x)\boldsymbol{N}_p^{-1}(x)]_{jj}\right.\\ &\quad \left. + K_{22}(\boldsymbol{\theta}(x))[\boldsymbol{J}^{-1}(\boldsymbol{\theta}(x))]_{22}[\boldsymbol{N}_p^{-1}(x)\boldsymbol{T}_p^{22}(x)\boldsymbol{N}_p^{-1}(x)]_{jj}\right\} + o(\frac{1}{nh_2}).\end{aligned}$$

Due to the compactness of the support of the kernel K, we do not need further restrictions on the choice of the bandwidths. An example of two-stage equations are the GEE2 equations (see, e.g., Zhao and Prentice, 1990) where the first estimating equation yields an estimator for the mean response and where the second estimating equation is used to obtain an estimator of the correlation structure of the multivariate response vector, given an estimator of the mean. By iterating between these two equations until convergence, estimators for both mean and correlation structure are obtained.

4.3 Multiparameter single covariate models II

In this section we propose an alternative set of local estimating equations for multiparameter models. For $k = 1, \ldots, \kappa$, obtain the estimators by solving the following

set of equations:

$$\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \beta_{10}(X_i), \ldots, \beta_{k-1,0}(X_i), \sum_{j=0}^{p_k} \beta_{kj}(x)(X_i - x)^j, \beta_{k+1,0}(X_i), \ldots, \beta_{\kappa 0}(X_i)\} \times K_{h_k}(X_i - x) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} \psi_k\{\boldsymbol{Y}_i; \beta_{10}(X_i), \ldots, \beta_{k-1,0}(X_i), \sum_{j=0}^{p_k} \beta_{kj}(x)(X_i - x)^j, \beta_{k+1,0}(X_i), \ldots, \beta_{\kappa 0}(X_i)\} \times (X_i - x)^{p_k} K_{h_k}(X_i - x) = 0 \tag{4.4}$$

An important difference with equations (4.1) is that we now have to obtain the estimators in *all* observations $X_1, \ldots, X_n$. In order to obtain the estimators in the data values, we need to solve a total of $n \times \sum_{k=1}^{\kappa}(p_k + 1)$ equations, by choosing x to be one of these values X_i, $i = 1, \ldots, n$. If estimators in a specific value x (not belonging to $\{X_1, \ldots, X_n\}$) are to be obtained, the above set of $\sum_{k=1}^{\kappa}(p_k + 1)$ equations in (4.4) needs to be added to the previous ones.

Although one might expect the solutions to the sets of equations (4.1) and (4.4) to be very similar, their asymptotic conditional bias and variance expressions are different. Defining the equations in this way, asymptotic properties of each of the estimators $\hat{\beta}_{kj}$ are the same as if those estimators would have been obtained by solving a one-parameter equation only.

Theorem 4.3 *For $p_1 = p_2 = p$ the asymptotic bias of $\hat{\beta}_{kj}(x)$ is for $p - j$ odd approximated by*

$$E[\hat{\beta}_{kj}(X_i) - \beta_{kj}(X_i)|\tilde{\boldsymbol{X}}] = h_k^{p+1} \frac{\theta_k^{(p+1)}(X_i)}{(p+1)!} \int_{\mathcal{R}_{X_i,h_k}} z^{p+1} K_{jp}(z) dz + O_P(h_k^{p+2}),$$

and the conditional variance by

$$Var(\hat{\beta}_{kj}(X_i)) = \frac{f_X^{-1}(X_i)}{nh_k} [\boldsymbol{J}^{-1}(\boldsymbol{\theta}(X_i)) \boldsymbol{K}(\boldsymbol{\theta}(X_i)) \boldsymbol{J}^{-1}(\boldsymbol{\theta}(X_i))^T]_{kk} \int_{\mathcal{R}_{X_i,h_k}} K_{jp}^2(z) dz + o_P(\frac{1}{nh_k}).$$

The simple formulae in the above theorem follow from the fact that, for each $k = 1, \ldots \kappa$, partial derivatives of the estimating equation ψ_k with respect to each of the intercept terms $\beta_{j0}(x_i)$ for $j \neq k$ and for all $i = 1, \ldots, n$ is $O[1/(nh_k)]$.

The estimating equations (4.4) are an interesting alternative to those proposed in Chapter 2.

It would be very interesting to see how the estimators obtained by (4.1) and (4.4) compare in practical applications. One expects the estimators obtained by (4.1) to be subject to more bias than the estimators obtained by solving (4.4). In an extreme case of very different curvature, the influence of one curve on the other should be more pronounced in equations (4.1).

Also bandwidth selection in the multiparameter models is an interesting challenge. The asymptotic bias and variance expressions in Theorem 4.3 have an important consequence on the selection of the optimal bandwidth. Since asymptotic mean squared error (AMSE) for the multiparameter model is simply a sum of the AMSE for each component, Theorem 4.3 implies that any of the bandwidth selectors for one-parameter models can be applied in this multiparameter setting. Alternatively, one can consider a multi-stage method which assumes the bandwidths for all but one curve to be fixed, and proceeds by selecting this one bandwidth, after which the procedure is repeated for each of the other components.

4.4 One-parameter additive models

In this section we now consider the multiple covariate case. To explain the idea of the estimation method we first consider models with only one parameter $\theta(\cdot)$.

The data we observe are $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ where $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Di})$, for a univariate response variable Y_i. In an additive model with $D \geq 1$ covariates, the unknown parameter of interest has the following form

$$\theta(x_1, \ldots, x_D) = \theta_1(x_1) + \ldots + \theta_D(x_D), \tag{4.5}$$

where, for example in a generalized linear model, $\theta(\boldsymbol{x}) = g(E[Y|\boldsymbol{X} = \boldsymbol{x}])$ for a known link function g. Additive and generalized additive models are nowadays very popular and often used for statistical modeling and estimation. An extensive account in the context of one-parameter exponential family models is given in Hastie and Tibshirani (1990).

If there is more than one covariate, that is, if $D > 1$, there is a problem of identifiability, which is typical for this kind of additive models. Since each of the functions $\theta_d(\cdot)$ is locally replaced by a polynomial of degree p_d, we obtain D intercept

terms. This calls for $D-1$ constraints on the estimators. For a random design, what one usually does is including an intercept term α in the model and assuming that each of the expected values $E[\theta_d(X_d)]$ is zero. This is also the approach that we take. More details are given in Section 4.4.2.

Local polynomial estimators of degree p_k of the curves $\theta_k(\cdot)$ at the point $\boldsymbol{x} = (x_1, \ldots, x_D)$, and of its derivatives up to order p_k, for $(k = 1, \ldots D)$, can be obtained by solving the following set of estimating equations for the coefficients $\beta_{dj}(\cdot)$:

$$\boldsymbol{\psi}(\boldsymbol{X}_1) = 0, \ldots, \boldsymbol{\psi}(\boldsymbol{X}_n) = 0$$

where $\boldsymbol{\psi}(\boldsymbol{x}) =$

$$\begin{cases} \sum_{i=1}^{n} \psi\{Y_i; \alpha + \sum_{j=0}^{p_d} \beta_{dj}(x_d)(X_{di} - x_d)^j + \sum_{k \neq d} \beta_{k0}(X_{ki})\} K_{h_d}(X_{di} - x_d) \\ \qquad \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{for } d = 1, \ldots D) \\ \sum_{i=1}^{n} \psi\{Y_i; \alpha + \sum_{j=0}^{p_d} \beta_{dj}(x_d)(X_{di} - x_d)^j + \sum_{k \neq d} \beta_{k0}(X_{ki})\} (X_{di} - x_d)^{p_d} K_{h_d}(X_{di} - x_d) \ . \end{cases} \tag{4.6}$$

The intercept term α is estimated by solving $\sum_{i=1}^{n} \psi(y_i; \alpha) = 0$.

If the likelihood of the data is known, we can define

$$\psi(Y_i; \theta) = (\partial/\partial\theta) \ln f(Y_i; \theta),$$

in which case, for $D = 1$, the above equations are the local p_dth degree log likelihood equations (Chapter 2). In case this likelihood belongs to a one-parameter exponential family, the resulting estimators are the local quasi-likelihood estimators (Fan, Heckman and Wand, 1995).

Although these equations resemble the set of equations (4.4), they are structurally very different. In the set (4.6) there is only one global parameter $\theta(\cdot)$, and hence a one-dimensional estimating function $\psi(\cdot)$ (see also Section 4.4.1), while in the set of equations (4.4) there are κ global parameters $\theta_1(\cdot), \ldots, \theta_\kappa(\cdot)$ and also κ estimating functions $\psi_1(\cdot), \ldots, \psi_\kappa(\cdot)$.

Another important difference with the multiparameter estimating equations (4.4) is that in the single covariate multiparameter models there are no constraints on the functions $\theta_k(\cdot)$, while in the multiple covariate additive models these constraints are really necessary.

4.4.1 Definitions and notations

The local polynomial design matrices of the model are defined as

$$\mathbb{X}_{dx} = \begin{pmatrix} 1 & (X_{d1} - x_d) & \cdots & (X_{d1} - x_d)^{p_d} \\ \vdots & \vdots & & \vdots \\ 1 & (X_{dn} - x_d) & \cdots & (X_{dn} - x_d)^{p_d} \end{pmatrix}.$$

Let $\boldsymbol{W}_{dx} = \mathrm{diag}_{1 \le i \le n}\{K_{h_d}(X_{di} - x_d)\}$ be a diagonal matrix containing the kernel weights at the point $\boldsymbol{x}$, and let the vector $\boldsymbol{q}_{1d}(\boldsymbol{x}, \boldsymbol{\beta})$, $d = 1, \ldots, D$ represent the set of (yet unweighted) estimating functions,

$$\boldsymbol{q}_{1d}(\boldsymbol{x}, \boldsymbol{\beta}) = \begin{pmatrix} \psi\{Y_1, \alpha + \boldsymbol{e}_1^T \mathbb{X}_{dx} \boldsymbol{\beta}_d + \sum_{k \neq d} \beta_{k0}(X_{k1})\} \\ \vdots \\ \psi\{Y_n, \alpha + \boldsymbol{e}_n^T \mathbb{X}_{dx} \boldsymbol{\beta}_d + \sum_{k \neq d} \beta_{k0}(X_{kn})\} \end{pmatrix},$$

where $\boldsymbol{e}_j$ is a vector of the appropriate length with a one at the jth entry and zeroes elsewhere and, for a particular choice of $\boldsymbol{x}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1x}^T, \ldots, \boldsymbol{\beta}_{Dx}^T)^T$, where $\boldsymbol{\beta}_{dx} = (\beta_{d0}(x_d), \ldots, \beta_{dp_d}(x_d))^T$.

With these notations, we may write $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{U}(\boldsymbol{x}, \boldsymbol{\beta})$ where

$$\boldsymbol{U}(\boldsymbol{x}, \boldsymbol{\beta}) \equiv \begin{pmatrix} \boldsymbol{U}_1(\boldsymbol{x}, \boldsymbol{\beta}) \\ \vdots \\ \boldsymbol{U}_D(\boldsymbol{x}, \boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} \mathbb{X}_{1x}^T \boldsymbol{W}_{1x} \boldsymbol{q}_{11}(\boldsymbol{x}, \boldsymbol{\beta}) \\ \vdots \\ \mathbb{X}_{Dx}^T \boldsymbol{W}_{Dx} \boldsymbol{q}_{1D}(\boldsymbol{x}, \boldsymbol{\beta}) \end{pmatrix}. \tag{4.7}$$

Denote $\psi'(y, \theta) = \frac{\partial}{\partial \theta} \psi(y, \theta)$,

$$\boldsymbol{q}_{2d}(\boldsymbol{x}, \boldsymbol{\beta}) = \begin{pmatrix} \psi'\{Y_1, \alpha + \boldsymbol{e}_1^T \mathbb{X}_{dx} \boldsymbol{\beta}_d + \sum_{k \neq d} \beta_{k0}(X_{k1})\} \\ \vdots \\ \psi'\{Y_n, \alpha + \boldsymbol{e}_n^T \mathbb{X}_{dx} \boldsymbol{\beta}_d + \sum_{k \neq d} \beta_{k0}(X_{kn})\} \end{pmatrix},$$

$\boldsymbol{P}_d(\boldsymbol{x}, \boldsymbol{\beta}) = -\mathrm{diag}\{\boldsymbol{q}_{2d}(\boldsymbol{x}, \boldsymbol{\beta})\}$, its conditional expected value is defined as $\boldsymbol{Q}_d(\boldsymbol{x}, \boldsymbol{\beta}) = -\mathrm{diag}\{E[\boldsymbol{q}_2(\boldsymbol{x}, \boldsymbol{\beta}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]\}$, and the matrix of first partial derivatives of the set of estimating equations $\boldsymbol{J}_d(\boldsymbol{x}, \boldsymbol{\beta}) = \mathbb{X}_{dx}^T \boldsymbol{W}_{dx} \boldsymbol{P}_d(\boldsymbol{x}, \boldsymbol{\beta}) \mathbb{X}_{dx}$.

To obtain expressions for the asymptotic conditional bias and variance of the estimators $\hat{\beta}_{kj}(X_{ki})$ it suffices to concentrate on a first order approximation of the equations $\boldsymbol{U}(\boldsymbol{x}, \boldsymbol{\beta}) = 0$, *if* second partial derivatives of the estimating equations

(e.g. third partial derivatives of the log likelihood) are uniformly bounded. Under this assumption, equation (4.7) can, by a one-step Taylor expansion, be approximated by (all other terms are of lower order),

$$\widehat{\boldsymbol{\beta}}_d(\boldsymbol{x}) \approx \boldsymbol{\beta}_d + \boldsymbol{J}_d^{-1}(\boldsymbol{x},\boldsymbol{\beta})\boldsymbol{U}_d(\boldsymbol{x},\boldsymbol{\beta}) = \boldsymbol{\beta}_d + (\mathbb{X}_{dx}^T \boldsymbol{W}_{dx}\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})\mathbb{X}_{dx})^{-1}\mathbb{X}_{dx}\boldsymbol{W}_{dx}\boldsymbol{q}_{1d}(\boldsymbol{x},\boldsymbol{\beta}),$$

or, equivalently,

$$\widehat{\boldsymbol{\beta}}_d(\boldsymbol{x}) \approx (\mathbb{X}_{dx}^T \boldsymbol{W}_{dx}\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})\mathbb{X}_{dx})^{-1}\mathbb{X}_{dx}\boldsymbol{W}_{dx}\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})\tilde{\boldsymbol{Y}}_{d\beta}$$

where

$$\tilde{\boldsymbol{Y}}_{d\beta}(\boldsymbol{x}) = \mathbb{X}_{dx}\boldsymbol{\beta} + \boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})^{-1}\boldsymbol{q}_{1d}(\boldsymbol{x},\boldsymbol{\beta})$$

is the *adjusted dependent variable* (refer to, e.g., Hastie and Tibshirani, 1990, p. 138, see also Remark 4.8). Alternatively, we could have used $\boldsymbol{Q}_d(\boldsymbol{x},\boldsymbol{\beta})$ instead of $\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})$. It is well-known that

$$j!\boldsymbol{e}_j^T\widehat{\boldsymbol{\beta}}_d = j!\widehat{\beta}_{dj} = \widehat{\theta}_d^{(j)}(x).$$

4.4.2 The iteratively reweighted backfitting algorithm

In "classical" additive models, estimators are most often obtained via a backfitting algorithm. For likelihood models, Hastie and Tibshirani (1990, p.149) motivate the use of an iterative backfitting scheme. In a one-parameter likelihood model they obtain the following expression:

$$\theta_d(X_d) = (E[P_d(\boldsymbol{x},\boldsymbol{\beta})|X_d])^{-1}\, E[P_d(\boldsymbol{x},\boldsymbol{\beta})\{\tilde{Y}_{d\beta} - \alpha - \sum_{k\neq d}\theta_k(X_k)\}|X_d],$$

where $P_d(\boldsymbol{x},\boldsymbol{\beta})$ and $\tilde{Y}_{d\beta}$ are the single observation likelihood versions of $\boldsymbol{P}_d(x,\boldsymbol{\beta})$ and $\tilde{\boldsymbol{Y}}_{d\beta}$, respectively, at the "true" parameter value. This is the basis of the backfitting algorithm. To explain how it proceeds, we can rely on the typical notations used in classical additive models. For $\boldsymbol{x}$ in the set $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$, define for $j = 1,\ldots,p_d+1$,

$$\boldsymbol{s}_{jd}^T(\boldsymbol{x}) = \boldsymbol{e}_j^T\left(\mathbb{X}_{dx}^T\boldsymbol{W}_{dx}\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})\mathbb{X}_{dx}\right)^{-1}\mathbb{X}_{dx}^T\boldsymbol{W}_{dx}\boldsymbol{P}_d(\boldsymbol{x},\boldsymbol{\beta})$$

and $\boldsymbol{S}_{jd}$ the matrix with rows $(\boldsymbol{s}_{jd}^T(\boldsymbol{X}_1),\ldots,\boldsymbol{s}_{jd}^T(\boldsymbol{X}_n))$. If $j = 1$, this subscript will be omitted. To include the condition of zero means of each of the marginal functions, the estimators are being centered, by subtracting from each estimator the

mean of all fitted values (i.e. $\sum_{i=1}^{n} \widehat{\theta}_d(\boldsymbol{X}_i)/n$). In matrix notation, this centering is calculated by the matrix

$$\boldsymbol{S}_d^* = (\boldsymbol{I} - \mathbf{1}_n \cdot \mathbf{1}_n^T/n)\boldsymbol{S}_d,$$

where $\mathbf{1}_n$ is an $(n \times 1)$ vector of ones. The backfitting scheme now looks as follows, for $\widehat{\boldsymbol{\theta}}_k = (\widehat{\theta}_k(\boldsymbol{X}_1), \ldots, \widehat{\theta}_k(\boldsymbol{X}_n))^T$ iterate between the following equations:

$$\begin{cases} \widehat{\boldsymbol{\theta}}_1 = \boldsymbol{S}_1^*(\check{\boldsymbol{Y}}_{1\beta} - \widehat{\boldsymbol{\alpha}} - \sum_{k\neq 1}^{D} \widehat{\boldsymbol{\theta}}_k) \\ \qquad \vdots \\ \widehat{\boldsymbol{\theta}}_D = \boldsymbol{S}_D^*(\check{\boldsymbol{Y}}_{D\beta} - \widehat{\boldsymbol{\alpha}} - \sum_{k\neq D}^{D} \widehat{\boldsymbol{\theta}}_k). \end{cases}$$

Or equivalently,

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_1 \\ \widehat{\boldsymbol{\theta}}_2 \\ \vdots \\ \widehat{\boldsymbol{\theta}}_D \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{S}_1^* & \ldots & \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* & \boldsymbol{I} & \ldots & \boldsymbol{S}_2^* \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_D^* & \boldsymbol{S}_D^* & \ldots & \boldsymbol{I} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{S}_1^* \\ \boldsymbol{S}_2^* \\ \vdots \\ \boldsymbol{S}_D^* \end{pmatrix} (\check{\boldsymbol{Y}}_\beta - \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\alpha}$ is obtained by maximizing the unweighted equations $\sum_{i=1}^{n} \psi\{\boldsymbol{Y}_i; \alpha\} = 0$. Note how this estimation scheme differs from the corresponding one in classical regression models by the "smoother" matrices $\boldsymbol{S}_d$ who depend on the unknown coefficients via the matrix $\boldsymbol{P}_d(\cdot, \boldsymbol{\beta})$ of partial derivatives of the estimating equations. Also the vector of adjusted dependent variables $\check{\boldsymbol{Y}}_{d\beta}$ depends on the coefficient vector $\boldsymbol{\beta}$.

To obtain a solution of the full set of equations, we have to apply the above backfitting scheme iteratively, resulting in the following algorithm:

Iteratively reweighted backfitting algorithm:

Step (1): Initialize all unknown parameters: $\alpha, \boldsymbol{\beta}$.
Step (2): Backfitting algorithm until convergence.
Step (3): Update $\boldsymbol{P}_d(\cdot, \boldsymbol{\beta})$ and $\check{\boldsymbol{Y}}_{d\beta}$, for $d = 1, \ldots, D$.
Step (4): Repeat Steps (2) and (3) until convergence

In the next section we obtain asymptotic bias and variance expressions of the resulting estimators.

4.4.3 Asymptotic properties

A novel aspect of the results in this section is that the asymptotic properties of estimators are obtained outside the framework of the classical, homoscedastic regression model. While the idea of obtaining the estimators via backfitting in a one-parameter, fully additive likelihood model has already been formulated by Hastie and Tibshirani (1990), the statistical properties are not yet known. In order to obtain our results, we rely on the important piece of work as provided by Opsomer and Ruppert (1997a).

We first focus on the bivariate case; later we will indicate how the case $D > 2$ can be handled. For theoretical purposes, we define

$$\boldsymbol{q}_1 = (\psi\{Y_1, \theta_1(X_{11}) + \ldots + \theta_D(X_{D1})\}, \ldots, \psi\{Y_n, \theta_1(X_{1n}) + \ldots + \theta_D(X_{Dn})\})^T .$$

$\tilde{\boldsymbol{X}}_j = (X_{j1}, \ldots, X_{jn})^T$, $\tilde{\boldsymbol{\theta}}_j = (\theta_j(X_{j1}), \ldots, \theta_j(X_{jn}))^T$, $\boldsymbol{q}_2 = \frac{\partial}{\partial\theta}\boldsymbol{q}_1$, $\boldsymbol{P} = -\operatorname{diag}(\boldsymbol{q}_2)$, $\boldsymbol{Q} = \operatorname{diag}\{E[-\boldsymbol{q}_2|\boldsymbol{X}_1, \boldsymbol{X}_2]\} = \operatorname{diag}_{1\leq i\leq n} J(\{X_{1i}, X_{2i}\})$, and the *adjusted dependent variable* is defined as $\tilde{\boldsymbol{Y}} = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 + \boldsymbol{P}^{-1}\boldsymbol{q}_1$. The presence of the functions $\theta_1(\cdot)$ and $\theta_2(\cdot)$ is practically taken care of by the iteratively reweighted backfitting algorithm as described earlier.

Under the usual regularity conditions it is easily obtained that, conditional on the covariates, the differences in asymptotic bias and variance expressions will be asymptotically negligible when these definitions come into force, instead of the previous ones. We immediately obtain that, for $r = 0, \ldots, p_d$, and for all d,

$$\boldsymbol{e}_{r+1}^T \frac{1}{n} \mathbb{X}_{dx}^T \boldsymbol{W}_{dz} \boldsymbol{P}_d(\boldsymbol{x}, \boldsymbol{\beta})(\tilde{\boldsymbol{Y}}_\beta - \tilde{\boldsymbol{Y}}) = o_P(\frac{h_d^r}{\sqrt{nh_d}}) + O_P(h_d^{p_d+r+1}) = o_P(\frac{h_d^r}{\sqrt{nh_d}}).$$

For $D = 2$ more explicit expressions of the estimators can be obtained. In the bivariate case, we can obtain the estimators by iterating between the following equations:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_1 &= \{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_1^*\boldsymbol{S}_2^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_1^*)\}\tilde{\boldsymbol{Y}}_{1\beta} \\ \hat{\boldsymbol{\theta}}_2 &= \{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{S}_2^*\boldsymbol{S}_1^*)^{-1}(\boldsymbol{I} - \boldsymbol{S}_2^*)\}\tilde{\boldsymbol{Y}}_{2\beta} \end{aligned} \tag{4.8}$$

The results to follow are valid under the following set of assumptions. We refer to Opsomer and Ruppert (1997a) for a similar set of assumptions in the homoscedastic regression model. Let p_k denote the degree of the local polynomial to estimate

curve $\theta_k(\cdot)$. The jth moment of the kernel function K is defined as $\nu_j(z,h_k) = \int_{\mathcal{R}_{z,h_k}} u^j K(u)du$, where the integration region is given by,

$$\mathcal{R}_{z,h_k} = \{t : (z + h_k t) \in \text{supp}(f_k)\} \cap \text{supp}(K),$$

where f_k is the marginal probability density function of covariate Z_k. This region will indicate the difference between interior and boundary points, the former are identified when $\mathcal{R}_{z,h_k} = \text{supp}(K)$, in which case the notational dependence of ν_j on z and h_k is omitted. Define the "Fisher" information $J(\{u,v\}) = E[\frac{\partial \psi}{\partial \theta}(y; \theta_1(u) + \theta_2(v))]$,

$$\begin{aligned}
J_1(z_1) &= \textstyle\int J(\{z_1,v\}) f_2(v)dv \; ; & L_2(z_2) &= \textstyle\int f_{1|2}(u|z_2) J(\{u,z_2\}) J_1(u)^{-1} du \\
J_2(z_2) &= \textstyle\int J(\{u,z_2\}) f_1(u)du \; ; & \tilde{L}_1(z_1) &= \textstyle\int f_{2|1}(v|z_1) J(z_1,v) J_1(z_1)^{-1} dv \\
L^* &= \textstyle\int f(u,v) J(\{u,v\}) J_1(u)^{-1} du dv \; ; & L(z_1,z_2) &= L_2(z_2) + \tilde{L}_1(z_1) - L^*
\end{aligned}$$

(A1) The kernel K is continuous, with support [-1, 1] and bounded variation on [-1, 1]. All odd moments of the kernel are zero and for p_k odd, $\nu_{p_k+1} \neq 0$.

(A2) The marginal design densities f_k and the joint design density f all are continuous functions on a compact support. Within their respective supports, the marginal design densities are strictly positive. The following condition holds:

$$\sup_{x,u,z} \left| \frac{J(\{x,z\})}{J_2(z)} \left\{ \frac{f(u,z) J(\{u,z\})}{f_1(u) J_1(u) f_2(z)} - L(u,z) \right\} \right| < 1.$$

(A3) For all $k = 1, \ldots, D$, as $n \to \infty, h_k \to 0$, and $n h_k / \log n \to \infty$.

(A4) The estimating function $\psi(y,\theta)$ is differentiable with respect to θ. The first component of each vector $\boldsymbol{q}_{2d}$ has a bounded second moment, conditional on all covariates except X_d. The following inequality holds for all u, for some $C > 0$ and for $\alpha > 2/(1+\delta)$ where $h_d = cn^\gamma$ for some $c > 0$ and $-1 < \gamma < 0$,

$$P(|q_{2d}(Y; \theta_1(X_1) + \theta_2(X_2)| \geq u) \leq C u^{-\alpha}.$$

(A5) The Fisher information $J(\{u,v\})$ is strictly positive and bounded over the support of f.

Lemma 4.1 *Under the above assumptions, the following asymptotic approximations hold uniformly over all elements of the matrices:*

$$\begin{aligned}[\boldsymbol{S}_k^*]_{ij} &= [\boldsymbol{S}_k]_{ij} - \frac{1}{n}\frac{J(\{Z_{1j}, Z_{2j}\})}{J_k(Z_{kj})} + o(\frac{1}{n}) \quad a.s.,\\ \boldsymbol{S}_1^*\boldsymbol{S}_2^* &= \boldsymbol{T}_{12}^* + o(\frac{\mathbf{1}\mathbf{1}^T}{n}) \quad a.s.,\end{aligned}$$

where $\boldsymbol{T}_{12}^*$ is a matrix whose (i,j)th element is given by

$$[\boldsymbol{T}_{12}^*]_{ij} = \frac{1}{n}\frac{J(\{Z_{1j}, Z_{2j}\})f(Z_{1i}, Z_{2j})J(\{Z_{1i}, Z_{2j}\})}{f_1(Z_{1i})J_1(Z_{1i})f_2(Z_{2j})J_2(Z_{2j})} - \frac{1}{n}\frac{J(\{Z_{1j}, Z_{2j}\})}{J_2(Z_{2j})}L(Z_{1i}, Z_{2j}).$$

In the proof of this lemma we use a slightly modified version of Proposition 1 of Schulman and Ruppert (1998) which we now formulate.

Lemma 4.2 *Let $g(\boldsymbol{x}, z)$ be a measurable function on $\chi \times \mathbb{R}$ where χ is a measurable subset of $\mathbb{R}^k$ and let $h = cn^\gamma$ for some $c > 0$ and $-1 < \gamma < 0$, such that $P(|g(\boldsymbol{x}, z)| \geq u) \leq Cu^{-\alpha}$ for all u some $C > 0$ and $\alpha > 2/(1+\gamma)$. Also assume that $E[g(\boldsymbol{X}, Z)^2|\boldsymbol{X} = \cdot]$ is bounded and that $(\boldsymbol{X}_1, Z_1), \ldots, (\boldsymbol{X}_n, Z_n)$ are i.i.d. pairs of random variables. Then, with probability 1,*

$$\sup_{n,z}\sqrt{\frac{nh}{\log n}}\left\{\frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{Z_i - z}{h}\right)\right\}^r\left(\frac{Z_i - z}{h}\right)^s g(\boldsymbol{X}_i, Z_i) - E\left[\frac{1}{h}\left\{K\left(\frac{Z_1 - z}{h}\right)\right\}^r\left(\frac{Z_1 - z}{h}\right)^s g(\boldsymbol{X}_1, Z_1)\right]\right\} < \infty$$

for each $r = 1, 2, \ldots$ and $s = 0, 1, \ldots$. A similar result holds when each of these functions is replaced by its absolute value.

Before we give the proof of Lemma 4.1, let us give some additional notations. Define $\boldsymbol{H}_k = \mathrm{diag}_{0\leq j\leq p_k}\{h^j\}$, and let $\boldsymbol{N}_{p_k}(x)$ be a matrix with (ℓ_1+1, ℓ_2+1)th entry equal to $\nu_{\ell_1+\ell_2}(x, h_k)$.

Proof of Lemma 4.1. We shall give the proof only for $k = 1$, the other case is completely analogous. Since, by calculations analogous to those in the proof of Theorem 2.1, we obtain that

$$\frac{1}{n}\mathbb{X}_{1x}^T\boldsymbol{W}_{1x}\boldsymbol{P}\mathbb{X}_{1x} = J_1(x_1)f_1(x_1)\boldsymbol{H}_1\boldsymbol{N}_{p_1}(x_1)\boldsymbol{H}_1 + o_P(\boldsymbol{H}_1\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{H}_1).$$

By writing the matrix product explicitly, we also have that

$$\begin{aligned}
[\boldsymbol{S}_1]_{ij} &= \boldsymbol{e}_1^T \left(\frac{1}{n} \mathbb{X}_{1x}^T \boldsymbol{W}_{1x} \boldsymbol{P} \mathbb{X}_{1,x} \right)^{-1} \frac{1}{n} \mathbb{X}_{1x}^T \boldsymbol{W}_{1x} \boldsymbol{P} \\
&= \sum_{\ell=0}^{p_1} \frac{1}{n h_1^{\ell+1}} \frac{J(\{X_{1j}, X_{2j}\})}{f_1(X_{1i}) J_1(X_{1i})} [N_{p_1}(X_{1i})^{-1}]_{1,\ell+1} (X_{1j} - X_{1i})^\ell K \left(\frac{X_{1i} - X_{1j}}{h_1} \right) \\
&\quad \times (1 + o_P(1)),
\end{aligned}$$

and by taking the centering matrices into account,

$$\begin{aligned}
[\boldsymbol{S}_1^*]_{ij} &= [\boldsymbol{S}_1]_{ij} - \frac{1}{n} \sum_{k=1}^{n} [\boldsymbol{S}_1]_{kj} \\
&= [\boldsymbol{S}_1]_{ij} - \sum_{k=1}^{n} \sum_{\ell=0}^{p_1} \frac{1}{n h_1^{\ell+1}} \frac{J(\{X_{1j}, X_{2j}\})}{f_1(X_{1i}) J_1(X_{1i})} [N_{p_1}(X_{1i})^{-1}]_{1,\ell+1} (X_{1j} - X_{1i})^\ell \\
&\quad \times K \left(\frac{X_{1i} - X_{1j}}{h_1} \right) (1 + o_P(1)).
\end{aligned}$$

By definition of the inverse of a matrix, this can be simplified to

$$[\boldsymbol{S}_1^*]_{ij} = [\boldsymbol{S}_1]_{ij} - \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\})}{J_1(X_{1i})} + o_P(\frac{1}{n}).$$

Similarly,

$$[\boldsymbol{S}_1 \boldsymbol{S}_2]_{ij} = [\boldsymbol{T}_{12}]_{ij} = \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\}) f(X_{1i}, X_{2j}) J(\{X_{1i}, X_{2j}\})}{f_1(X_{1i}) J_1(X_{1i}) f_2(X_{2j}) J_2(X_{2j})} + o_P(\frac{1}{n})$$

and

$$[(\mathbf{1}_n \mathbf{1}_n^T \boldsymbol{S}_1 / n) \boldsymbol{S}_2]_{ij} = \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\})}{J_2(X_{2j})} L(X_{2j}) (1 + o_P(1))$$

$$[\boldsymbol{S}_1 (\mathbf{1}_n \mathbf{1}_n^T \boldsymbol{S}_2 / n)]_{ij} = \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\})}{J_2(X_{2j})} \tilde{L}(X_{1i}) (1 + o_P(1))$$

$$[(\mathbf{1}_n \mathbf{1}_n^T \boldsymbol{S}_1 / n)(\mathbf{1}_n \mathbf{1}_n^T \boldsymbol{S}_2 / n)]_{ij} = \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\})}{J_2(X_{2j})} L^* \cdot (1 + o_P(1)),$$

from which it follows that

$$[\boldsymbol{S}_1^* \boldsymbol{S}_2^*]_{ij} = [\boldsymbol{T}_{12}]_{ij} - \frac{1}{n} \frac{J(\{X_{1j}, X_{2j}\})}{J_2(X_{2j})} L(X_{1i}, X_{2j}) + o_P(\frac{1}{n}).$$

To show that these results hold uniformly for all indices i, j with remainder terms $o(1)$, apply Lemma 4.2.

Lemma 3.2 of Opsomer and Ruppert (1997a) holds under the previous set of conditions. That lemma guarantees the existence of the necessary inverse matrices in (4.9). Note that condition (A2), under Gaussianity, simplifies to

$$\sup_{x_1,x_2}\left|\frac{f(x_1,x_2)}{f_1(x_1)f_2(x_2)}-1\right|<1$$

and hence, is a condition on the design of the covariates. In our context, also the first partial derivatives of the estimating function should be taken into account (e.g. the curvature of a log likelihood function). The presence of this function is the most important difference between the Gaussian equations (least squares) and this more general structure.

In order to obtain the asymptotic bias and variance expressions we need one further condition:

(A6) The function $\theta_d(x)$ has a (p_d+1)th ((p_d+2)nd) derivative for p_d odd (p_d even), $d=1,\ldots,D$, which is continuous and bounded.

Define $\boldsymbol{t}_i$ as the ith row of $(\boldsymbol{I}-\boldsymbol{T}_{12}^*)^{-1}$ and $\boldsymbol{v}_i$ as the ith row of $(\boldsymbol{I}-\boldsymbol{T}_{21}^*)^{-1}$. To simplify notation, denote K_{0p_d} as K_{p_d}. Let $R_j(K)=\int u^jK^2(u)du$,

$$v(z_1,z_2)=J^{-1}(\{z_1,z_2\})E[\psi^2(Y,\theta_1(Z_1)+\theta_2(Z_2))|Z_1=z_1,Z_2=z_2)]J^{-1}(\{z_1,z_2\}),$$

and

$$\mathcal{D}^{p+1}\boldsymbol{\theta}_d=\left(\frac{\partial^{p+1}\theta_d(X_{d1})}{\partial x_d^{p+1}},\ldots,\frac{\partial^{p+1}\theta_d(X_{dn})}{\partial x_d^{p+1}}\right)^T.$$

Theorem 4.4 *If conditions (A1)–(A6) hold, the conditional bias of the estimators can be approximated by:*

$$E[\hat{\alpha}-\alpha|\tilde{\boldsymbol{X}}_1,\tilde{\boldsymbol{X}}_2]=O_P(\frac{1}{\sqrt{n}}),$$

and for p_1 and p_2 both odd,

$$\begin{aligned}E[\hat{\theta}_1(X_{1i})-\theta_1(X_{1i})|\tilde{\boldsymbol{X}}_1,\tilde{\boldsymbol{X}}_2]&=\frac{h_1^{p_1+1}}{(p_1+1)!}\nu_{p_1+1}(\boldsymbol{X}_{1i},h_1)\theta_1^{(p_1+1)}(X_{1i})\\&+\frac{h_1^{p_1+1}}{(p_1+1)!}\nu_{p_1+1}\left((\boldsymbol{t}_i^T-\boldsymbol{e}_i^t)\mathcal{D}^{p_1+1}\boldsymbol{\theta}_1-E[\theta_1^{(p_1+1)}(X_{1i})]\right)\\&-\frac{h_2^{p_2+1}}{(p_2+1)!}\nu_{p_2+1}\left((\boldsymbol{t}_i^TE[\theta_2^{(p_2+1)}(X_{2i})J(\{X_{1i},X_{2i}\})|\tilde{\boldsymbol{X}}_1]J_1^{-1}(X_{1i})-\right.\\&\left.E[\theta_2^{(p_2+1)}(X_{2i})\frac{J(\{X_{1i},X_{2i}\})}{J_1(Z_{1i})}]\right)+O_P(\frac{1}{\sqrt{n}})+o_P(h_1^{p_1+1}+h_2^{p_2+1}),\end{aligned}$$

a similar expression is obtained for $E[\hat{\theta}_2(X_{2i}) - \theta_2(X_{2i})|\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2]$, *such that*

$$
\begin{aligned}
E[\hat{\theta}(X_{1i}, X_{1i}) - \theta(X_{1i}, X_{2i})|\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2] &= \frac{h_1^{p_1+1}}{(p_1+1)!}\Big\{\nu_{p_1+1}(\boldsymbol{X}_{1i}, h_1)\theta_1^{(p_1+1)}(X_{1i}) \\
&+\nu_{p_1+1}\Big((\boldsymbol{t}_i^T - \boldsymbol{e}_i^t)\mathcal{D}^{p_1+1}\boldsymbol{\theta}_1 - \boldsymbol{v}_i^T E[\theta_1^{(p_1+1)}(X_{1i})J(\{X_{1i}, X_{2i}\})|\tilde{\boldsymbol{X}}_2]J_2^{-1}(X_{2i}) \\
&+ E[\theta_1^{(p_1+1)}(X_{1i})\frac{J(\{X_{1i}, X_{2i}\})}{J_2(X_{2i})}] - E[\theta_1^{(p_1+1)}(X_{1i})]\Big)\Big\} \\
&+\frac{h_2^{p_2+1}}{(p_2+1)!}\Big\{\nu_{p_2+1}(\boldsymbol{X}_{2i}, h_2)\theta_2^{(p_2+1)}(X_{2i}) \\
&+\nu_{p_2+1}\Big((\boldsymbol{v}_i^T - \boldsymbol{e}_i^t)\mathcal{D}^{p_2+1}\boldsymbol{\theta}_2 - \boldsymbol{t}_i^T E[\theta_2^{(p_2+1)}(X_{2i})J(\{X_{1i}, X_{2i}\})|\tilde{\boldsymbol{X}}_1]J_1^{-1}(X_{1i}) \\
&+ E[\theta_2^{(p_2+1)}(X_{1i})\frac{J(\{X_{1i}, X_{2i}\})}{J_1(X_{1i})}] - E[\theta_2^{(p_2+1)}(X_{2i})]\Big)\Big\} + o_P(h_1^{p_1+1} + h_2^{p_2+1}).
\end{aligned}
$$

For the conditional variances we obtain the following expressions:

$$
Var(\hat{\alpha}) = \frac{1}{n^2}\sum_{i=1}^{n} v(X_{1i}, X_{2i}),
$$

$$
Var(\hat{\theta}_1(X_{1i})|\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2) = \frac{1}{nh_1}R_0(K_{p_1}, X_{1i})f_1^{-1}(X_{1i})v(X_{1i}, X_{2i})\frac{E[J^2(\{X_{1i}, X_{2i}\})|X_{1i}]}{J_1(X_{1i})^2} + o_P(\frac{1}{nh_1})
$$

and similarly for $Var(\hat{\theta}_1(X_{1i})|\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2)$, *such that*

$$
\begin{aligned}
Var(\hat{\theta}(X_{1i}, X_{2i})|\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2) &= v(X_{1i}, X_{2i})\Big\{\frac{1}{nh_1}R(K_{p_1}, X_{1i})f_1^{-1}(X_{1i}) \\
&\times\frac{E[J^2(\{X_{1i}, X_{2i}\})|X_{1i}]}{J_1(X_{1i})^2} + \frac{1}{nh_2}R_0(K_{p_2}, X_{2i})f_2^{-1}(X_{2i})\frac{E[J^2(\{X_{1i}, X_{2i}\})|X_{2i}]}{J_2(X_{2i})^2}\Big\} \\
&+o_P(\frac{1}{nh_1} + \frac{1}{nh_2})
\end{aligned}
$$

By using approximations similar to those used in the proof of Lemma 4.1, the proof of this theorem can be obtained along the same lines as the proof of Theorem 4.1 in Opsomer and Ruppert (1997a), details are therefore omitted. All results obtained above reduce to those of the classical additive model under Gaussianity. This theorem clearly demonstrates the undesirable property that the bias of the estimators depends on all curves in the model. Note also that in general there is no simplification when $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ are independent, because of the function J.

A possible solution for the bias problem could be to define a multi-step procedure where for estimation of, say, $\theta_1(x)$, a too small bandwidth is used for estimation of the other components, such that their contribution to the bias will be, at least asymptotically, negligible. In the next steps, the same procedure is repeated for each of the other components of the additive model. It is yet unknown how this procedure will perform in practical situations.

Other approaches for estimating additive models have recently been developed and studied. The "non-oracle" efficiency of backfitting estimators has also been noticed by Linton and Nielsen (1995), who defined an alternative estimation scheme, based on a marginal integration approaches. Properties of this type of estimator are also studied by Fan, Härdle and Mammen (1998). This estimator, however, is not fully efficient in mean squared error sense. There are indications that an approach which combines marginal integration and backfitting is fully efficient in mean squared error sense, see, e.g., Linton (1997) and Kim, Linton and Hengartner (1998). Linton (1998) studies estimation in generalized additive models.

Remark 4.3. If p_1 and p_2 are both even, a little bit more effort is necessary to arrive at an approximate bias expression which is, similar to the results of Theorem 4.4, an extension of the bias expression obtained by Opsomer and Ruppert (1997a). For a mixture of odd and even degrees of the polynomial, it can be guessed what the asymptotic bias expressions will be.

Remark 4.4. For the case $D > 2$ we refer to Opsomer (1999) where the classical additive model is studied. In our situation, we would start from the same equations, but replace $\boldsymbol{Y}$ by $\tilde{\boldsymbol{Y}}$, and in the rest of the calculations, incorporate the derivatives of the estimating equations in the appropriate way, as is illustrated in the above results. Results for additive models with more than two covariates are obtained by a recursive formula which allows to write the smoother matrix for the D-dimensional model as a function of the smoother matrix for a $(D-1)$-dimensional submodel and the smoother matrix of a one-dimensional submodel. A similar decomposition is obtained for penalized regression splines, see Theorem 5.1 of Chapter 5.

Remark 4.5. If an estimator of the curve $\theta(\cdot)$ in a point $\boldsymbol{x}$, not being one of the observed data points, is needed, this can be obtained by adding an additional set of estimating equations:

$$\hat{\theta}_1(x_1) \quad = \quad \boldsymbol{s}_1^T(\boldsymbol{x})(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\theta}}_1)$$

$$\vdots$$

$$\hat{\theta}_D(x_D) \;=\; \boldsymbol{s}_D^T(\boldsymbol{x})(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\theta}}_D),$$

where $\boldsymbol{\alpha}$ is a vector of length n where each element is equal to α. An important remark is that when using this estimation scheme the estimators in all data points are really necessary, one cannot just obtain an estimator in an arbitrary value $\boldsymbol{x}$ without having to calculate estimates in all data values.

Remark 4.6. Although there were no stringent restrictions on the orders of the bandwidth parameters, it turns out that by taking the "optimal bandwidths", which minimize the asymptotic mean averaged squared error (AMASE), all bandwidths are of the same order, see Opsomer and Ruppert (1998, p.607). The same phenomenon is observed in this more general setting.

Remark 4.7. Note that we here focused attention to conditional bias and variance expressions for the estimators of the curves $\theta_d(\cdot)$. Also the derivatives of these curves, up to order p_d can be studied. In this case we use the smoother matrices $\boldsymbol{S}_{jd}$ instead of $\boldsymbol{S}_d$, see Opsomer and Ruppert (1998) or Opsomer (1995, Section 3.4) for more details in classical homoscedastic regression models.

Remark 4.8. We now would like to give an intuitive (heuristic) motivation of the definition of the adjusted dependent variable. Using a local estimation method implies working with a saturated model, where each observation has its own parameter, this in contrast to a parametric model where there are only a few global parameters. In a classical regression model where the response can be written in the additive error structure: $Y = \theta(x_i) + \varepsilon_i$, note that, under saturation, $\theta(x_i) = \theta_i$. Solving the set of normal equations for this model implies that

$$\hat{\theta}_i = \theta_i + (Y_i - \theta_i) = \theta(x_i) + \varepsilon_i = Y_i.$$

For an equivalent formulation of this equation in the more general likelihood context where the log likelihood contribution of the i th observation is $\log f(Y_i, \theta_i)$, we have to put a step backwards to see where the normal (least squares) equations are coming from. For Gaussian data it is easily shown that the least squares estimators are exactly the same as the maximum likelihood (ML) estimators. Let us now consider a Taylor expansion of the log likelihood equations at the ML estimator about the value $(\theta_1, \ldots, \theta_n)$.

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \log f(Y_i, \hat{\theta}_i) = 0 \;=\; \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \log f(Y_i, \theta_i) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta_i^2} \log f(Y_i, \theta_i)(\hat{\theta}_i - \theta_i)$$

$$+\frac{1}{2}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta_i^3}\log f(Y_i,\xi_i)(\hat{\theta}_i-\theta_i)^2 \tag{4.9}$$

where ξ_i is on the line segment between $\hat{\theta}_i$ and θ_i. Most proofs of consistency and asymptotic normality of ML estimators start from an equation such as (4.9). See, e.g., Serfling (1980) for a proof for parametric estimators and Chapter 2 for a proof in the local likelihood context. Next, by assuming the third derivatives of the log likelihood to be bounded, these proofs proceed by showing that asymptotic bias, variance and the asymptotic distribution are completely determined by the first two terms at the right of equation (4.9). This implies that the formerly mentioned asymptotic properties of the ML estimators $\hat{\theta}_i$ are the same as those of

$$\theta_i-\left(\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_i^2}\log f(Y_i,\theta_i)\right)^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_i}\log f(Y_i,\theta_i),$$

which in turn are the same as those of

$$\tilde{\theta}_i=\theta_i+I^{-1}(\theta_i)\sum_{i=1}^{n}\frac{\partial}{\partial\theta_i}\log f(Y_i,\theta_i),$$

where $I(\theta_i)=E[-\frac{\partial^2}{\partial\theta_i^2}\log f(Y_i,\theta_i)]$. The smoothing step in classical regression models is performed by multiplying the vector of response values $(Y_1,\ldots,Y_n)$, which is equal to the vector $(\hat{\theta}_1,\ldots,\hat{\theta}_n)$, with a smoothing matrix. Next, properties of the resulting estimator are studied. This gives a motivation to study in a likelihood context the properties of local estimators which are obtained by smoothing the vector

$$\tilde{\boldsymbol{Y}}=(\tilde{\theta}_1,\ldots,\tilde{\theta}_n)^T,$$

which is commonly called the adjusted dependent variable in iteratively reweighted least squares regression models. In the local additive models, this vector will be used in conjuction with a backfitting algorithm, resulting in iteratively reweighted backfitting estimators.

The main conclusion from this section is that backfitting estimators do not work optimally for the purpose they are defined: estimating components of additive models. It should be stressed that this is not only true for additive models having structure (4.5), but for *any* additive structure, e.g., $\theta(x_1,\ldots,x_D)=\theta_1(x_1,\ldots,x_k)+\theta_2(x_{k+1},\ldots,x_D)$.

4.5 Multiparameter additive models

In this section we combine the results from the section on multiparameter models with the results of the previous section on additive models, which leads us to studying multiparameter additive models.

For simplicity of notation, assume that we have a two-parameter model $\kappa = 2$, where each parameter is an additive function of the covariates:

$$\theta_k(\boldsymbol{x}_k) = \theta_{k1}(x_{k1}) + \ldots + \theta_{kd_k}(x_{kD_k})$$

where $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kD_k})$, $k = 1, \ldots, \kappa$. Note that it is *not* necessary that the components of $\boldsymbol{x}_1$ be different from components of $\boldsymbol{x}_k$, $k \neq 1$. Since each parameter is an additive function of the covariates, we here have the same identifiability problem as before in Section 4.4. Therefore we assume that the expectation of each of these component functions $\theta_{kd}(X_{kd})$ equals zero, and include an intercept term, such that, for the ith observation, the contribution to the global, unweighted estimating function for the kth parameter at the true parameter values is given by

$$\psi_k\{\boldsymbol{Y}_i; \alpha_1 + \theta_{11}(X_{11i}) + \ldots + \theta_{1D_1}(X_{1D_1i}), \alpha_2 + \theta_{21}(X_{21i}) + \ldots, \theta_{2D_2}(X_{2D_2i})\}.$$

In the definition of the set of local estimating equations for the multiparameter additive model, we take a similar approach as in equations (4.4) and (4.6). Define the following sets of equations: $\boldsymbol{\psi}_1(\boldsymbol{x}_1) =$

$$\sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \alpha_1 + \sum_{j=0}^{p_{11}} \beta_{11j}(x_{11})(X_{11i} - x_{11})^j + \beta_{120}(X_{12i}) + \ldots + \beta_{1D_10}(X_{1D_1i}),$$

$$\alpha_2 + \beta_{210}(X_{21i}) + \ldots + \beta_{2D_20}(X_{2D_2i})\} K\left(\frac{X_{11i} - x_{11}}{h_1}\right) \begin{pmatrix} 1 \\ \vdots \\ (X_{11i} - x_{11})^{p_{11}} \end{pmatrix}$$

$$\vdots$$

$$\sum_{i=1}^{n} \psi_1\{\boldsymbol{Y}_i; \alpha_1 + \beta_{110}(X_{11i}) + \ldots + \beta_{1,D_1-1,0}(X_{1,D_1-1,i})$$

$$+ \sum_{j=0}^{p_{1D_1}} \beta_{1D_1j}(x_{1D_1})(X_{1D_1i} - x_{1D_1})^j, \alpha_2 + \beta_{210}(X_{21i}) + \ldots + \beta_{2D_20}(X_{2D_2i})\}$$

$$\times K\left(\frac{X_{1D_1i}-x_{1D_1}}{h_1}\right)\begin{pmatrix}1\\ \vdots\\ (X_{1D_1i}-x_{1D_1})^{p_{1D_1}}\end{pmatrix}.$$

And similarly, $\boldsymbol{\psi}_2(\boldsymbol{x}_2) =$

$$\sum_{i=1}^{n}\psi_2\{\boldsymbol{Y}_i;\alpha_1+\beta_{110}(X_{11i})+\ldots+\beta_{1D_10}(X_{1D_1i}),\alpha_2+\sum_{j=0}^{p_{21}}\beta_{21j}(x_{21})(X_{21i}-x_{21})^j+$$

$$\beta_{220}(X_{22i})+\ldots+\beta_{2D_20}(X_{2D_2i})\}K\left(\frac{X_{21i}-x_{21}}{h_2}\right)\begin{pmatrix}1\\ \vdots\\ (X_{21i}-x_{21})^{p_{21}}\end{pmatrix}$$

$$\vdots$$

$$\sum_{i=1}^{n}\psi_2\{\boldsymbol{Y}_i;\alpha_1+\beta_{110}(X_{11i})+\ldots+\beta_{1D_10}(X_{1D_1i}),\ \alpha_2+\beta_{210}(X_{21i})+\ldots$$

$$+\beta_{2,D_2-1,0}(X_{2,D_2-1,i})+\sum_{j=0}^{p_{2D_2}}\beta_{2D_2j}(x_{2D_2})(X_{2D_2i}-x_{2D_2})^j)\}$$

$$\times K\left(\frac{X_{2D_2i}-x_{2D_2}}{h_2}\right)\begin{pmatrix}1\\ \vdots\\ (X_{2D_2i}-x_{2D_2})^{p_{2D_2}}\end{pmatrix}.$$

The estimators of the coefficients $\beta_{kdj}(\cdot)$ at the data values are obtained by solving the set of equations

$$\boldsymbol{\psi}_1(\boldsymbol{X}_{11})=\ldots=\boldsymbol{\psi}_1(\boldsymbol{X}_{1n})=0=\boldsymbol{\psi}_2(\boldsymbol{X}_{21})=\ldots=\boldsymbol{\psi}_2(\boldsymbol{X}_{2n}). \tag{4.10}$$

Note that this is a collection of

$$\{n\times\sum_{k=1}^{\kappa}\sum_{d=1}^{D_k}(p_{kd}+1)\}$$

equations. The constant terms α_1 and α_2 are estimated via the parametric estimating equations

$$\sum_{i=1}^{n}\psi_1(\boldsymbol{Y}_i;\alpha_1,\alpha_2)=0\ ,\qquad\sum_{i=1}^{n}\psi_2(\boldsymbol{Y}_i;\alpha_1,\alpha_2)=0.$$

The local polynomial design matrices of the multiparameter additive model are defined similarly as in a one-parameter additive model, i.e., for $k = 1, \ldots, \kappa$ and $d = 1, \ldots, D_k$,

$$\mathbb{X}_{kdx} = \begin{pmatrix} 1 & (X_{kd1} - x_{kd}) & \cdots & (X_{kd1} - x_{kd})^{p_{kd}} \\ \vdots & \vdots & & \vdots \\ 1 & (X_{kdn} - x_{kd}) & \cdots & (X_{kdn} - x_{kd})^{p_{kd}} \end{pmatrix}.$$

Define $\boldsymbol{W}_{kdx} = \text{diag}_{1 \le i \le n}\{K_{h_{kd}}(X_{kdi} - x_{kd})\}$ and let $\boldsymbol{q}_{1kd}$ represent the set of unweighted estimating functions corresponding to $\theta_{kd}(\cdot)$, at the true parameter values, e.g. for $k = 1$ we have

$$\boldsymbol{q}_{1kd} = \begin{pmatrix} \psi\{\boldsymbol{Y}_1, \alpha_1 + \sum_{d=1}^{D_1} \theta_{1d}(X_{1d1}), \alpha_2 + \sum_{d=1}^{D_2} \theta_{2d}(X_{2d1})\} \\ \vdots \\ \psi\{\boldsymbol{Y}_n, \alpha_1 + \sum_{d=1}^{D_1} \theta_{1d}(X_{1dn}), \alpha_2 + \sum_{d=1}^{D_2} \theta_{2d}(X_{2dn})\} \end{pmatrix}.$$

Let $\boldsymbol{P}_{kd} = \text{diag}\left\{-\frac{\partial \boldsymbol{q}_{1kd}}{\partial \theta_k}\right\}$.

Based on a first order Taylor series approximation of the local estimating equations (4.10) and ignoring all terms of order $O[1/(nh_{kd})]$, we obtain that

$$\widehat{\boldsymbol{\beta}}_{kd}(\boldsymbol{x}) \approx (\mathbb{X}_{kdx}^T \boldsymbol{W}_{kdx} \boldsymbol{P}_{kd} \mathbb{X}_{kdx})^{-1} \mathbb{X}_{kdx} \boldsymbol{W}_{kdx} \boldsymbol{P}_{kd} \tilde{\boldsymbol{Y}}_{kd}$$

where

$$\tilde{\boldsymbol{Y}}_{kd} = \boldsymbol{\theta}_{k1} + \ldots + \boldsymbol{\theta}_{kD_k} + \boldsymbol{P}_{kd}^{-1} \boldsymbol{q}_{1kd}.$$

Approximate conditional bias and variance expressions for the local polynomial estimators obtained by solving equations (4.10) are, for the bivariate case and for all polynomials of odd degree, given in Theorem 4.4, where the Fisher information matrix $J(\{X_{1i}, X_{2i}\})$ is replaced by

$$J_{kk}(\{X_{11i}, X_{1D_1 i}\}, \{X_{21i}, X_{2D_2 i}\}),$$

and the functions $J_1(\cdot)$ and $J_2(\cdot)$ are redefined according to this definition of J_{kk}. For more than two covariates or for polynomials of even degree, the notation is getting rather cumbersome, but one can easily conjecture how these results should look like.

From the previous discussion it should be clear that also for multiparameter additive models problems are to be expected when the backfitting estimation scheme

is used to estimate the additive components. An example of such a model is a multivariate response (generalized) linear model where each component of the mean response is assumed to be additive in the covariates. Estimation by backfitting methods of the additive components of each mean curve separately will give the same problems as mentioned in Section 4.4.3. It should be stressed that the optimality property of the "separated" components (e.g. the marginal means in previous example) remains to hold. Moreover, this property also holds if we don't assume an additive structure but use multivariate smoothing techniques instead to estimate each of these separated curves. For example, for a bivariate linear model with two covariates, we can estimate $\theta_1(x_1, x_2)$ and $\theta_2(x_1, x_2)$ each with its own optimal bandwidth matrix by solving a set of local estimating equations similar to (4.4).

For the classical additive model, Wand (1998) obtains a central limit theorem for local polynomial backfitting estimators. It would be very interesting to also obtain such a result for the local polynomial estimators in the general multivariate additive models as defined above.

4.6 Discussion

From the results of this chapter, it is clear that the way estimating equations are defined is important. A wrong choice of estimating equations will yield estimators with undesirable statistical properties.

In additive models we observe, from the conditional bias in Theorem 4.4, that the backfitting approach does not achieve the same bias as the so-called oracle estimator, based on knowing all other components. In classical regression models (as studied by Opsomer and Ruppert, 1997), this dependence of the bias on all components of the additive model disappears when the covariates are mutually independent. In this case, the bias of the estimator for, say, $\theta_1(\cdot)$ does not depend on additive components other than $\theta_1(\cdot)$. In general models, this property does not hold. The variance, though, is optimal in the sense that it is the same as the variance of the oracle estimator.

The main message of this chapter is that for a multiparameter model, the set of estimating equations II (see equation (4.4)) is the right choice. It allows full flexibility in selecting the different optimal smoothing parameters for the different curves. An interesting topic for further research is to investigate the performance of

data-driven bandwidth selectors, based on set II.

Since most datasets contain a mixture of discrete and continuous covariates, there really is a need for semiparametric models, which include the non-continuous variables in a parametric way, while the continuous variables can be handled using smoothing ideas, or also parametrically.

Recently, there has been a lot of attention to semiparametric models. For a one-dimensional continuous covariate X, and some other vector of covariates $\boldsymbol{Z}$, Hastie and Tibshirani (1990) describe a generalized additive model $\eta(X, \boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{\beta} + \gamma(X)$, where, for a known link function $g(\cdot)$, and Y the response vector, $\eta(X, \boldsymbol{Z}) = g(E[Y|X, \boldsymbol{Z}])$. This idea has been generalized by Carroll, Fan, Gijbels and Wand (1997), to allow for more than one variable in the function $\gamma(\cdot)$, but yet remain the ease of interpretation and avoid the curse of dimensionality. Their generalized partially linear single-index model looks as follows: $\eta(\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{\beta} + \gamma(\boldsymbol{X}^t\boldsymbol{\alpha})$, for a D dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_D)^T$ and $\|\boldsymbol{\alpha}\| = 1$. In the context of classical regression models, where the function $g(\cdot)$ is the identity function, Opsomer and Ruppert (1997b) propose the following semiparametric additive model $E[Y|\boldsymbol{X}, \boldsymbol{Z}] = \boldsymbol{Z}^T\boldsymbol{\beta} + \sum_{d=1}^{D} \gamma_d(X_d)$.

A straightforward extension is to reformulate this model in the context of generalized linear models: $\eta(\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{\beta} + \sum_{d=1}^{D} \gamma_d(X_d)$, with an obvious extension to the quasi likelihood context. This is different from the semiparametric model of Severini and Staniswalis (1994), where $\eta(\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{\beta} + \boldsymbol{\gamma}(\boldsymbol{X})$, for a D-variate function $\boldsymbol{\gamma}(\cdot) : I\!R^D \to I\!R$. The additive model has the advantage that we can avoid struggling with high dimensional functions, while they still are more general than the single-index model of Carroll, Fan, Gijbels and Wand (1997). Moreover, additive models are easy to interpret, since each of the estimated functions γ_d can be plotted against its corresponding covariate. The graphical representation of the Severini and Staniswalis (1994) model is only possible for a one or two dimensional covariate $\boldsymbol{X}$.

In a one-parameter semiparametric model $\psi\{\boldsymbol{Y}; \theta(\boldsymbol{x}, \boldsymbol{z})\}$ we define the parameter $\theta(\cdot)$ as

$$\theta(\boldsymbol{X}, \boldsymbol{Z}) = \alpha + \boldsymbol{Z}^T\boldsymbol{\beta} + \sum_{d=1}^{D} \gamma_d(X_d).$$

The estimators might be obtained via the set of equations (4.11), see also Opsomer and Ruppert (1997b) for similar equations in the framework of classical regression

models, where $\theta(\boldsymbol{X}, \boldsymbol{Z}) = E[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}]$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbb{Z}^T \boldsymbol{P} \mathbb{Z})^{-1} \mathbb{Z}^T \boldsymbol{P} (\check{\boldsymbol{Y}} - \widehat{\boldsymbol{\alpha}} - \sum_{d=1}^{D} \widehat{\boldsymbol{\gamma}}_d(\boldsymbol{x})) \\
\widehat{\boldsymbol{\gamma}}_1 &= \boldsymbol{S}_1^*(\check{\boldsymbol{Y}} - \widehat{\boldsymbol{\alpha}} - \boldsymbol{Z}^T \hat{\boldsymbol{\beta}} - \sum_{k \neq 1} \widehat{\boldsymbol{\gamma}}_k) \qquad (4.11) \\
&\vdots \\
\widehat{\boldsymbol{\gamma}}_D &= \boldsymbol{S}_D^*(\check{\boldsymbol{Y}} - \widehat{\boldsymbol{\alpha}} - \boldsymbol{Z}^T \hat{\boldsymbol{\beta}} - \sum_{k \neq D} \widehat{\boldsymbol{\gamma}}_k)
\end{aligned}$$

where $\widehat{\boldsymbol{\gamma}}_d = (\hat{\gamma}_d(X_{d1}), \ldots, \hat{\gamma}_d(X_{dn}))^T$, $\boldsymbol{P} = \text{diag}_{1 \leq i \leq n}(-\frac{\partial \psi}{\partial \theta}\{\boldsymbol{Y}_i; \theta(\boldsymbol{X}_i, \boldsymbol{Z}_i)\})$ and where $\mathbb{Z}$ is the design matrix of the parametrically modeled variables.

It would be worth-while to perform a detailed study about asymptotic properties of estimators in semiparametric multiparameter models, where either one or more of the parameters can be modeled parametrically, semiparametrically or nonparametrically. It is to be expected that, by using the above estimation scheme, the order of the bandwidth parameters, which, in fully nonparametric models, for p odd is $O(n^{-1/(2p+3)})$ and for p even $O(n^{-1/(2p+5)})$, does no longer hold in semiparametric models *if* we want to obtain $\sqrt{n}$-consistent estimators for the coefficient vector of parametrically modeled variables, see, e.g., Opsomer and Ruppert (1997b). We conjecture that this problem might be solved by using generalized profile likelihood methods (Severini and Wong, 1992).

Chapter 5

Penalized Regression Splines for Additive Models

Generalized additive models have become one of the most widely used modern statistical tools. Traditionally they are fitted through scatterplot smoothing and the backfitting algorithm. However, a more recent development is the direct fitting through the use of low-rank smoothers such as penalized splines (Hastie, 1996, Marx and Eilers, 1998). There are a number of advantages of such an approach, particularly regarding computation. In this chapter we exploit the explicitness of penalized spline additive models to derive some useful and revealing theoretical approximations.

These results can also be found in Aerts, Claeskens and Wand (1999).

5.1 Introduction

Generalized additive models (GAM) are among the most practically used modern statistical techniques. Examples of their use in applications includes political science (Beck and Jackman, 1998), economics (Linton and Härdle, 1996) and environmental epidemiology (Schwartz, 1994). The main catalysts for this widespread use by practitioners is the exemplary monograph on the topic, Hastie and Tibshirani (1990), and the availability of the function `gam()` in the `S-PLUS` language for fitting such models (see e.g. Chambers and Hastie, 1991).

One aspect of GAM that has been slow to develop is the statistical properties of the estimation strategies for fitting such a model. Perhaps the main reason for this is the non-explicit nature of the most common type of GAM fitting procedure: *backfitting* combined with *local scoring* (Hastie and Tibshirani, 1990). This non-explicitness appears to be the main motivation for the *marginal integration* approach to additive model fitting (Linton and Nielsen, 1995) and a sophisticated theory now exists for this strategy (e.g. Fan, Härdle and Mammen, 1998). The statistical properties of additive models based on backfitting and local polynomial estimators have since been derived by Opsomer and Ruppert (1997) and Opsomer (1999). Except for the results on local polynomial estimators, which are presented in Chapter 4, there appears to be little or no theory on the statistical properties of estimators in a *generalized* additive model.

An attractive alternative to backfitting and marginal integration is direct fitting based on low rank smoothers such as penalized splines (Hastie, 1996, Marx and Eilers, 1998). This approach has chiefly been motivated by computational expediency. However, the directness of the method also means that the estimator has an explicit form. This paper exploits this fact. We derive simple closed form approximations to the bias and variance of the estimator and its components, not just for ordinary additive models but for GAM. The results are both useful and revealing. For example, they can be used to provide rough starting values for the smoothing parameters (see Section 5.3.1). They also provide some backup for commonly used degrees of freedom approximations (see Section 5.3.2).

Section 5.2 defines the penalized regression splines estimators in classical regression models with only one covariate. Section 5.3 treats the case of classical additive regression models. Extensions to semiparametric models, generalized additive models and additive generalized estimating equations are described in Section 5.4.

5.2 Single covariate models

To explain the idea of penalized regression splines and to introduce some notation, first assume that we have a classical regression model

$$\boldsymbol{Y} = \boldsymbol{f} + \varepsilon, \qquad \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $\boldsymbol{I}$ is the identity matrix.

A parametric regression model (of degree p) for these data is

$$f(x_j) = \beta_0 + \beta_1 x_j + \ldots + \beta_p x_j^p,$$

where the unknown coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ can for example be estimated by minimizing a least squares criterion, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{Y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2$$

where $\tilde{\boldsymbol{X}}$ is the design matrix

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix}.$$

To allow for more flexibility, in addition to the polynomial basis functions, we can add spline basis functions $\{(x_j - \kappa_1)_+\}^p, \ldots, \{(x_j - \kappa_K)_+\}^p$, and use a regression spline model

$$f(x_j) = \beta_0 + \beta_1 x_j + \ldots + \beta_p x_j^p + \sum_{k=1}^{K} \beta_{p+k}\{(x_k - \kappa_k)_+\}^p,$$

where $\kappa_1, \ldots, \kappa_K$ is a set of knots and $(x_j - \kappa_k)_+$ represents the positive part of $(x_j - \kappa_k)$, i.e. $(x_j - \kappa_k)_+ = 0$ when $x_j \leq \kappa_k$ and is equal to $(x_j - \kappa_k)$ otherwise. Estimators of the coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p+K})^T$ can be obtained in the same way as before (using the method of least squares) where we now define the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p & \{(x_1 - \kappa_1)_+\}^p & \cdots & \{(x_1 - \kappa_K)_+\}^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p & \{(x_n - \kappa_1)_+\}^p & \cdots & \{(x_n - \kappa_K)_+\}^p \end{bmatrix}.$$

The smoothness of the fitted model depends on the degree p, the location of the knots and on how many knots there are. Several papers address the selection of the knots, see, e.g., Friedman and Silverman (1989), Stone, Hansen, Kooperberg and Truong (1997) and Smith and Kohn (1996). Wand (1997) gives a comparison of some of these approaches. Ruppert and Carroll (1997) recommend K between 5 and 40 and κ_k the $k/(K+1)$th sample quantile of the covariate values $x_1, \ldots, x_n$.

An alternative to knot selection is to keep all the knots, but to restrict their influence on the fitted model, that is, to constrain the values of these coefficients

which correspond to the spline basis functions. For example, we can bound the sum of the squared coefficients (i.e. the squared L_2 norm) by some constant. In this case, the complete vector of coefficients is obtained as follows,

$$\hat{\boldsymbol{\beta}} = \arg\min\{\|\boldsymbol{Y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \alpha\boldsymbol{\beta}^T\boldsymbol{D}\boldsymbol{\beta}\} \tag{5.1}$$

where $\boldsymbol{D} = \text{diag}(\mathbf{0}_{p\times 1}, \mathbf{1}_K)$. Since the amount of smoothing is determined by α, this is called the smoothing parameter. Some other penalty functions are proposed by Ruppert and Carroll (1997, 1999).

The fitted values corresponding to equation (5.1) are given by

$$\widehat{\boldsymbol{f}}_\alpha = \boldsymbol{G}_\alpha\boldsymbol{Y} \qquad \text{where} \qquad \boldsymbol{G}_\alpha = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{A}_\alpha)^{-1}\boldsymbol{X}^T$$

and $\boldsymbol{A}_\alpha = \alpha\boldsymbol{D}$. Some properties of these estimators are obtained in the next sections.

5.3 Additive models

In this section we will study the penalized regression spline estimators in the standard additive models framework. In these models the response Y_i $(i = 1, \ldots, n)$ depends in an additive way on the d covariates, $x_{1i}, \ldots, x_{di}$, through arbitrary univariate functions f_j,

$$Y_i = \beta_0 + \sum_{j=1}^{d} f_j(x_{ji}) + \varepsilon_i. \tag{5.2}$$

It is assumed that the i.i.d. errors have mean zero, variance σ^2 and are independent of the covariates. Each of the functions f_j will be estimated by a degree p_j penalized spline estimator with smoothing parameter α_j.

Note that, due to identifiability requirements, the f_j in (5.2) are defined only up to an additive constant. Therefore, they can be replaced by $f_j(x_{ji}) - \frac{1}{n}\sum_{i=1}^{n} f_j(x_{ji})$. This makes β_0 orthogonal to the f_j and its estimate, $\widehat{\beta}_0 = \overline{Y}$, is independent of the x_{ji}'s. Since $\overline{Y}$ can be subtracted from the $Y_i's$ without affecting the model fitting we can assume, without loss of generality, that $\beta_0 = 0$. This convention will be made from here onwards. For any $(n \times m)$ matrix $\mathbf{C}$, denote by $\mathbf{C}^*$ the centered matrix

$$\mathbf{C}^* = (\boldsymbol{I} - \mathbf{1}_n \cdot \mathbf{1}_n^T/n)\mathbf{C}.$$

In matrix notation we can write the model as

$$\boldsymbol{Y} = \sum_{j=1}^{d} \boldsymbol{f}_j + \varepsilon, \qquad \text{cov}(\varepsilon) = \sigma^2\boldsymbol{I} \tag{5.3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{f}_j = (f_j(x_{j1}), \ldots, f_j(x_{jn}))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Let $\boldsymbol{f} = \sum_{j=1}^{d} \boldsymbol{f}_j$. The penalized spline fit to these data is

$$\widehat{\boldsymbol{f}}_{\boldsymbol{\alpha}} = \boldsymbol{G}_{\boldsymbol{\alpha}}\boldsymbol{Y} \qquad \text{where} \qquad \boldsymbol{G}_{\boldsymbol{\alpha}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{A}_{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}^T$$

with centered design matrix

$$\boldsymbol{X} = [\boldsymbol{X}_1^* \cdots \boldsymbol{X}_d^*], \qquad \boldsymbol{A}_{\boldsymbol{\alpha}} = \underset{1 \le j \le d}{\text{blockdiag}}(\alpha_j \boldsymbol{D}_j)\ , \qquad \boldsymbol{D}_j = \text{diag}(\boldsymbol{0}_{p_j \times 1}, \boldsymbol{1}_{K_j})$$

$$\boldsymbol{X}_j = \begin{bmatrix} x_{j1} & \cdots & x_{j1}^{p_j} & \{(x_{j1} - \kappa_{j1})_+\}^{p_j} & \cdots & \{(x_{j1} - \kappa_{jK_j})_+\}^{p_j} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{jn} & \cdots & x_{jn}^{p_j} & \{(x_{jn} - \kappa_{j1})_+\}^{p_j} & \cdots & \{(x_{jn} - \kappa_{jK_j})_+\}^{p_j} \end{bmatrix}$$

and $\kappa_{j1}, \ldots, \kappa_{jK_j}$ $(j = 1, \ldots, d)$ is a set of knots in the jth direction. The knots are usually taken to be relatively "dense" among the observations in an attempt to capture the curvature in $\boldsymbol{f}_j$. A reasonable allocation rule is one knot for every 4-5 observations, up to a maximum of about 40 knots. Subscripts denoting the dependence of matrices on the smoothing vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$ will be omitted, unless $\boldsymbol{\alpha} = \boldsymbol{0}$, in which case a subscript $\boldsymbol{0}$ will be used.

For any $j = 1, \ldots, d$, the full smoother matrix $\boldsymbol{G}$ can be decomposed as

$$\boldsymbol{G} = \boldsymbol{G}_j + \boldsymbol{G}_{[-j]}$$

where

$$\boldsymbol{G}_j = [\boldsymbol{0} \cdots \boldsymbol{0}\ \boldsymbol{X}_j\ \boldsymbol{0} \cdots \boldsymbol{0}](\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T \tag{5.4}$$

$$\boldsymbol{G}_{[-j]} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_{j-1}\ \boldsymbol{0}\ \boldsymbol{X}_{j+1} \cdots \boldsymbol{X}_d](\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T. \tag{5.5}$$

The corresponding additive component fits are

$$\widehat{\boldsymbol{f}}_j = \boldsymbol{G}_j\boldsymbol{Y} \quad \text{and} \quad \widehat{\boldsymbol{f}}_{[-j]} = \boldsymbol{G}_{[-j]}\boldsymbol{Y}.$$

In the model with only covariate x_j, the regression spline smoother matrix is

$$\boldsymbol{S}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{X}_j + \boldsymbol{A}_j)^{-1}\boldsymbol{X}_j^T. \tag{5.6}$$

In the model with all covariates except x_j, the regression spline smoother matrix is

$$\boldsymbol{S}_{[-j]} = \boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T\boldsymbol{X}_{[-j]} + \boldsymbol{A}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T. \tag{5.7}$$

where

$$\boldsymbol{X}_{[-j]} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_{j-1}\ \boldsymbol{X}_{j+1} \cdots \boldsymbol{X}_d] \qquad \text{and} \qquad \boldsymbol{A}_{[-j]} = \underset{1\le i\le d, i\neq j}{\text{blockdiag}}(\alpha_i \boldsymbol{D}_i).$$

The following recursive formula is an important stepping stone towards obtaining approximations in penalized spline additive models. For any covariate x_j ($j = 1, \ldots, d$), the result allows one to write the full smoother matrix $\boldsymbol{G}$ in terms of the design matrix $\boldsymbol{X}_j$ and the smoother matrix in the sub-model corresponding to deletion of the jth covariate.

Theorem 5.1 *For any $j = 1, \ldots, d$,*

$$\boldsymbol{G} = \boldsymbol{S}_{[-j]} + (\boldsymbol{I} - \boldsymbol{S}_{[-j]})\boldsymbol{G}_j \tag{5.8}$$

with

$$\begin{aligned} \boldsymbol{G}_j &= \boldsymbol{X}_j\{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{S}_{[-j]})\boldsymbol{X}_j + \boldsymbol{A}_j\}^{-1}\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{S}_{[-j]}) \\ &= \boldsymbol{S}_j[\boldsymbol{I} - \boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T(\boldsymbol{I} - \boldsymbol{S}_j)\boldsymbol{X}_{[-j]} + \boldsymbol{A}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T(\boldsymbol{I} - \boldsymbol{S}_j)], \end{aligned}$$

if the inverse matrices exist.

To construct the additive model smoother matrix $\boldsymbol{G}_j$, it is sufficient to know the smoother matrix $\boldsymbol{S}_{[-j]}$ and the design matrix $\boldsymbol{X}_j$, or, using the equivalent expression (5.9), the univariate smoother matrix $\boldsymbol{S}_j$ and the design matrix $\boldsymbol{X}_{[-j]}$ corresponding to the $d-1$ other covariates. Hence, this recursive formula shows how the full smoother matrix $\boldsymbol{G}$ can be built from lower dimensional pieces. Equation (5.8) is comparable to the result of Lemma 2.1 in Opsomer (1999), where, for local polynomial estimators using backfitting, a similar identity can be obtained.

Using formula (5.8), we obtain:

Theorem 5.2 *Assuming that $\boldsymbol{\alpha} \to \boldsymbol{0}$ we have*

$$\boldsymbol{G} = \boldsymbol{G_0} - \sum_{j=1}^{d} \alpha_j \widetilde{\boldsymbol{B}}_j + o\left(\sum_{j=1}^{d} \alpha_j \widetilde{\boldsymbol{B}}_j\right), \tag{5.9}$$

where

$$\widetilde{\boldsymbol{B}}_j = \hat{\boldsymbol{X}}_j(\hat{\boldsymbol{X}}_j^T\hat{\boldsymbol{X}}_j)^{-1}\boldsymbol{D}_j(\hat{\boldsymbol{X}}_j^T\hat{\boldsymbol{X}}_j)^{-1}\hat{\boldsymbol{X}}_j^T \quad \textit{and} \quad \hat{\boldsymbol{X}}_j = (\boldsymbol{I} - \boldsymbol{S}_{\boldsymbol{0},[-j]})\boldsymbol{X}_j,$$

provided that all inverse matrices exist.

The matrix $\boldsymbol{G_0}$ is the usual "hat" matrix for fully parametric models ($\boldsymbol{\alpha} = \mathbf{0}$). The matrix $\tilde{\boldsymbol{X}}_j$ is obtained by projecting $\boldsymbol{X}_j$ orthogonal to the subspace determined by $\boldsymbol{X}_{[-j]}$. Details on the derivation of Theorem 5.2 can be found in the Appendix.

In the next two subsections we show how approximation (5.9) can be used to aid practical implementation of additive models.

5.3.1 Approximation of the risk

Our first application of the results of the preceding section is to approximate the risk in an additive model. Such approximations have the advantage of being simpler to optimize and can, perhaps, aid the practical selection of the smoothing parameters. For convenience we will work with the Mean Average Squared Error (MASE)

$$\text{MASE}(\widehat{\boldsymbol{f}}) = \tfrac{1}{n} E\|\widehat{\boldsymbol{f}} - \boldsymbol{f}\|^2.$$

This can be decomposed into the average variance plus the average squared bias and then simplified to give:

$$\text{MASE}(\widehat{\boldsymbol{f}}) = \tfrac{\sigma^2}{n}\text{tr}\{(\boldsymbol{G})^2\} + \tfrac{1}{n}\|(\boldsymbol{G} - \boldsymbol{I})\boldsymbol{f}\|^2.$$

Theorem 5.3 follows from this expression and (5.9):

Theorem 5.3 *The leading terms in the asymptotic expansion of* $\text{MASE}(\widehat{\boldsymbol{f}})$ *as* $\boldsymbol{\alpha} \to \mathbf{0}$ *are*

$$\text{AMASE}(\widehat{\boldsymbol{f}}) = \frac{1}{n}\left[\sigma^2\left\{\sum_{j=1}^{d}(p_j + K_j) - 2\boldsymbol{\alpha}^T\boldsymbol{q}\right\} + \boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha}\right]$$

where the entries of $\boldsymbol{q}$ $(d \times 1)$ *and* $\boldsymbol{Q}$ $(d \times d)$ *are* $\boldsymbol{q}_j = \text{tr}(\widetilde{\boldsymbol{B}}_j)$, *and*

$$\boldsymbol{Q}_{jj'} = (\widetilde{\boldsymbol{B}}_j\boldsymbol{f})^T(\widetilde{\boldsymbol{B}}_{j'}\boldsymbol{f}) + \sigma^2\text{tr}(\widetilde{\boldsymbol{B}}_j\widetilde{\boldsymbol{B}}_{j'}).$$

The AMASE*-optimal smoothing parameters are therefore given by*

$$\boldsymbol{\alpha}_{\text{AMASE}} = \sigma^2\boldsymbol{Q}^{-1}\boldsymbol{q}. \tag{5.10}$$

An application of this result is depicted in Figure 5.1. It shows the result of applying (5.10) to the Californian air pollution data from Breiman and Friedman (1985), and used for illustratory purposes by Hastie (1996). Of course (5.10) requires knowledge of $\boldsymbol{f}$ and σ^2, so preliminary estimates of those were plugged in. These were

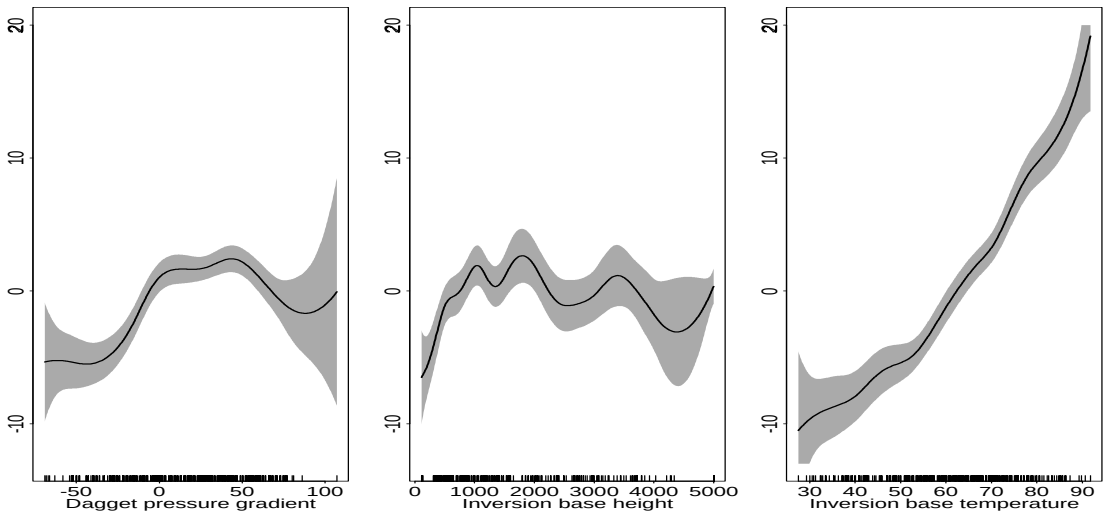


Figure 5.1: Penalized spline additive model fit to Californian air pollution data. The amount of smoothing is obtained by application $\boldsymbol{\alpha}_{\text{AMASE}}$ *with a preliminary estimate of the* f_j*'s obtained via piecewise quadratic fitting and Mallows'* C_p*.*

obtained using blockwise quadratic fits as suggested by Härdle and Marron (1995) and Ruppert, Sheather and Wand (1995). As in the latter reference, Mallows' C_p was used to choose among all blockwise quadratic fits with 4 or less blocks.

Comparison of Figure 5.1 with Figure 5 of Hastie (1996) shows the estimated functions for `Dagget pressure gradient` and `Inversion base temperature` being roughly comparable. However, that for `Inversion base height` is quite a bit more wiggly. As a check, we fit the data using `S-PLUS`'s `gam()` with default smoothing parameter choice and subtracted off the fitted values for `Dagget pressure gradient` and `Inversion base temperature` from the response. We then applied `smooth.spline()` to resulting scatterplot, with generalized cross-validation for smoothing parameter choice. It chose an even larger number of degrees of freedom, so there is some corroboration for the $\boldsymbol{\alpha}_{\text{AMASE}}$-based result obtained here.

While this is just one example, it indicates that $\boldsymbol{\alpha}_{\text{AMASE}}$ can be useful for getting an idea of the amount of smoothing required for fitting an additive model.

5.3.2 Approximation of the degrees of freedom

In an additive model with d components, the *degrees of freedom* for component $\widehat{\boldsymbol{f}}_j$ is defined to be (see, e.g., Hastie and Tibshirani, 1990, p. 128)

$$\mathrm{df}_j = \mathrm{tr}(\boldsymbol{G}_j).$$

While this is straightforward to compute using the full design matrix $\boldsymbol{X}$, it is often desirable to have an approximation to this quantity that uses only information about component j. In this way, the smoothing parameter corresponding to a particular degrees of freedom value can be specified. A natural candidate for this is

$$\mathrm{tr}(\boldsymbol{S}_j) = \mathrm{tr}\{(\boldsymbol{X}_j^T\boldsymbol{X}_j + \alpha_j\boldsymbol{D}_j)^{-1}\boldsymbol{X}_j^T\boldsymbol{X}_j\}$$

which has the advantage that it depends only on α_j, and thus allows for easier determination of the smoothing parameter corresponding to the specified degrees of freedom value. This approximation is used, for example, by the function `gam()` in the `S-PLUS` computing package (see e.g. Hastie and Tibshirani, 1990, p. 158)

From (5.9) we obtain:

Theorem 5.4

$$\frac{\mathrm{tr}(\boldsymbol{G}_j)}{\mathrm{tr}(\boldsymbol{S}_j)} - 1 = -\alpha_j \frac{\mathrm{tr}[\boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T(\boldsymbol{I} - \boldsymbol{S}_{j0})\boldsymbol{X}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T\boldsymbol{B}_j]}{p_j + K_j - \alpha_j\mathrm{tr}\{(\boldsymbol{B}_j\}}$$

where $\boldsymbol{B}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\boldsymbol{D}_j(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^T$.

This theorem shows that, for $\alpha_j \to 0$, $\mathrm{tr}(\boldsymbol{S}_j)$ has the same limit as $\mathrm{tr}(\boldsymbol{G}_j)$, giving some justification for its use.

The derivation of Theorem 5.4 is given in the Appendix.

The accuracy of this approximation is tested for the fit shown in Figure 5.1. The results are given in Table 5.1. It shows that the accuracy is very reasonable in this case.

5.4 Extensions

In this section we consider two important extensions of the classical additive model. The first extension allows one to model some of the covariates in a parametric way,

	$j=1$	$j=2$	$j=3$
tr($\boldsymbol{G}_j$) (exact)	9.036	11.803	9.884
tr($\boldsymbol{S}_j$) (approx.)	9.136	11.886	9.651
tr($\boldsymbol{G}_j$)/tr($\boldsymbol{S}_j$)	0.989	0.993	1.024

Table 5.1: Comparison between tr($\boldsymbol{S}_j$) and tr($\boldsymbol{G}_j$) for fit to Californian air pollution data.

while others are modeled nonparametrically using penalized regression splines. It turns out that the theoretical results for this semiparametric model are very similar to the full nonparametric case. In a second subsection, we will extend the above calculations to the broad class of generalized additive models.

5.4.1 Semiparametric models

When penalized regression splines are used to model the nonparametric components of a semiparametric model, the same methodology as in fully nonparametric models can be used. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ ($q < d$) denote the design matrices of the parametric components, setting $\alpha_1 = \ldots = \alpha_q = 0$ ensures that these components are not being penalized. With this simple adjustment, we can estimate simultaneously all parametric and nonparametric components.

A partitioning of the design matrix in the following way,

$$\boldsymbol{X} = [\boldsymbol{X}_{\text{parm}}\ \boldsymbol{X}_{\text{nonp}}], \text{ where } \boldsymbol{X}_{\text{parm}} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_q] \text{ and } \boldsymbol{X}_{\text{nonp}} = [\boldsymbol{X}_{q+1} \cdots \boldsymbol{X}_d]\ ,$$

provides us an easy formula to study separately the parametric and nonparametric parts of the model. Using formulas for the inverse of a partitioned matrix, we obtain the following expression for the smoother matrix $\boldsymbol{G}$,

$$\boldsymbol{G} = \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T + \boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R} + \boldsymbol{A}_{\text{nonp}})^{-1}\boldsymbol{R}^T \tag{5.11}$$

where

$$\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T)\boldsymbol{X}_{\text{nonp}} \text{ and } \boldsymbol{A}_{\text{nonp}} = \underset{q+1\le j\le d}{\text{blockdiag}}(\alpha_j\boldsymbol{D}_j).$$

This implies that the estimator $\widehat{\boldsymbol{f}}$ can be written as the sum of the regression estimator for the parametric part in $\boldsymbol{X}_{\text{parm}}$ and the regression spline estimator for the

other part, after projection of $\boldsymbol{X}_{\text{nonp}}$ orthogonal to the $\boldsymbol{X}_{\text{parm}}$-subspace. Matrix $\boldsymbol{G}$ can also be written as the sum of a parametric part

$$\boldsymbol{G}_{\text{parm}} = \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T\{\boldsymbol{I} - \boldsymbol{X}_{\text{nonp}}(\boldsymbol{R}^T\boldsymbol{R} + \boldsymbol{A}_{\text{nonp}})^{-1}\boldsymbol{R}^T\}$$

and a nonparametric part

$$\boldsymbol{G}_{\text{nonp}} \quad = \quad \boldsymbol{X}_{\text{nonp}}(\boldsymbol{R}^T\boldsymbol{R} + \boldsymbol{A}_{\text{nonp}})^{-1}\boldsymbol{R}^T \tag{5.12}$$

which can be used to obtain separate estimators $\widehat{\boldsymbol{f}}_{\text{parm}} = \boldsymbol{G}_{\text{parm}}\boldsymbol{Y}$ and $\widehat{\boldsymbol{f}}_{\text{nonp}} = \boldsymbol{G}_{\text{nonp}}\boldsymbol{Y}$ for, respectively, the parametric and nonparametric components. Note that if there is no parametric part ($\boldsymbol{X}_{\text{parm}} = \boldsymbol{0}$), everything reduces to the results of Section 5.3, whereas for a true semiparametric model, the estimator of the nonparametric part is given by (see equation (5.12))

$$\boldsymbol{X}_{\text{nonp}}(\boldsymbol{X}_{\text{nonp}}^T\boldsymbol{W}\boldsymbol{X}_{\text{nonp}} + \boldsymbol{A}_{\text{nonp}})^{-1}\boldsymbol{X}_{\text{nonp}}^T\boldsymbol{W}\boldsymbol{Y} \ ,$$

where $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T$.

Also in a semiparametric model an optimal smoothing parameter can be obtained by minimizing the MASE with respect to $\boldsymbol{\alpha}$. The second term in the right-hand side of equation (5.11) shows that this boils down to replacing $\boldsymbol{X}$ by $\boldsymbol{R}$ in Section 5.3.1. Another strategy is to select the smoothing parameters optimally for estimation of the nonparametric part only. In this case we focus on matrix (5.12), which differs from the second term in (5.11) by the matrix $\boldsymbol{W}$. Because of this, the expression for the asymptotically optimal smoothing parameter becomes more complicated and resembles the formula which will be given in the next section.

5.4.2 Generalized additive models

Model (5.3) is mainly used for normally distributed errors. This Gaussian regression model is a member of the class of generalized linear models (GLM), see, e.g., McCullagh and Nelder (1989). In all these GLM, the likelihood of the response belongs to an exponential family and can be written as follows,

$$\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \ ,$$

where θ is the so-called natural parameter, which is related to the mean response in the following way, $db(\theta)/d\theta = E(Y) = \mu$, and ϕ is a dispersion parameter. A

generalized linear model is further specified by a known "link" function $g(\cdot)$, and $\eta = g(\mu)$ is called the systematic component.

Instead of modeling η as a linear function of the covariates, which would result in the classical generalized linear model, we can use a nonparametric estimator in an additive models framework. More specifically, as in Hastie and Tibshirani (1990), we assume $\eta(x_1, \ldots, x_d) = \eta_1(x_1) + \ldots + \eta_d(x_d)$. While Eilers and Marx (1998) use B-splines to estimate the η_j's, we will, similar to Ruppert and Carroll (1997), estimate each of these functions η_j using penalized regression splines of degree p_j. This means that each additive component $\boldsymbol{\eta}_j = [\eta_j(x_{j1}), \ldots, \eta_j(x_{jn})]^T$ is modeled as $\boldsymbol{X}_j\boldsymbol{\beta}_j$ with $\boldsymbol{X}_j$ as in Section 5.3. The parameter vector $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T]^T$ can now be estimated by maximizing the following penalized log likelihood function:

$$\sum_{i=1}^{n} \left[Y_i\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\beta}) - b\{\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\beta})\} \right] / a(\phi) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}.$$

For $\boldsymbol{U}_\beta$ the vector of first partial derivatives and $\boldsymbol{J}_\beta$ the matrix of minus second partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$, we immediately obtain that the $(k+1)$st update in a Newton-Raphson iterative procedure is given by

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{J}_{\beta^{(k)}} + \boldsymbol{A})^{-1}(\boldsymbol{J}_{\beta^{(k)}}\boldsymbol{\beta}^{(k)} + \boldsymbol{U}_{\beta^{(k)}}),$$

or, by the chain rule,

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}^T \boldsymbol{J}_{\eta^{(k)}} \boldsymbol{X} + \boldsymbol{A})^{-1} \boldsymbol{X}^T \boldsymbol{J}_{\eta^{(k)}} (\boldsymbol{\eta}^{(k)} + \boldsymbol{J}_{\eta^{(k)}}^{-1} \boldsymbol{U}_{\eta^{(k)}}), \tag{5.13}$$

with $\boldsymbol{A}$ as in Section 5.3 and with $\log L(\cdot)$ denoting the log likelihood of the data,

$$\boldsymbol{J}_\eta = \operatorname*{diag}_{1 \le i \le n} \left(\frac{\partial^2 \log L(Y_i; \boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \eta^2} \right)$$

and

$$\boldsymbol{U}_\eta = \left(\frac{\partial \log L(Y_1; \boldsymbol{x}_1, \boldsymbol{\beta})}{\partial \eta}, \ldots, \frac{\partial \log L(Y_n; \boldsymbol{x}_n, \boldsymbol{\beta})}{\partial \eta} \right)^T.$$

If the observed Fisher information matrix is replaced by its expectation $\boldsymbol{I}_\beta = E(\boldsymbol{J}_\beta)$, we obtain the iterative solutions of a Fisher scoring procedure. Note that these two algorithms coincide if the canonical link function is used, that is, if $\eta = \theta$.

From equation (5.13) it is immediately clear that an equivalent way of obtaining the estimators $\widehat{\boldsymbol{\beta}}$ is via iteratively reweighted ridge regression, where the adjusted

dependent variable is defined as $\boldsymbol{Z}_{\eta^{(k)}} = \boldsymbol{\eta}^{(k)} + \boldsymbol{J}^{-1}_{\eta^{(k)}}\boldsymbol{U}_{\eta^{(k)}}$. Writing the estimator of the coefficient $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{J}_\eta\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T\boldsymbol{J}_\eta\boldsymbol{Z}_\eta,$$

which allows us to extend the asymptotic results of Section 5.3 to generalized additive models. For Gaussian responses, all results shown below simplify to those of Section 5.3.

Approximation of the risk

The smoothing parameter $\boldsymbol{\alpha}$ will be selected by extending the definition of MASE to the context of generalized additive models. The overall risk is now measured by

$$\text{MASE}(\widehat{\boldsymbol{\eta}}) = \frac{1}{n}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|^2.$$

This can be rewritten as,

$$\text{MASE}(\widehat{\boldsymbol{\eta}}) = \tfrac{1}{n}\text{tr}\,\{\boldsymbol{G}\text{Var}(\boldsymbol{U}_\eta)\boldsymbol{G}\} + \tfrac{1}{n}\|\boldsymbol{G}\boldsymbol{I}_\eta\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\eta}\|^2,$$

where $\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{I}_\beta + \boldsymbol{A})^{-1}\boldsymbol{X}^T$. The asymptotic approximation to MASE is as follows:

Theorem 5.5 *For $\boldsymbol{\alpha}$ tending to $\mathbf{0}$,*

$$\text{AMASE}(\widehat{\boldsymbol{\eta}}) = \tfrac{1}{n}\left[\boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\boldsymbol{q} + \text{tr}\{\boldsymbol{G_0}\text{Var}(\boldsymbol{U}_\eta)\boldsymbol{G_0}\}\right]$$

where $\boldsymbol{q}$ $(d \times 1)$ and $\boldsymbol{Q}$ $(d \times d)$ have entries $\boldsymbol{q}_j = \text{tr}\{\widetilde{\boldsymbol{B}}_j\text{Var}(\boldsymbol{U}_\eta)\boldsymbol{G_0}\}$,

$$\boldsymbol{Q}_{jj'} = (\widetilde{\boldsymbol{B}}_j\boldsymbol{I}_\eta\boldsymbol{X}\boldsymbol{\beta})^T(\widetilde{\boldsymbol{B}}_{j'}\boldsymbol{I}_\eta\boldsymbol{X}\boldsymbol{\beta}) + tr\{(\widetilde{\boldsymbol{B}}_j)^T\text{Var}(\boldsymbol{U}_\eta)\widetilde{\boldsymbol{B}}_{j'}\},$$

$$\widetilde{\boldsymbol{B}}_j = \check{\boldsymbol{X}}_j(\boldsymbol{X}_j^T\boldsymbol{I}_\eta\check{\boldsymbol{X}}_j)^{-1}\boldsymbol{D}_j(\boldsymbol{X}_j^T\boldsymbol{I}_\eta\check{\boldsymbol{X}}_j)^{-1}\check{\boldsymbol{X}}_j^T \quad \textit{and} \quad \check{\boldsymbol{X}}_j = (\boldsymbol{I} - \boldsymbol{S}_{\mathbf{0},[-j]}\boldsymbol{I}_\eta)\boldsymbol{X}_j.$$

The AMASE-optimal smoothing parameters are

$$\boldsymbol{\alpha}_{AMASE} = \boldsymbol{Q}^{-1}\boldsymbol{q}.$$

Approximation of the degrees of freedom

We follow Hastie and Tibshirani (1990) in defining the degrees of freedom for the jth component in a generalized additive model as $\text{df}_j = \text{tr}(\boldsymbol{G}_j\boldsymbol{F})$ where now

$$\boldsymbol{G}_j = [\mathbf{0}\cdots\mathbf{0}\ \boldsymbol{X}_j\ \mathbf{0}\cdots\mathbf{0}](\boldsymbol{X}^T\boldsymbol{F}\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T$$

with $\boldsymbol{F}$ either the observed or the expected Fisher information matrix with respect to $\boldsymbol{\eta}$.

As demonstrated in the Appendix, since we can show that

$$\frac{\mathrm{tr}(\boldsymbol{G}_j\boldsymbol{F})}{\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{F})} - 1 = -\alpha_j \frac{\mathrm{tr}[\boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{S}_{j0}\boldsymbol{F})\boldsymbol{X}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T\boldsymbol{F}\boldsymbol{B}_j\boldsymbol{F}]}{p_j + K_j - \alpha_j\mathrm{tr}\{(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{D}_j\}}, \quad (5.14)$$

where

$$\boldsymbol{B}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{D}_j(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^T \quad \text{and} \quad \boldsymbol{S}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j + \boldsymbol{A}_j)^{-1}\boldsymbol{X}_j^T, \quad (5.15)$$

the degree of freedom value df_j might be approximated by $\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{F})$. This has the computational advantage that only that part of the design matrix related to the jth covariate needs to be used.

Semiparametric models

If some of the covariates of a generalized additive model are modeled parametrically and others nonparametrically, the resulting semiparametric model can be handled similarly, as before. Assume that the first q covariates are the components of the parametric part, and that $\alpha_1 = \ldots = \alpha_q = 0$.
All results of Section 5.4.2 remain valid, and a similar decomposition as in Section 5.4.1 holds. Using the same partitioning of the design matrix as in Section 5.4.1, the smoother matrix $\boldsymbol{G}$ is now obtained as

$$\boldsymbol{G} = \boldsymbol{X}_{\mathrm{parm}}(\boldsymbol{X}_{\mathrm{parm}}^T\boldsymbol{F}\boldsymbol{X}_{\mathrm{parm}})^{-1}\boldsymbol{X}_{\mathrm{parm}}^T + \boldsymbol{R}(\boldsymbol{X}_{\mathrm{nonp}}^T\boldsymbol{F}\boldsymbol{R} + \boldsymbol{A}_{\mathrm{nonp}})^{-1}\boldsymbol{R}^T$$

where $\boldsymbol{R} = \{\boldsymbol{I} - \boldsymbol{X}_{\mathrm{parm}}(\boldsymbol{X}_{\mathrm{parm}}^T\boldsymbol{F}\boldsymbol{X}_{\mathrm{parm}})^{-1}\boldsymbol{X}_{\mathrm{parm}}^T\boldsymbol{F}\}\boldsymbol{X}_{\mathrm{nonp}}$. Separate estimators for the parametric and the nonparametric components can be obtained using the following smoother matrices:

$$\boldsymbol{G}_{\mathrm{parm}} = \boldsymbol{X}_{\mathrm{parm}}(\boldsymbol{X}_{\mathrm{parm}}^T\boldsymbol{F}\boldsymbol{X}_{\mathrm{parm}})^{-1}\boldsymbol{X}_{\mathrm{parm}}^T\{\boldsymbol{I} - \boldsymbol{X}_{\mathrm{nonp}}(\boldsymbol{R}^T\boldsymbol{F}\boldsymbol{R} + \boldsymbol{A}_{\mathrm{nonp}})^{-1}\boldsymbol{R}^T\}$$

and

$$\boldsymbol{G}_{\mathrm{nonp}} \;=\; \boldsymbol{X}_{\mathrm{nonp}}(\boldsymbol{R}^T\boldsymbol{F}\boldsymbol{R} + \boldsymbol{A}_{\mathrm{nonp}})^{-1}\boldsymbol{R}^T.$$

If we focus on the nonparametric part only, we can write the estimator of the nonparametric part of $\boldsymbol{\eta}$ as follows,

$$\boldsymbol{X}_{\mathrm{nonp}}(\boldsymbol{X}_{\mathrm{nonp}}^T\boldsymbol{W}\boldsymbol{X}_{\mathrm{nonp}} + \boldsymbol{A}_{\mathrm{nonp}})^{-1}\boldsymbol{X}_{\mathrm{nonp}}^T\boldsymbol{W}\boldsymbol{Z}\ ,$$

where now

$$\boldsymbol{W} = \boldsymbol{F}\{\boldsymbol{I} - \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T \boldsymbol{F} \boldsymbol{X}_{\text{parm}})^{-1} \boldsymbol{X}_{\text{parm}}^T \boldsymbol{F}\}. \tag{5.16}$$

5.4.3 Multiparameter models and generalized estimating equations

Wild and Yee (1996) and Yee and Wild (1996) introduced the use of vector smoothing splines in the multivariate regression and generalized estimating equations (GEE) context, where the parameter vector of interest is modeled in an additive way. The selection of the smoothing parameters and the approximation of the degrees of freedom can be obtained similarly as in the previous section, after introducing the following notation. In most cases, an estimator for the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is obtained by solving a set of estimating equations:

$$\sum_{i=1}^{n} \psi_k(\boldsymbol{Y}_i; \boldsymbol{x}_i, \boldsymbol{\theta}) = 0, \quad k = 1, \ldots, m. \tag{5.17}$$

If the likelihood of the data is known, ψ_k might be the partial derivative of the log likelihood with respect to θ_k. For example, for $m = 2$, θ_1 can be the mean and θ_2 the log variance of normally distributed data. For robust estimators, equation (5.17) might lead to M-estimators of $\boldsymbol{\theta}$ (Huber, 1981), or (5.17) can represent some set of generalized estimating equations (Liang and Zeger, 1986). In the latter case, usually the response vector is multidimensional.

In a regression model, the parameters are modeled as function of the covariates $\boldsymbol{x}_i = [\boldsymbol{x}_{1i}^T \ldots, \boldsymbol{x}_{mi}^T]^T$. Let

$$\begin{bmatrix} \theta_1(\boldsymbol{x}_{1i}) \\ \vdots \\ \theta_m(\boldsymbol{x}_{mi}) \end{bmatrix} = \begin{bmatrix} \theta_1(x_{1i1}, \ldots, x_{1id_1}) \\ \vdots \\ \theta_m(x_{mi1} \ldots, x_{mid_m}) \end{bmatrix}$$

be the parameter vector of interest. The covariate vectors $\boldsymbol{x}_{ji}$ $(j = 1, \ldots, m)$ can be the same, different for each parameter θ_j or partly overlapping with another vector $\boldsymbol{x}_{ki}$ $(k \neq j)$. For example, in toxicity studies, typically resulting in clustered binary data, a given dose might have its influence on both proportion of success (e.g., inverse logit of θ_1) and association between outcomes (represented by θ_2), but, say, individual weight of the subjects might be included only in the success probability parameter, and not in the association part.

In additive models, each of these parameter functions is, for some unknown functions f_{kj}, $(k = 1, \ldots, m; j = 1, \ldots, d_k)$,

$$\theta_k(\boldsymbol{x}_{ki}) = \beta_{k0} + \sum_{j=1}^{d_k} f_{kj}(x_{kij}).$$

For identifiability purposes, we subtract the mean of the function values from each of the f_{kj}. By introducing a regression spline design matrix $\boldsymbol{X}_{kj}$ (defined similarly as the matrix $\boldsymbol{X}_j$ in Section 5.3), we define the design matrix for θ_k as $\boldsymbol{X}_k = [\boldsymbol{X}_{k1} \cdots \boldsymbol{X}_{kd_k}]$, such that $\theta_k(\boldsymbol{x}_{ki}) = \boldsymbol{X}_k^T \boldsymbol{\beta}_k$. The design matrix $\boldsymbol{X}$ is now defined as $\boldsymbol{X} = \text{blockdiag}_{1 \le k \le m}(\boldsymbol{X}_k)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_m^T)^T$. The (observed) "Fisher Information matrix" is $\boldsymbol{J}_\beta = \boldsymbol{X}^T \boldsymbol{J}_\theta \boldsymbol{X}$ where $\boldsymbol{J}_\theta$ is a partitioned matrix with (j,k)th block

$$\boldsymbol{J}_{\theta,jk} = \underset{1 \le i \le n}{\text{diag}} \frac{\partial \psi_j}{\partial \theta_k}(\boldsymbol{Y}_i; \boldsymbol{X}_i, \boldsymbol{\theta}).$$

From this, the expected matrices $\boldsymbol{I}_\beta$ and $\boldsymbol{I}_\theta$ are easily obtained. The "score" vector $\boldsymbol{U}_\theta$ is the column vector

$$[\psi_1(\boldsymbol{Y}_1; \boldsymbol{x}_1, \boldsymbol{\theta}), \ldots, \psi_1(\boldsymbol{Y}_n; \boldsymbol{x}_n, \boldsymbol{\theta}), \ldots, \psi_m(\boldsymbol{Y}_n; \boldsymbol{X}_n, \boldsymbol{\theta})]^T$$

and $\boldsymbol{U}_\beta = \boldsymbol{X}^T \boldsymbol{U}_\theta$.

With these ingredients we can obtain an estimator for $\boldsymbol{\beta}$ via the iteratively reweighted ridge regression scheme, as presented in Section 5.4.2. To prove the results of Section 5.3 for general estimating equations, take $\boldsymbol{W} = \boldsymbol{J}_\beta$ in the Appendix. The semiparametric case can be handled by similar adjustments as explained earlier.

5.5 Proofs of Theorems

All calculations will be presented very generally, using a symmetric weight matrix $\boldsymbol{W}$. To obtain the results of Section 5.3 take $\boldsymbol{W} = \boldsymbol{I}$. For the nonparametric part of a semiparametric model, $\boldsymbol{W}$ is defined in Section 5.4.1, $\boldsymbol{S}_{[-j]}$ and $\boldsymbol{G}_j$ are defined similarly as in (5.4) and (5.7), but with the matrices $\boldsymbol{X}$ and $\boldsymbol{A}$ replaced by $\boldsymbol{X}_{\text{nonp}}$ and $\boldsymbol{A}_{\text{nonp}}$ respectively. For the generalized additive model, $\boldsymbol{W}$ is the Fisher information matrix. And for the semiparametric generalized additive model, $\boldsymbol{W}$ is defined in (5.16). If $\boldsymbol{W}$ is positive definite or a projection matrix, the following results hold.

5.5.1 Proof of Theorem 5.1

For $j = 1, \ldots, d$, we rewrite $\boldsymbol{G}$ in the following way,

$$\begin{aligned}\boldsymbol{G} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T = \boldsymbol{G}_{[-j]} + \boldsymbol{G}_j \\ &= (\boldsymbol{X}_{[-j]},\ \boldsymbol{X}_j)\begin{pmatrix} (\boldsymbol{X}_{[-j]})^T\boldsymbol{W}\boldsymbol{X}_{[-j]} + \boldsymbol{A}_{[-j]} & (\boldsymbol{X}_{[-j]})^T\boldsymbol{W}\boldsymbol{X}_j \\ \boldsymbol{X}_j^T\boldsymbol{W}\boldsymbol{X}_{[-j]} & \boldsymbol{X}_j^T\boldsymbol{W}\boldsymbol{X}_j + \boldsymbol{A}_j \end{pmatrix}^{-1}\boldsymbol{X}^T\end{aligned}$$

Using formulae for the inverse of a partitioned matrix, (see, e.g., Searle 1982, p. 260),

$$\boldsymbol{G}_{[-j]} = \boldsymbol{S}_{[-j]}[\boldsymbol{I} - \boldsymbol{W}\boldsymbol{X}_j\{\boldsymbol{X}_j^T\boldsymbol{W}(\boldsymbol{I} - \boldsymbol{S}_{[-j]}\boldsymbol{W})\boldsymbol{X}_j + \boldsymbol{A}_j\}^{-1}\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{S}_{[-j]})] \quad (5.18)$$

$$\boldsymbol{G}_j = \boldsymbol{X}_j\{\boldsymbol{X}_j^T\boldsymbol{W}(\boldsymbol{I} - \boldsymbol{S}_{[-j]}\boldsymbol{W})\boldsymbol{X}_j + \boldsymbol{A}_j\}^{-1}\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{S}_{[-j]}). \quad (5.19)$$

Note that for $\boldsymbol{W} = \boldsymbol{I}$, the expression for $\boldsymbol{G}_j$ reduces to (5.9). From (5.18) and (5.19), the recursive formula

$$\boldsymbol{G} = \boldsymbol{S}_{[-j]} + (\boldsymbol{I} - \boldsymbol{S}_{[-j]}\boldsymbol{W})\boldsymbol{G}_j \quad (5.20)$$

is easily obtained.

Next, we will use this formula to obtain the decomposition (5.9).

5.5.2 Proof of Theorem 5.2

If $d = 1$, the result is shown by extending the result of Wand (1999) to allow for a general weight matrix $\boldsymbol{W}$. Suppose that (5.9) is true for $j = d - 1$, that is,

$$\boldsymbol{S}_{[-d]} = \boldsymbol{S}_{\boldsymbol{0},[-d]} - \sum_{j=1}^{d-1}\alpha_j\widetilde{\boldsymbol{B}}_j^{[-d]} + o\left(\sum_{j=1}^{d-1}\alpha_j\widetilde{\boldsymbol{B}}_j^{[-d]}\right)\ , \quad (5.21)$$

where $\widetilde{B}_j^{[-d]}$ is defined similar to $\widetilde{B}_j$, but now in the model with all covariates except x_d.

Starting from (5.19) and using (5.21), we can approximate $\boldsymbol{G}_d$ by

$$\begin{aligned}\boldsymbol{G}_d \approx\ & \boldsymbol{X}_d\{\boldsymbol{I} + (\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I} - \boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d)^{-1}\boldsymbol{L}_{\boldsymbol{\alpha}}\}^{-1}(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{I} - \boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d)^{-1} \\ & \times\boldsymbol{X}_d^T(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{S}_{[-d]})\ ,\end{aligned}$$

where

$$\boldsymbol{L}_{\boldsymbol{\alpha}} = \alpha_d\boldsymbol{D}_d + \sum_{k=1,k\neq j}^{d-1}\alpha_k\boldsymbol{X}_j^T\boldsymbol{W}\widetilde{\boldsymbol{B}}_k^{[-d]}\boldsymbol{W}\boldsymbol{X}_j\ .$$

Then, (5.20) leads to the following approximation,

$$\begin{aligned}\boldsymbol{G} \;\approx\; & \boldsymbol{G_0}\sum_{j=1}^{d-1}\alpha_j\Big\{\widetilde{\boldsymbol{B}}_j^{[-d]} + \Big[(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\\ & \times \boldsymbol{X}_d^T\boldsymbol{W}-\boldsymbol{I}\Big]\widetilde{\boldsymbol{B}}_j^{[-d]}\boldsymbol{W}\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\boldsymbol{X}_d^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{S}_{\boldsymbol{0},[-d]})\\ & -(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\boldsymbol{X}_d^T\boldsymbol{W}\widetilde{\boldsymbol{B}}_j^{[-d]}\Big\} - \alpha_d\widetilde{\boldsymbol{B}}_d\;,\end{aligned}$$

where, from (5.19) and (5.20) with $j=d$ and $\boldsymbol{\alpha}=\boldsymbol{0}$,

$$\boldsymbol{G_0} = \boldsymbol{S}_{\boldsymbol{0},[-d]} + (\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\boldsymbol{X}_d^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{S}_{\boldsymbol{0},[-d]})\;.$$

Since

$$\begin{aligned}\widetilde{\boldsymbol{B}}_j \;=\; & \Big(\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\boldsymbol{X}_d^T\boldsymbol{W}\Big)\widetilde{\boldsymbol{B}}_j^{[-d]}\\ & \times\Big(\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\left\{\boldsymbol{X}_d^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_{\boldsymbol{0},[-d]}\boldsymbol{W})\boldsymbol{X}_d\right\}^{-1}\boldsymbol{X}_d^T\boldsymbol{W}\Big)^T\;,\end{aligned}$$

the result is proven.

5.5.3 Proof of Theorem 5.4

Using the equivalent definition of the inverse of $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+\boldsymbol{A}$, we obtain, similarly as (5.18) and (5.19), that

$$\boldsymbol{G}_j = \boldsymbol{S}_j\{\boldsymbol{I}-\boldsymbol{W}\boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_j\boldsymbol{W})\boldsymbol{X}_{[-j]}+\boldsymbol{A}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{S}_j)\}.$$

Then,

$$\begin{aligned}&\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{W}-\boldsymbol{G}_j\boldsymbol{W}) =\\ &\mathrm{tr}\left(\boldsymbol{X}_{[-j]}\{(\boldsymbol{X}_{[-j]})^T\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{S}_j\boldsymbol{W})\boldsymbol{X}_{[-j]}+\boldsymbol{A}_{[-j]}\}^{-1}(\boldsymbol{X}_{[-j]})^T(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{S}_j)\boldsymbol{W}\boldsymbol{S}_j\boldsymbol{W}\right).\end{aligned}$$

For $\boldsymbol{\alpha}\to\boldsymbol{0}$, by (5.9), we have the approximation

$$(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{S}_j)\boldsymbol{W}\boldsymbol{S}_j\boldsymbol{W}\approx\alpha_j\boldsymbol{B}_j\boldsymbol{W}\;,$$

from which, by (5.9), the numerator of (5.14) is easily obtained.

Part II

Lack of Fit Tests

Chapter 6

Testing the Fit of a Parametric Function: Order Selection Tests

6.1 Introduction

In this and in the following chapters we focus attention to testing the hypothesis that a function has a prescribed parametric form. The function of interest can be one of the parameters in a regression model (typically the mean of the response, but also its variance, or the correlation between different outcomes,...) or it can be a complete density function of which we want to investigate the goodness-of-fit. The methods for testing such hypotheses fall mostly into two categories:

1. Parametric methods designed to detect specific types of departures from the prescribed model (see Chapters 8 and 9)

2. Omnibus nonparametric methods (see Chapters 6 and 7).

Tests of the latter type are appealing in that they are consistent against virtually any departure from the hypothesized parametric model.

In this chapter we propose omnibus methods of testing function fit that are generally applicable and easy to calculate. In contrast to some other methods, for example most tests based on kernel estimators, the test statistics that we propose do not require, in the definition of the test statistic, the specification of a smoothing

parameter, such as the bandwidth in kernel based methods, which is mainly considered as a nuisance parameter. In fact, the smoothing parameter in our tests will play a key role.

The basic idea of the construction of the tests comes from Eubank and Hart (1992) who proposed the so-called *order selection* test for checking the fit of a linear model in fixed-design regression. In their test, the difference between the hypothesized linear function and the true regression is expressed as an orthogonal series. This difference can be estimated by an orthogonal series with a finite number of terms, with 0 terms corresponding to the hypothesized linear model. The order selection test proceeds by using a modified Mallow's criterion (1973) to choose the number of terms in the orthogonal series estimate, and rejecting the null hypothesis if and only if the number of terms selected is larger than 0. The tests proposed in the current chapter are analogous to the order selection test of Eubank and Hart (1992), but are applicable to a much wider range of statistical problems. One important area of application is in generalized linear models, where our ideas can be used to check the fit of a linear model for a given link function. Our procedure also leads to new tests of the hypothesis that a covariance stationary time series is white noise. An important aspect of the tests is that they can be applied in any general multiparameter model.

In recent years there has been considerable interest in using smoothing ideas to construct lack-of-fit tests (see Hart, 1997 and references therein). The existing tests that make use of smoothers are mostly in the "classical" regression context and mostly based on squared error discrepancy measures; a number of references to such tests may be found in Eubank, Hart and LaRiccia (1993) and Hart (1997). A body of work has arisen on omnibus lack-of-fit and goodness-of-fit tests that use data-driven model selection criteria; see Yanagimoto and Yanagimoto (1987), Barry and Hartigan (1990), Eubank and Hart (1992), Hart and Wehrly (1992), Kim (1992), Barry (1993), Ledwina (1994), Eubank, Hart, Simpson and Stefanski (1995), Kallenberg and Ledwina (1995, 1997), Fan (1996), Kuchibhatla and Hart (1996), Eubank (1997), Eubank, Li and Wang (1997), Inglot, Kallenberg and Ledwina (1997), Lee and Hart (1998) and Bogdan (1999). In all these works, model selection criteria play an important role in tests of the null hypothesis that a function has a prescribed form. These criteria include cross-validation, Akaike's information criterion (AIC), Bayes information criterion (BIC), and estimators of risk. We refer to Zhang (1992)

for a study of such criteria in the closely related context of model selection. Recently, Simonoff and Tsai (1999) also proposed a test based on an improved AIC criterion for additive and semiparametric "classical" regression models.

Inasmuch as orthogonal series estimators are smoothers, the tests proposed in the current chapter may also be regarded as *smoothing-based*. The first two tests we propose are applicable when the joint density of the data is known up to the function of interest and a finite number of nuisance parameters. The first method makes use of a modified version of AIC, Akaike's (1974) criterion for selecting the dimension of a statistical model. In addition, we propose another test that is a score-statistic analog of the likelihood-based AIC test. These tests use information divergence (i.e., likelihood) to measure discrepancy between the true and fitted models. Staniswalis and Severini (1991) and Staniswalis, Severini and Moschopoulos (1993) likewise use likelihood-based diagnostics to assess the fit of regression models. Quasi-likelihood functions are used by Härdle, Mammen and Müller (1998).

In general, the validity of both of these proposed tests depends upon a correct specification of the likelihood function. To address this problem, we propose a robust version of the score-based test whose asymptotic significance level is correct even when the likelihood is misspecified, or when quasi-likelihood or generalized estimating equations methods are being used.

The rest of the chapter proceeds as follows. In Section 6.2 we describe and analyze tests that assume a correct specification of the likelihood function. In Section 6.3 we propose tests that retain their validity when this assumption does not hold. Section 6.4 provides several examples of the settings in which the proposed tests may be used. Section 6.5 is devoted to a simulation study and Section 6.6 to data examples.

Most of the results of this chapter can also be found in Aerts, Claeskens and Hart (1999).

6.2 Tests in full likelihood models

In this section we introduce three tests for checking the fit of a parametric function. We begin by laying down some groundwork in Section 6.2.1. In Sections 6.2.2 and 6.2.3 we describe two versions of an AIC-based test that are equivalent and yet have different interpretations. Section 6.2.4 provides some aymptotic distribution theory

for the *AIC*-based test. Finally, in Section 6.2.5 we introduce a score-based test of function fit.

6.2.1 Preliminaries

We suppose that the observed data $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ have a joint density of the form

$$f_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \gamma(\cdot), \boldsymbol{\eta}),$$

where f_n is known up to $\gamma(\cdot)$, the function of interest, and $\boldsymbol{\eta}$, a k-dimensional vector of nuisance parameters ($k < \infty$). It follows that the log-likelihood function has the form $L(\gamma(\cdot), \boldsymbol{\eta}) = \log(f_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \gamma(\cdot), \boldsymbol{\eta}))$. Our interest is in testing the following null hypothesis:

$$H_0 : \ \gamma(\cdot) \in \mathcal{G}, \tag{6.1}$$

where $\mathcal{G} = \{\gamma(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta\}$ and Θ is a subset of a p-dimensional Euclidean space. As alternative models for $\gamma(\cdot)$ we consider sequences of approximators $\{\gamma(\cdot; \theta_1, \ldots, \theta_{p+j}) : j = 1, 2, \ldots\}$, (an example is given below) which are assumed to satisfy the following properties:

(C0a) For all possible vectors $(\theta_1, \ldots, \theta_{p+j})$ and each $j = 0, 1, \ldots$:

$$\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+j}) \equiv \gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+j}, 0).$$

(C0b) There exist approximators $\gamma(\,\cdot\,; \theta_1^r, \ldots, \theta_{p+r}^r)$, $r = 1, 2, \ldots$, that converge (in some appropriate sense) to $\gamma(\cdot)$ as $r \to \infty$.

Condition (C0a) implies that the parametric family $\mathcal{G}$ is nested within the alternative models, which in turn form a sequence of nested models having more and more parameters. The second condition is the first step to ensuring that $\gamma(\cdot)$ may be consistently estimated whether or not it is a member of $\mathcal{G}$. Since we are not primarily interested in function *estimation*, (C0b) is not necessary for our subsequent results and is included only to suggest a schema by which alternative models may be constructed. We refer to Efromovich (1996) for more on estimating functions in a quite general setting.

As just described, the approximators to the underlying function $\gamma(\cdot)$ could be quite general, but now we give a concrete example. Define, for $r = 1, 2, \ldots$, and all t in the domain of $\gamma(\cdot)$,

$$\gamma(t; \theta_1, \ldots, \theta_{p+r}) = \gamma(t; \theta_1, \ldots, \theta_p) + \sum_{j=1}^{r} \theta_{p+j} u_j(t),$$

where $u_1, u_2, \ldots$ are known functions that span some 'large' space of functions. If t is real-valued, possibilities for the u_j's are

- polynomials: $u_j(t) = t^j$, $j = 1, 2, \ldots$,
- trigonometric functions: $u_j(t) = \cos(Ajt)$, $j = 1, 2, \ldots$,
- or linear combinations of polynomials and/or trig functions that are orthogonal in some sense.

It is implicit here that the functions in $\{\gamma(\cdot) : \gamma \in \mathcal{G}\}$ are not of the form $\sum_{j=1}^{p} b_j u_{i_j}(t)$ with $i_j \in \{1, 2, \ldots, r\}$, $j = 1, \ldots, p$. If this were the case, we could simply discard $u_{i_1}, \ldots, u_{i_p}$ from $\sum_{j=1}^{r} \theta_{p+j} u_j(t)$. For example, suppose we wish to test the hypothesis that $\gamma(t)$ has the form $\theta_1 + \theta_2 t$ and we want to use polynomial alternatives. Then we could take $u_j(t) = t^{j+1}$, $j = 1, 2, \ldots$, since there is no point in including 1 or t in our set of alternative models.

6.2.2 An AIC-based test

The basic idea of our tests is to use a data-driven method of selecting a model for $\gamma(\cdot)$, and to reject the null hypothesis (6.1) if the selected model contains more than p parameters, with p being the number of parameters in H_0. In a likelihood context a popular method of model selection is AIC, Akaike's Information Criterion (1974). Given an approximator $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r})$ of $\gamma(\cdot)$, our log-likelihood is

$$L_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r}) = \log f_n\left(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r}), \boldsymbol{\eta}\right), \quad r = 0, 1, \ldots \ . \tag{6.2}$$

Suppressing the dependence of the likelihood on k and p, we write

$$\mathcal{L}_r = \sup_{\eta, \theta_1, \ldots, \theta_{p+r}} L_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r}), \quad (r = 0, 1, \ldots).$$

The AIC function is the penalized likelihood

$$AIC(r) = \mathcal{L}_r - (k + p + r), \quad r = 0, 1, \ldots,$$

in which r is the number of parameters in $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r})$ corresponding to the alternative hypothesis. An estimate of $\gamma(\cdot)$ may be obtained by choosing r to maximize $AIC(r)$ over some set of the form $\{0, 1, \ldots, r_n\}$, where r_n could either be fixed

or tending to infinity with n. For future reference we note that the maximizer of $AIC(r)$ is equal to the maximizer of $2(\mathcal{L}_r - \mathcal{L}_0) - 2r$.

A possible test of H_0 against a general alternative is to reject H_0 if the maximizer, $\hat{r}$, of $AIC(r)$ is larger than 0. Under certain regularity conditions (given in Theorem 6.1 below), the limiting level of this test (as $n \to \infty$) is about 0.29. By most standards, a type I error probability of 0.29 is quite high. To obtain control of the test level, we thus propose a modification of AIC that parallels a proposal in Eubank and Hart (1992). Define the *likelihood information criterion*, LIC, by

$$LIC(r; C_n) = 2(\mathcal{L}_r - \mathcal{L}_0) - C_n r, \quad r = 0, 1, \ldots,$$

where C_n is some constant larger than 1, and let $\hat{r}_{C_n}$ be the maximizer of $LIC(r; C_n)$. By appropriate choice of C_n, the asymptotic type I error probability of the test

$$\text{"reject } H_0 \text{ when } \hat{r}_{C_n} > 0\text{"} \tag{6.3}$$

can be any number between 0 and 1. For example, a test of asymptotic level 0.05 is obtained by using $C_n = 4.18$. (See Hart (1997), p. 178 for values of C_n leading to other test levels.)

6.2.3 An equivalent version of the test

The test described above in (6.3) rejects H_0 if and only if $LIC(r; C_n)$ is larger than 0 for some r in $\{1, \ldots, r_n\}$, which is equivalent to rejecting H_0 when $T_n \geq C_n$, with

$$T_n = \max_{1 \leq r \leq r_n} \frac{2(\mathcal{L}_r - \mathcal{L}_0)}{r}.$$

Hence, in addition to playing the role of penalty constant, C_n is a critical value of the statistic T_n. Using this version of the test one may approximate the P-value corresponding to an observed T_n by using either a large-sample distribution or the bootstrap.

The test based on T_n has a nice interpretation in terms of likelihood ratio statistics. Notice that $\mathcal{L}_r - \mathcal{L}_0$ is the log of the likelihood ratio that is used to test hypothesis (6.1) against the alternative that $\gamma(\cdot)$ has the form $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r})$. Since our test of H_0 is omnibus, T_n is not a single likelihood ratio but rather the largest of a set of *weighted* log-likelihood ratio statistics. The largest weights are

placed on the models with the fewest parameters. This has a similar effect to using a prior distribution that places higher probability on alternatives with fewer parameters.

It is worth noting that T_n's limit distribution can be guessed from the fact that T_n depends on the data only through $D_r = 2(\mathcal{L}_r - \mathcal{L}_{r-1})$, $r = 1, \ldots, r_n$. Under H_0 (and the regularity conditions in Theorem 6.1 below) the statistics $D_1, D_2, \ldots$ are asymptotically distributed as independent and identically distributed χ^2_1 random variables.

6.2.4 Asymptotic distribution theory

The data-driven smoothing parameter $\hat{r}_{C_n}$ represents the number of basis elements added to the null model parameters. Since we want to use $\hat{r}_{C_n}$ as the test statistic, we are interested in its distribution. It turns out that, under H_0, $\hat{r}_{C_n}$ has a generalized arc-sine distribution, as described by Woodroofe (1982). Our Theorem 6.1 below is, for the most part, a consequence of Woodroofe's results. The key distinction between our setting and Woodroofe's is in how these results are applied. We are interested in tests for model fit using series estimators, while Woodroofe sought to characterize the number of superfluous parameters chosen by AIC. A novel aspect of Theorem 6.1 is the penalty constant C_n that appears in the definition of the likelihood based information criterion (LIC). This constant plays the role of critical value and can be made sample size dependent. Woodroofe's results are based on the traditional AIC in which $C_n = 2$ for all n.

We need essentially the same assumptions as in Woodroofe (1982). Let $\mathbf{0}_r$ be a vector of r 0's, and let $(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)$ denote the true parameter under H_0 and $\mathcal{B}_r(\varepsilon)$ the $(k+p+r)$-dimensional sphere centered at $(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r)$ with radius ε. The notation A^c indicates complement of the set A.

(C1) For each $\varepsilon > 0$ and all $r = 0, 1, \ldots$

$$\sup\left\{\left(L_r(\boldsymbol{\delta}_r) - L_0(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)\right) \ : \ \boldsymbol{\delta}_r \in \mathcal{B}_r^c(\varepsilon)\right\} \xrightarrow{P} -\infty, \text{ as } n \to \infty.$$

(C2) There exists an $\varepsilon_0 > 0$ such that for all $r \geq 0$, $L_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ is 2 times continuously differentiable in $\mathcal{B}_r(\varepsilon_0)$, for all n sufficiently large. Define $\boldsymbol{U}_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ to be the vector of first and $-\boldsymbol{J}_{nr}(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ to be the matrix of second partial derivatives of the log-likelihood L_r, where the derivatives are with respect to the parameters $(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$.

(C3) There exist non-random, continuous matrices $\boldsymbol{J}_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ that are positive definite in $\mathcal{B}_r(\varepsilon_1)$, and such that for $r \geq 0$

$$\sup \left\{ \| \frac{1}{n} \boldsymbol{J}_{nr}(\boldsymbol{\delta}_r) - \boldsymbol{J}_r(\boldsymbol{\delta}_r) \| \ : \ \boldsymbol{\delta}_r \in \mathcal{B}_r(\varepsilon_1) \right\} \xrightarrow{P} 0, \text{ as } n \to \infty$$

for some $0 < \varepsilon_1 < \varepsilon_0$.

(C4) There exist continuous, positive definite and non-random matrices $\boldsymbol{K}_r(\cdot)$ such that, for each $r \geq 0$, $n^{-1/2}\boldsymbol{U}_r(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r)$ converges in distribution to $(\mathcal{Z}_1, \ldots, \mathcal{Z}_{k+p+r}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}_r(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0))$ as $n \to \infty$.

(C5) The sequence $\{C_n\}$ tends to $C > 1$ as n and r_n tend to ∞, and for every $\varepsilon > 0$ there is a positive integer $r_0(\varepsilon) = r_0$ such that under H_0

$$P\left(\max_{r_0 \leq r \leq r_n} \left[\frac{2\left(\mathcal{L}_r - L_0(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)\right)}{r} \right] \geq \frac{C+1}{2} \right) < \varepsilon,$$

for all n sufficiently large.

In this section, where we assume the likelihood model to be correctly specified, the matrices $\boldsymbol{J}_r(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r)$ and $\boldsymbol{K}_r(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)$ are exactly the same by Bartlett's identities (Bartlett, 1954).

Theorem 6.1 *Assume that conditions (C1)-(C5) hold. Then, under the null hypothesis (6.1),*

i) the stochastic smoothing parameter $\hat{r}_{C_n}$ as defined in Section 2.2 is asymptotically distributed according to a generalized arc sine distribution, i.e., for $P_0 = 1$ and $P_r = P(S_1 > 0, \ldots, S_r > 0)$, $r = 1, 2, \ldots$

$$\lim_{n \to \infty} P(\hat{r}_{C_n} = r) = P_r \exp\left(- \sum_{s=1}^{\infty} \frac{P(S_s > 0)}{s} \right)$$

where $S_s = \sum_{j=1}^{s} \left(V_j^2 - C\right)$, $s = 1, 2, \ldots$, and $V_1, V_2, \ldots$ are independent and identically distributed standard normal random variables.

ii) The asymptotic distribution of the test statistic T_n is given by

$$P(T \leq x) = \exp\left[- \sum_{s=1}^{\infty} \frac{P(\chi_s^2 > sx)}{s} \right],$$

where χ_s^2 has the chi-square distribution with s degrees of freedom.

We omit a proof of Theorem 6.1, since it would be essentially the same as Woodroofe's proof of his Theorems 3 and Theorem 4. The condition that r_n tend to ∞ is not necessary to obtain either a valid or a powerful test. If r_n is fixed at R, the limiting null distribution of T_n changes somewhat, but for test levels no bigger than .10, the distribution in (ii) is still an excellent approximation as long as $R \geq 5$ (Hart, 1997, p. 177-178). As noted by Woodroofe (1982), condition (C5) can be replaced in particular cases by restrictions on the rate at which r_n tends to ∞. The setting of Section 6.4.3 is one such case. Finally, it is worth noting that for test (6.3) to have reasonable power, the limiting value of C_n should be finite. Indeed, if the sequence C_n is unbounded as $n \to \infty$, it follows that, under the null hypothesis, $P(\hat{r}_{C_n} = 0)$ converges to 1, in which case test (6.3) would not be useful.

In the following theorem we establish that the test based on T_n is consistent under very general conditions.

Theorem 6.2 *Suppose the null hypothesis is false in the sense that there exist* $r_a \geq 1$, $\boldsymbol{\delta}_{r_a} = (\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r_a})$ *and a positive number* ζ *for which* $\lim_{n\to\infty} P\left[\left(L_{r_a}(\boldsymbol{\delta}_{r_a}) - \mathcal{L}_0\right)/n > \zeta\right] = 1$. *If in addition* $r_n \to \infty$ *and* C_n *tends to a finite constant, then the test with rejection region* $T_n \geq C_n$ *has power tending to 1 as* $n \to \infty$.

Proof. For all n sufficiently large $r_n > r_a$ and hence

$$P(T_n \geq C_n) \geq P\left(\frac{2(\mathcal{L}_{r_a} - \mathcal{L}_0)}{r_a} \geq C_n\right) \geq P(L_{r_a}(\boldsymbol{\delta}_{r_a}) - \mathcal{L}_0 \geq r_a C_n/2).$$

By assumption the very last probability tends to 1 as $n \to \infty$, and so the result is proven.

Consistency of the *LIC* test is mainly desirable on the assumption that the likelihood has been correctly specified but γ is not in $\mathcal{G}$. Under such a circumstance, the main condition in Theorem 6.2 is generally a consequence of condition (C0b). However, as remarked before, (C0b) is not necessary for consistency of the test. Also, the condition $r_n \to \infty$ is not necessary for consistency. Fixing r_n at, say, 10 would still yield a test that is consistent against all but rather unusual alternatives.

6.2.5 Tests based on score statistics

Both versions of the proposed *AIC*-based tests can be written in terms of the likelihood ratio statistic $2(\mathcal{L}_r - \mathcal{L}_0)$ for testing hypothesis (6.1) against the alternative

that $\gamma(\cdot)$ has the form $\gamma(\cdot;\theta_1,\ldots,\theta_{p+r})$. The Wald and score test statistic are two computationally attractive quadratic approximations of the likelihood ratio statistic. Either one could be used instead of $2(\mathcal{L}_r - \mathcal{L}_0)$. The Wald statistic (Boos, 1992) needs the "unrestricted" ML-estimators, while the score statistic only requires fitting the null model. The last property is appealing in our setting where one considers a large number of alternative models. We thus focus on score-based tests. A parallel development is possible for Wald statistics, which are, however, known not to be invariant to equivalent reparametrizations of nonlinear restrictions (see, e.g., Phillips and Park, 1988).

Analogous to the definition of *LIC*, we define the *score-based information criterion*

$$SIC(r;C_n) = \mathcal{S}_r - C_n r, \quad r = 0,1,\ldots,r_n,$$

where

$$\mathcal{S}_r = \begin{cases} 0 & \text{if } r = 0 \\ \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})^T \left(\boldsymbol{J}_{nr}(\hat{\boldsymbol{\delta}}_{r0})\right)^{-1} \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0}) & \text{if } r = 1,2,\ldots,r_n, \end{cases}$$

$\hat{\boldsymbol{\delta}}_{r0} = (\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0, \mathbf{0}_r)$ and $(\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0)$ is the null estimate, i.e. the estimate of $(\boldsymbol{\eta}, \theta_1, \ldots, \theta_p)$ gotten by maximizing the log-likelihood L_0. An apparently sensible test is that which rejects H_0 when the maximizer, $\tilde{r}_{C_n}$, of $SIC(r, C_n)$ is larger than 0. This test is equivalent to one that rejects H_0 for $\tilde{T}_n \geq C_n$, where

$$\tilde{T}_n = \max_{1 \leq r \leq r_n} \frac{\mathcal{S}_r}{r}.$$

The next theorem states that, under H_0, $\tilde{r}_{C_n}$ and $\tilde{T}_n$ have the same limiting distributions as $\hat{r}_{C_n}$ and T_n, respectively.

Theorem 6.3 *Let conditions (C1)-(C4) hold, and assume that, as $n \to \infty$, $C_n \to C > 1$ and $r_n \to \infty$. Furthermore, suppose that for every $\varepsilon > 0$ there exists a positive integer $r_0(\varepsilon) = r_0$ such that under H_0*

$$(C6) \qquad P\left(\max_{r_0 \leq r \leq r_n} \frac{\mathcal{S}_r}{r} \geq \frac{C+1}{2}\right) < \varepsilon$$

for all n sufficiently large. Then the conclusion of Theorem 6.1 remains true if $\hat{r}_{C_n}$ and T_n are replaced by $\tilde{r}_{C_n}$ and $\tilde{T}_n$, respectively.

Proof. The proof of this theorem is easily obtained by using the following result. For $k > 1$, let $\boldsymbol{Z}_k = (Z_1,\ldots,Z_k)$ have a multivariate Gaussian distribution with

mean zero and covariance matrix $\boldsymbol{\Sigma}$. Denote by $\boldsymbol{\Sigma}_j$ $(j = 1, 2, \ldots, k)$ the upper $(j \times j)$ dimensional submatrix of $\boldsymbol{\Sigma}$. Straightforward calculations show that, for each $j = 1, 2, \ldots, k$, the following random variable

$$\boldsymbol{Z}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{Z}_j - \boldsymbol{Z}_{j-1}^T \boldsymbol{\Sigma}_{j-1}^{-1} \boldsymbol{Z}_{j-1}$$

has a χ_1^2 distribution and is independent of $\boldsymbol{Z}_{j-1}^T \boldsymbol{\Sigma}_{j-1}^{-1} \boldsymbol{Z}_{j-1}$. Theorem 6.3 is now a consequence of condition (C6) *and* convergence in distribution of $\mathcal{S}_1, \ldots, \mathcal{S}_r$ to $V_1^2, V_1^2 + V_2^2, \ldots, V_1^2 + \cdots + V_r^2$ for each r, where $V_1, V_2, \ldots$ are i.i.d. standard normal random variables. The latter result is established by using the fact that, by definition, $\mathcal{S}_r$ is a quadratic form in $\mathbf{U}_r(\hat{\boldsymbol{\delta}}_{r0})_r$, the last r components of $\mathbf{U}_r(\hat{\boldsymbol{\delta}}_{r0})$.

The matrix $\boldsymbol{J}_{nr}(\cdot)$ in $\mathcal{S}_r$ is the *observed* Fisher information evaluated at the null parameter estimates. In some cases it is possible to obtain explicit expressions for the *expected* Fisher information. In such cases one could replace $\boldsymbol{J}_{nr}(\cdot)$ by the expected information evaluated at the null estimates. There is no general consensus in the literature as to which of these two approaches is better (Efron and Hinkley, 1978 and Boos, 1992). Asymptotically the two versions of $\mathcal{S}_r$ are generally equivalent to first order. An appealing aspect of using expected information is that it often leads to simpler and more readily interpretable expressions for $\mathcal{S}_r$. The last point will be illustrated by example in Sections 6.4.2–6.4.4.

In the following theorem we provide conditions under which the score-based test is consistent. We require the following condition:

(C7) There exists r such that $||\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})||/n$ is bounded away from 0 and ∞ in probability as $n \to \infty$.

Theorem 6.4 *Suppose that* $(\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0)$ *converges in probability to some vector* $(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)$, *let* $\mathcal{B}_r(\varepsilon)$ *be defined as in Section 6.2.4 in terms of this vector, and assume that conditions (C2),(C3) and (C7) hold. If in addition* $r_n \to \infty$ *and* $C_n \to C < \infty$, *then the power of the test with rejection region* $\{\max_{1 \le j \le r_n} \{\mathcal{S}_j/j\} \ge C_n\}$ *tends to 1 as* $n \to \infty$.

Proof. For all n sufficiently large we have $r_n > r$, and so
$P(\tilde{T}_n \ge C_n) \ge P(\mathcal{S}_r \ge rC_n)$. Defining

$$\xi_n = n^{-2} \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})^T \boldsymbol{J}_r^{-1}(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r) \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})$$

and

$$\Delta_n = n^{-2}\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})^T\{n\boldsymbol{J}_{nr}^{-1}(\hat{\boldsymbol{\delta}}_{r0}) - \boldsymbol{J}_r^{-1}(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r)\}\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0}),$$

we have

$$P(\mathcal{S}_r \geq rC_n) = P(\xi_n + \Delta_n \geq rC_n/n).$$

By conditions (C3) and (C7), it follows that there exists a positive number a such that $P(\xi_n \geq a) \to 1$ as $n \to \infty$. Using the consistency of the null MLE for $(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0)$ and conditions (C3) and (C7), it is easily shown that Δ_n tends to 0 in probability. Together, the last two facts imply that $P(\xi_n + \Delta_n \geq rC_n/n) \to 1$ as $n \to \infty$, and the result follows.

Condition (C7) in Theorem 6.4 represents the negation of the null hypothesis. Intuitively this condition will hold as long as there exists an r for which $(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_r)$ is not a local maximum of $\lim_{n\to\infty}$ log-likelihood$/n$.

6.3 Robust tests

Maximum likelihood methods, as described in Section 6.2, are very generally applicable. An important assumption, however, is that the true joint probability density function (pdf) $g_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ of the observations $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ belongs to a family having the form $f_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \gamma(\cdot), \boldsymbol{\eta})$, as specified by (6.2). In this section we will construct lack-of-fit tests for cases where this assumption does not necessarily hold. First we consider the case of a misspecified likelihood, and then discuss a test statistic derived from general estimating equations.

6.3.1 Likelihood misspecification

When the true data generating process is not completely described by the assumed likelihood, the likelihood model is misspecified. More information about misspecified likelihood models can be found in White (1994) and references therein. When the likelihood is misspecified, it turns out (see Theorem 6.5 below) that the asymptotic distribution of test statistics resulting from *LIC* and *SIC* can be quite complicated and, more importantly, can depend on unknown model parameters. Here we propose a robustified score statistic that has the same asymptotic distribution as that appearing in Theorem 6.1.

As before, our interest is in testing the hypothesis (6.1), where now the "true parameters" are the best approximating values of $(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ in the Kullback-Leibler sense. In other words, the true parameters minimize the Kullback-Leibler distance

$$E\left[\log g_n(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) - \log f_n(\boldsymbol{Z}_n, \ldots, \boldsymbol{Z}_n; \gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r}), \boldsymbol{\eta})\right], \tag{6.4}$$

between f_n and g_n, where, here and elsewhere, all expectations are with respect to g_n. If there exists a unique vector $(\boldsymbol{\eta}, \boldsymbol{\theta})$ such that $f_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n; \gamma(\,\cdot\,; \boldsymbol{\theta}), \boldsymbol{\eta}) = g_n(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, this is also the best approximating parameter obtained by minimizing (6.4).

Theorem 6.5 *Assume conditions (C1)-(C4) and let $r_n = R$, a fixed positive integer. The maximizers of $LIC(r; C_n)$ and $SIC(r; C_n)$ both have the limiting distribution $\mathcal{P}$, given by*

$$\mathcal{P}(r) = P(\tilde{S}_r - Cr \geq \tilde{S}_j - Cj, \text{ for all } j = 0, \ldots, R), \quad r = 0, 1, \ldots, R,$$

where

$$\tilde{S}_j = (\mathbf{0}_{k+p}, \mathcal{Z}_{k+p+1}, \ldots, \mathcal{Z}_{k+p+j}) \boldsymbol{J}_j^{-1}(\boldsymbol{\eta}^0, \boldsymbol{\theta}^0, \mathbf{0}_j)(\mathbf{0}_{k+p}, \mathcal{Z}_1, \ldots, \mathcal{Z}_{k+p+j})^T,$$

using the notation as in (C3) and (C4).

Proof. If the likelihood model is not the true likelihood of the data, the matrices $\boldsymbol{J}_r$ and $\boldsymbol{K}_r$ as defined in (C3) and (C4) are in general not the same. As a consequence, the asymptotic distribution of $\tilde{\mathcal{S}}_j$ is a weighted sum of chi-squared random variables, and (in general) no further simplification of the limiting distribution is possible.

Note that the assumptions (C3) and (C4) imply that for a correctly specified model Theorem 6.5 reduces to Theorem 6.1, since the matrices $\boldsymbol{J}_r$ and $\boldsymbol{K}_r$ are in that case exactly the same.

To overcome the practical problems with the asymptotic distribution in Theorem 6.5, we define a score statistic in analogy to those of Kent (1982) and Engle (1984). Let $\mathcal{R}_0 = 0$ and $\mathcal{R}_r =$

$$\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})_r^T \left(\left(\boldsymbol{J}_{nr}^{-1}(\hat{\boldsymbol{\delta}}_{r0}) \right)_r \left[\left(\boldsymbol{J}_{nr}^{-1}(\hat{\boldsymbol{\delta}}_{r0}) \boldsymbol{K}_{nr}(\hat{\boldsymbol{\delta}}_{r0}) \boldsymbol{J}_{nr}^{-1}(\hat{\boldsymbol{\delta}}_{r0}) \right)_r \right]^{-1} \left(\boldsymbol{J}_{nr}^{-1}(\hat{\boldsymbol{\delta}}_{r0}) \right)_r \right) \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})_r$$

for $r = 1, 2, \ldots$, where $\boldsymbol{K}_{nr}(\cdot)$ is a consistent estimator of $\boldsymbol{K}_r(\cdot)$, $\boldsymbol{D}_r(\cdot)_r$ the lower right $(r \times r)$ submatrix of a $(k+p+r)$ dimensional matrix $\boldsymbol{D}_r(\cdot)$, and $\boldsymbol{U}_r(\cdot)_r$ the last r components of the length $k+p+r$ vector $\boldsymbol{U}_r(\cdot)$. Now define the *robust score-based information criterion*

$$RIC(r; C_n) = \mathcal{R}_r - C_n r, \quad r = 0, 1, \ldots, r_n.$$

We may then reject H_0 whenever the maximizer of $RIC(r; C_n)$ exceeds 0. This test is equivalent to one that rejects H_0 whenever

$$\bar{T}_n = \max_{1 \le r \le r_n} \frac{\mathcal{R}_r}{r} \ge C_n.$$

Under H_0 and appropriate regularity conditions, the statistic $\bar{T}_n$ will have the same limit distribution as that of T_n in Theorem 6.1. Moreover, this result will not in general depend upon a correct specification of the likelihood. Rather than stating a theorem to this effect, we will use simulation in Section 6.5 to investigate the adequacy of the large sample distribution for finite n.

In certain settings where series estimators are used to approximate γ, the matrices $\boldsymbol{J}_{nr}(\hat{\boldsymbol{\delta}}_{r0})$ can be greatly simplified by an orthogonalization process, an example of which will be given in Section 6.4.4. Finally, since the statistic $\bar{T}_n$ only requires the score equations, and not knowledge of the likelihood itself, the *RIC* test will also be applicable in more general statistical models, which we consider next.

6.3.2 Estimating equations

Parameter estimators can be obtained by solving more general estimating equations than likelihood equations. If the parameter of interest is the (conditional) mean of the response variables, the idea of solving a set of score equations $\boldsymbol{U}_r(\cdot) = 0$ in likelihood models is generalized to the construction of quasi-likelihood equations (Wedderburn, 1974), and to the multivariate version, the generalized estimating equations (GEE) (Liang and Zeger, 1986). Other frequently used estimating equations are found in the context of M-estimation (Huber, 1981), resulting in robust regression models.

In the absence of a likelihood function, it is clear that an *LIC*-based test can no longer be constructed. Instead, an *RIC* criterion may be defined as follows. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be independent observations, and consider the sets of estimating

equations $\sum_{i=1}^{n} \boldsymbol{\psi}_r(\boldsymbol{Z}_i; \boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r}) = \mathbf{0}_{p+k+r}$, where $\boldsymbol{\psi}_r$ is a $p+k+r$ vector of statistics, $r = 0, 1, \ldots$. Let $(\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0)$ be the solution to the set of equations corresponding to $r = 0$, define $\widehat{\boldsymbol{\delta}}_{r0} = (\hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0, \mathbf{0}_r)$, and take $\boldsymbol{\xi}_r$ to be the length $p+k+r$ vector equal to $\sum_{i=1}^{n} \boldsymbol{\psi}_r(\boldsymbol{Z}_i; \hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0, \mathbf{0}_r)$. Now define $\mathcal{R}_0 = 0$ and

$$\mathcal{R}_r = (\boldsymbol{\xi}_r)_r^T \left\{ \left(\check{\boldsymbol{J}}_{nr}^{-1}(\widehat{\boldsymbol{\delta}}_{r0}) \right)_r \left[\left(\check{\boldsymbol{J}}_{nr}^{-1}(\widehat{\boldsymbol{\delta}}_{r0}) \check{\boldsymbol{K}}_{nr}(\widehat{\boldsymbol{\delta}}_{r0}) \check{\boldsymbol{J}}_{nr}^{-1}(\widehat{\boldsymbol{\delta}}_{r0}) \right)_r \right]^{-1} \left(\check{\boldsymbol{J}}_{nr}^{-1}(\widehat{\boldsymbol{\delta}}_{r0}) \right)_r \right\} (\boldsymbol{\xi}_r)_r$$

for $r = 1, 2, \ldots$, where $\check{\boldsymbol{J}}_{nr}(\cdot)$ is a $(p+k+r) \times (p+k+r)$ matrix of partial derivatives of $\boldsymbol{\psi}_r$ and

$$\check{\boldsymbol{K}}_{nr}(\widehat{\boldsymbol{\delta}}_{r0}) = \sum_{i=1}^{n} \boldsymbol{\psi}_r(\boldsymbol{Z}_i; \hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0, \mathbf{0}_r) \boldsymbol{\psi}_r(\boldsymbol{Z}_i; \hat{\boldsymbol{\eta}}_0, \hat{\boldsymbol{\theta}}_0, \mathbf{0}_r)^T.$$

We may now define a criterion and a statistic in analogy to RIC and $\bar{T}_n$ of Section 6.3.1. Presumably one can develop parallel distribution theory for these statistics, though we do not pursue this problem here.

6.4 Applications

The tests of model fit we have introduced have a wide variety of applications. In this section we will illustrate how they can be used in several situations. In Sections 6.4.1 and 6.4.2 we consider goodness of fit and white noise tests, respectively. In regard to regression, we consider Gaussian likelihoods in Section 6.4.3 and tests for homoscedasticity in Section 6.4.4. Section 6.4.5 describes an application to longitudinal data.

6.4.1 Testing goodness of fit

Suppose we have independent observations $X_1, \ldots, X_n$ having unknown density $f(x)$. Writing $f(x) = C_\gamma \exp(\gamma(x))$ we want to test the hypothesis $H_0 : \gamma(x) \in \mathcal{G}$ for some parametric family of functions $\mathcal{G} = \{\gamma(x; \theta_1, \ldots, \theta_p) : (\theta_1, \ldots, \theta_p) \in \Theta\}$. Typical examples of interest are testing for exponentiality and testing for normality.

Consider approximators of $\gamma(x)$ of the form

$$\gamma(x; \theta_1, \ldots, \theta_p) + \sum_{j=1}^{r} \theta_j u_j(x).$$

We may then proceed to test H_0 using either *LIC* or *SIC* as described in Section 6.2. This test seems to be an interesting alternative to the data-driven smooth tests proposed by Kallenberg and Ledwina (1997).

6.4.2 White noise tests

Consider a sequence of time-ordered observations $X_1, \ldots, X_n$, and suppose we wish to test the hypothesis that these observations are independent, i.e., that they are a white noise sequence. A quite general model for dependence is that which assumes $\{X_t\}$ to have the linear form

$$X_t = \mu + \sum_{i=0}^{\infty} \xi_i \epsilon_{t-i}, \quad t = 1, 2, \ldots, \tag{6.5}$$

where the ϵ_j's are i.i.d. random variables with mean 0 and finite variance σ^2 and $\sum_{i=0}^{\infty} \xi_i^2 < \infty$. If model (6.5) holds, the process $\{X_t\}$ is covariance stationary in the sense that $\mathrm{Cov}(X_i, X_{i+j})$ depends on j but not i. If each ξ_i other than ξ_0 is 0, then the null hypothesis of independence is true.

Model (6.5) is characterized by the probability distribution of ϵ_i and the power spectrum

$$\gamma(\omega) = \left| \sum_{j=0}^{\infty} \xi_j \exp(2\pi i j \omega) \right|^2, \quad 0 \le \omega \le 1.$$

A popular way of approximating such spectra (Newton, 1988) is to use autoregressive representations, which imply that the process $\{X_t\}$ has the form

$$X_t - \mu = \sum_{i=1}^{r} \theta_i (X_{t-i} - \mu) + \epsilon_t, \quad t = 1, 2, \ldots$$

for some $r \ge 0$, where the ϵ_t's are i.i.d. with mean 0 and finite variance and the zeroes of $1 - \theta_1 z - \cdots - \theta_r z^r$ are outside the unit circle in the complex plane. The case $r = 0$ corresponds to the null hypothesis of independence.

Suppose we entertain the autoregressive model along with the assumption that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. If we take $X_i = \mu$ for $i < 1$ (a common practice in time series analysis), then the joint density of the observations can be written as in Section 6.2.1 with $\boldsymbol{\eta} = (\mu, \sigma^2)$ and γ the spectral density of the process $\{X_t\}$. The white noise hypothesis may be tested by applying the method of Section 6.2.2. The quantity *LIC* in this case has the form

$$LIC(r; C_n) = n(\log \hat{\sigma}_0^2 - \log \hat{\sigma}_r^2) - C_n r, \quad r = 0, 1, \ldots, r_n,$$

where $\hat{\sigma}_r^2$ denotes the maximum likelihood estimate of σ^2 for the order r autoregressive model. By contrast, *SIC*, based on expected Fisher information, takes the form

$$SIC(r; C_n) = \left[n \sum_{j=1}^{r} \frac{\hat{\rho}_j^2}{(1 - j/n)} \right] - C_n r, \quad r = 0, 1, \ldots, r_n$$

where $\hat{\rho}_j$ is the sample autocorrelation function, i.e.,

$$\hat{\rho}_j = \frac{\sum_{i=1}^{n-j} (X_i - \bar{X})(X_{i+j} - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \quad j = 1, \ldots, n-1.$$

It is noteworthy that application of the above *LIC*- and *SIC*-based tests is not limited to the case of Gaussian errors. The limit distribution of each test statistic is invariant under general moment conditions on $X_1, \ldots, X_n$. Another interesting aspect of this example is that both tests are apparently reasonable tests of white noise, but for different reasons. The quantity *LIC* is just a penalty-modified version of Akaike's criterion for selecting autoregressive order. By contrast, *SIC* is closely related to criteria for selecting the order of a *moving average* model for the time series $X_1, X_2, \ldots$ (Hart, 1997, pp. 242-244). One thus anticipates that the power properties of the two tests will be different. The *LIC* test would be expected to have better power than the *SIC* test against autoregressive alternatives to white noise, while *SIC* should tend to have better power when the process is of moving average type.

Existing knowledge of *AIC* gives us reason to believe that, in general, the maximizer of $LIC(\cdot\,; 2)$ is close to the dimension of a best approximating function among $\gamma(\cdot\,; \theta_1, \ldots, \theta_{p+r})$, $r = 0, 1, \ldots$. One might expect that the same is true of the maximizer of $SIC(\cdot\,; 2)$. However, the example in this section shows that this is not always the case. Suppose that the true process is first order autoregressive (AR(1)) with $\theta_1 \neq 0$. Then the maximizer of *SIC* will tend to be larger than $p(= 1)$; indeed it will tend to ∞ as $n \to \infty$. This is because all the autocorrelations of the underlying AR(1) process are nonzero, and hence each $\hat{\rho}_j$ estimates a nonzero quantity. Furthermore, the maximizer of *SIC* will tend to be larger the closer θ_1 is to 1.

6.4.3 Regression based on Gaussian likelihood

Suppose that $Y_1, \ldots, Y_n$ are independent observations such that $Y_i \sim \mathcal{N}(\gamma(x_i), \eta)$, $i = 1, \ldots, n$, where $x_1, \ldots, x_n$ are fixed design points lying in the interval $[0, 1]$. Let

$\gamma_1, \ldots, \gamma_p$ be known functions and consider testing the null hypothesis

$$H_0 : \gamma(x) = \sum_{i=1}^{p} \theta_i \gamma_i(x), \quad 0 \le x \le 1, \tag{6.6}$$

for some set of parameters $\theta_1, \ldots, \theta_p$. Now suppose that the set of functions $\{u_1, u_2, \ldots\}$ is complete for the functions that are continuous on $[0,1]$. (We also assume that $\gamma_1, \ldots, \gamma_p$, $u_1, \ldots, u_s$ are linearly independent for every finite s.) As our approximators for γ we use

$$\gamma(x; \theta_1, \ldots, \theta_{p+r}) = \sum_{j=1}^{p} \theta_j \gamma_j(x) + \sum_{j=1}^{r} \theta_{p+j} u_j(x), \quad r = 1, 2, \ldots .$$

The test of (6.6) based on LIC rejects H_0 for large values of the statistic

$$T_n = \max_{1 \le r \le r_n} \frac{n(\log SSE_p - \log SSE_{p+r})}{r},$$

where SSE_{p+r} denotes the familiar error sum of squares from the least squares analysis of the linear model $\sum_{j=1}^{p} \theta_j \gamma_j(x) + \sum_{j=1}^{r} \theta_{p+j} u_j(x)$.

We now consider a version of the SIC-based test from Section 6.2.5. Here we use the expected Fisher information evaluated at the maximum likelihood estimates under H_0. This test rejects H_0 for large values of

$$\tilde{T}_n = \frac{n}{SSE_p} \max_{1 \le r \le r_n} \frac{SSE_p - SSE_{p+r}}{r}.$$

This statistic is identical to that proposed by Eubank and Hart (1992) for testing the fit of linear models. When the Gaussian likelihood assumption is correct and the null hypothesis is true, the exact distribution of $\tilde{T}_n$ (for any n) can be obtained (Hart, 1997, pp. 177-180). For non-Gaussian errors, Eubank and Hart (1992) provide conditions such that $\tilde{T}_n$ converges in distribution (under H_0) to a random variable with the distribution given in (ii) of Theorem 6.1. Hence, this is another example of where the Gaussian-likelihood score-based test can be asymptotically valid even though the true likelihood is not Gaussian.

How do the likelihood and score-based statistics compare in this example? Under certain conditions the two tests will be asymptotically equivalent. Suppose that r_n is arbitrarily large but fixed. Then by using a Taylor series expansion of $\log(SSE_{p+r}/SSE_p)$ about 1, it is easy to see that T_n and $\tilde{T}_n$ are the same to first order, as $n \to \infty$. However, when $r_n = n - p - 1$, we have

$$T_n \ge \left(\frac{n}{n-p-1} \right) \left[\log(n-p) + \log(SSE_p/(n-p)) - \log SSE_{n-1} \right],$$

and, in general, the last quantity tends to infinity in probability as $n \to \infty$. Eubank and Hart (1992) show that in certain cases $\tilde{T}_n$ has a proper limiting distribution even if we take $r_n = n - p$. The fact that $\tilde{T}_n$ is less sensitive than T_n to the choice of r_n seems to be a clear advantage for the former statistic.

6.4.4 Tests for homoscedasticity

In classical linear regression one assumes that the Gaussian error terms have constant variance. More generally we have the model $Y_i \sim \mathcal{N}(\eta(x_i), \gamma(x_i)), \quad i = 1, \ldots, n$, where $Y_1, \ldots, Y_n$ are independent and $x_1, \ldots, x_n$ are fixed design points on, say, $[0, 1]$. The tests based on LIC and SIC are in this case tests for homoscedasticity where now the vector $\boldsymbol{\eta}^T = (\eta_1, \ldots, \eta_k)$ in the mean function $\eta(x) = \sum_{\ell=1}^{k} \eta_\ell \psi_\ell(x)$ is considered the nuisance parameter and the function of interest is

$$\gamma(x; \theta_1, \ldots, \theta_r) = \exp\left(\theta_1 + \sum_{j=1}^{r} \theta_{1+j} u_j(x)\right).$$

We consider k fixed and assume that the orthogonality property $\sum_{i=1}^{n} \psi_j(x_i)\psi_\ell(x_i) = n\delta_{j\ell}$ holds for each $j, \ell = 0, \ldots, k$. Next, suppose that the set $\{u_1, u_2, \ldots\}$ is complete and orthogonalized such that $\sum_{i=1}^{n} u_j(x_i)u_\ell(x_i) = n\delta_{j\ell}$, for $j, \ell = 0, \ldots, r$ for every r, where $u_0 = 1$. The orthogonalization of the ψ_j's and u_j's can be done using the Gram-Schmidt process (see, e.g., Rao, 1973).

We focus attention on the SIC-based test (using expected Fisher information) for the null hypothesis of homoscedasticity, i.e.,

$$H_0 \ : \ \gamma(x) = \theta_1, \quad 0 \le x \le 1$$

for some value of $\theta_1 > 0$. Define the residuals $e_i = Y_i - \sum_{\ell=1}^{k} \hat{\eta}_\ell \psi_\ell(x_i)$, $i = 1, \ldots, n$, where $\hat{\eta}_1, \ldots, \hat{\eta}_k$ are the ordinary least squares estimates of $\eta_1, \ldots, \eta_k$, and let $\hat{\sigma}_0^2 = (\sum_{i=1}^{n} e_i^2)/n$. The SIC test rejects H_0 for large values of

$$\tilde{T}_n = \frac{n}{2\hat{\sigma}_0^4} \max_{1 \le r \le r_n} \frac{1}{r} \sum_{j=1}^{r} \hat{\phi}_j^2,$$

where $\hat{\phi}_j = (\sum_{i=1}^{n} e_i^2 u_j(x_i))/n$ for $j = 1, \ldots, r$. This test statistic closely resembles the statistic suggested by Liaw (1997). The only difference is that our statistic uses $2\hat{\sigma}_0^4$ to estimate the null variance of e_i^2. This is a consistent estimator when assuming Gaussian errors but not in general, indicating that $\tilde{T}_n$ is not robust against

nonnormality. However, the *RIC* based statistic from Section 6.3 is identical to Liaw's statistic, which Liaw (1997) shows to be asymptotically valid even for non-Gaussian errors and $r_n = n - k - 1$.

The simple form taken by $\tilde{T}_n$ and $\bar{T}_n$ in this case is due to using the Gram-Schmidt process in defining the functions ψ_j and u_j, and also to using expected, rather than observed, Fisher information.

6.4.5 Longitudinal data

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^T$ and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})^T$ denote a vector of n repeated observations and a vector of covariates, respectively, measured on the ith subject, $i = 1, \ldots, m$. A possible Gaussian model for longitudinal data is $\boldsymbol{Y} = \gamma(\boldsymbol{X}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}^T = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_m^T)$, $\boldsymbol{X}^T = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_m^T)$, $\gamma(\boldsymbol{X}; \boldsymbol{\theta})$ is the nm vector obtained by applying $\gamma(\cdot; \boldsymbol{\theta})$ elementwise to X, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ a block-diagonal matrix having $n \times n$ blocks $\boldsymbol{\Sigma}_i$ representing the covariance matrix of $\boldsymbol{Y}_i$, $i = 1, \ldots, m$. The parameter vector $\boldsymbol{\theta}$ can be estimated by maximum likelihood.

This Gaussian model can be extended in different ways to GLMs for longitudinal responses (see Diggle, Liang and Zeger, 1994). A so-called marginal approach which does not require specification of the entire likelihood is the GEE-approach, introduced by Liang and Zeger (1986) and Zeger and Liang (1986). In this approach, the function of interest is $\gamma(x) = E(Y_{ij}|X_{ij} = x)$ which can be approximated as before by

$$\gamma(x; \theta_1, \ldots, \theta_{p+r}) = h^{-1}\left(\sum_{i=1}^{p} \theta_i \gamma_i(x) + \sum_{j=1}^{r} \theta_{p+j} u_j(x)\right)$$

for some (specified) link function h. A GEE estimator of $(\theta_1, \ldots, \theta_{p+r})$ is the solution to

$$\boldsymbol{U}(\theta_1, \ldots, \theta_{p+r}) = \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{\Sigma}_i^{-1} \left(\boldsymbol{Y}_i - \gamma(\boldsymbol{X}_i; \theta_1, \ldots, \theta_{p+r})\right) \tag{6.7}$$

where

$$\boldsymbol{D}_i = \partial \gamma(\boldsymbol{X}_i; \theta_1, \ldots, \theta_{p+r}) / \partial(\theta_1, \ldots, \theta_{p+r}),$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{J}_i^{1/2} \boldsymbol{R}_i \boldsymbol{J}_i^{1/2},$$

$$\boldsymbol{J}_i = \mathrm{diag}(\mathrm{Var}(\epsilon_{i1}) \ldots, \mathrm{Var}(\epsilon_{in})),$$

and $\boldsymbol{R}_i$ is an $n \times n$ correlation matrix. It is assumed that $\mathrm{Var}(\epsilon_{ij}) = \nu(E(Y_{ij}))\eta_1$, where ν is a known function and η_1 a scale parameter. The "working" correlation

matrix $\boldsymbol{R}_i$ may also depend on a vector of unknown parameters $(\eta_2, \ldots, \eta_k)$. A solution $(\hat{\theta}_1, \ldots, \hat{\theta}_{p+r})$ is obtained by iterating between solutions of (6.7) and consistent estimators of $\boldsymbol{\eta}$.

The ideas of Section 6.3.2 could be applied here to obtain a test of the null hypothesis

$$H_0 : \gamma(x) = h^{-1}(\sum_{i=1}^{p} \theta_i \gamma_i(x)).$$

Robust score statistics $\mathcal{R}_r$ may be constructed and a test statistic defined as $\max_{1 \le r \le r_n} \mathcal{R}_r / r$, the largest of a set of weighted robust score statistics. In the context of cluster correlated data, Rotnitzky and Jewell (1990) examined the asymptotic properties of a statistic analogous to $\mathcal{R}_r$.

6.5 Simulation study

This section is intended to illustrate certain aspects of the finite sample behavior of our proposed tests. These include level properties of the various tests, a power comparison of the proposed nonparametric tests for correctly specified models, and a power comparison with a classical parametric test.

6.5.1 Type I error probabilities

When the assumed likelihood model is incorrect, the distributional properties as formulated in Theorems 6.1 and 6.3 are no longer valid. Therefore it is expected that the *LIC*- and *SIC*-based tests will have inaccurate levels. We will illustrate this by simulations for generalized linear models (GLM), see e.g. McCullagh and Nelder (1989). Gourieroux, Monfort and Trognon (1984) show that when a generalized linear model is used, the mean can still be consistently estimated even if the true pdf does not belong to the specified class of exponential family models.

In a first setting we generated data from independent, zero-inflated Poisson response variables $Y_i = U_i \times V_i$, $i = 1, \ldots, n$, where

$$U_i \sim \text{Poisson}(\lambda(x_i)),$$

$$\lambda(x) = \exp(2(x + 0.2)(x - 0.3)),$$

$$V_i \sim \text{Bin}(1, \pi),$$

		Polynomials			Cosines		
n	π	LIC	SIC	RIC	LIC	SIC	RIC
30	1	4.2	3.8	4.4	4.2	3.6	4.8
	0.9	9.5	7.7	4.4	9.1	7.8	4.4
	0.8	15.5	12.2	5.4	14.5	11.4	5.4
	0.7	22.1	15.7	5.9	21.2	15.0	6.0
	0.6	32.4	20.8	5.4	30.8	19.7	5.4
60	1	5.4	5.1	5.8	5.6	5.4	5.7
	0.9	11.5	10.0	6.5	10.6	9.5	6.4
	0.8	16.1	14.7	5.9	15.9	14.7	6.2
	0.7	21.4	19.0	5.3	20.6	19.4	5.1
	0.6	29.8	25.3	5.5	27.6	23.9	5.3

Table 6.1: Simulation results for zero-inflated Poisson data

and U_i and V_i are independent. The design points $x_1, \ldots, x_n$ were chosen to be equidistant between zero and one. The null hypothesis

$$H_0 : \ln(\lambda(x)) = \theta_0 + \theta_1 x + \theta_2 x^2$$

is tested at a 5% level of significance. Table 6.1 shows estimated type I error probabilities (as %) based on 1000 simulation runs.

Note first that the case $\pi = 1$ corresponds to a correctly specified model. Here, each of the three tests maintain the nominal 5% significance level. As π decreases from 1 (i.e., misspecified models), the type I error probabilities of the LIC and SIC tests increase. In contrast, the RIC test nicely holds the prescribed level of 5%.

As a second example we consider overdispersed binomial data, generated from a beta-binomial model $BB(N, \pi(x), \rho)$ where $N = 13$,

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -5 + bx$$

and

$$\ln\left(\frac{1+\rho(x)}{1-\rho(x)}\right) = c.$$

			Polynomials		Cosines	
n	b	c	*SIC*	*RIC*	*SIC*	*RIC*
30	3	0	4.8	6.1	6.2	4.2
		0.05	8.0	4.0	9.9	3.9
		0.10	11.8	3.4	14.5	3.5
		0.20	19.6	2.1	24.0	2.4
		0.30	28.2	2.5	34.1	2.4
	5	0	4.4	6.4	5.0	4.8
		0.05	9.1	5.6	10.7	4.5
		0.10	14.6	5.6	15.3	4.5
		0.20	25.2	5.6	27.1	4.5
		0.30	34.6	3.4	40.2	3.8
60	3	0	4.9	6.7	4.5	5.6
		0.05	10.1	5.8	10.8	4.5
		0.10	13.6	5.1	14.5	4.8
		0.20	22.8	4.1	25.4	4.5
		0.30	32.4	3.4	34.6	4.2
	5	0	5.4	6.8	6.2	5.6
		0.05	8.8	5.9	10.7	5.7
		0.10	14.1	5.2	14.6	4.6
		0.20	23.8	4.4	28.0	4.6
		0.30	35.6	4.4	39.4	5.2

Table 6.2: Simulation results for beta-binomial data

Here, $\pi(x)$ represents success probability, $\rho(x)$ the intracluster correlation, N the cluster size, and x a covariate. (For more details on the beta-binomial model, see Skellam, 1948 and Section 2.6.1). Using the binomial likelihood in *LIC*, *SIC* and *RIC*, we tested the hypothesis that the logit of $\pi(x)$ is linear in x at a 5% level of significance. If $\rho \equiv 0$, the beta-binomial model reduces to the binomial model (no overdispersion), but for $\rho > 0$ the binomial likelihood is incorrect. Simulated type I error probabilities (as %) based on 1000 simulation runs are shown in Table 6.2.

When the model is correctly specified ($c = 0$), both score tests have empirical levels close to the nominal level. However, even a small amount of intracluster correlation causes highly inflated type I error probabilities for the *SIC* test. The *RIC*-test nicely corrects for the model misspecification and has the correct type I error rate. The *LIC* test was not considered in this part of the simulation because of numerical problems with the maximum likelihood estimates.

6.5.2 Power comparisons when the likelihood is correctly specified

Here we address the question: "How does the power of an *RIC* test compare with that of *LIC* and *SIC* tests for correctly specified models?" In general, one might expect the *RIC* test to have less power, but, since there are many ways to specify an alternative model, there is probably no simple answer. To investigate this question, we generated Poisson data $Z_i = (Y_i, x_i)$, $i = 1, \ldots, n = 30$, where the Y_i's were independent with

$$Y_i \sim \text{Poisson}(\lambda(x_i)),$$

$$\ln(\lambda(x)) = -1 + 4x^a, \tag{6.8}$$

and the x_i's were equally spaced in [0,1]. We tested the following null hypothesis at a 5% level of significance:

$$H_0 : \ln(\lambda(x)) = \theta_0 + \theta_1 x.$$

Table 6.3 shows the simulated rejection probabilities (as %) based on 1000 simulation runs.

In this example, losses in power for the *RIC* test are significant, meaning that the unnecessary use of the robust procedure is quite costly.

	Polynomials			Cosines		
a	LIC	SIC	RIC	LIC	SIC	RIC
1.0	5.5	5.4	5.1	6.0	5.5	5.4
1.2	9.8	10.7	6.5	6.8	8.0	6.0
1.4	16.3	20.5	8.9	11.0	11.8	7.2
1.6	26.4	33.6	13.0	18.1	22.4	10.4
1.8	39.8	45.9	19.7	29.9	34.9	15.6
2.0	50.7	57.4	26.5	38.6	44.7	22.4

Table 6.3: Empirical power for Poisson model with $\ln(\lambda(x))$ *as in (6.8).*

In a second example, data were generated from the alternative

$$\lambda(x) = \exp\{-1 + 4x + a\ln(x+1)\}, \tag{6.9}$$

with n=30. The simulated rejection probabilities (as %) over 1000 simulation runs are shown in Table 6.4. For this alternative, the RIC test has either comparable or somewhat *higher* power than the LIC and SIC tests, which are again comparable to each other. Obviously then, one does not always pay a price when using a robust test unnecessarily.

	Polynomials			Cosines		
a	LIC	SIC	RIC	LIC	SIC	RIC
0.0	5.5	5.4	5.1	6.0	5.5	5.4
1.0	4.8	3.7	6.2	4.9	4.5	4.4
3.0	9.8	8.4	12.9	9.2	8.1	9.0
5.0	39.1	36.4	42.6	32.9	30.8	31.2
7.0	93.5	92.5	92.5	86.3	85.5	84.3

Table 6.4: Empirical power for Poisson model with $\ln(\lambda(x))$ *as in (6.9).*

6.5.3 Comparison with a parametric test

Here we compare the *LIC* test with a likelihood ratio test in the context of longitudinal data. One would expect that no test will have substantially higher power than the likelihood ratio test based on a correctly specified alternative. We used a special case of the model in Section 6.4.5 in which Gaussian observations $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{i5})$ were generated from 15 (independent) experimental units (so $n = 5$ and $m = 15$). We took $X_{ij} = j$ for all i and $j = 1, \ldots, 5$, which corresponds to using time as the covariate and observing all experimental units at the same time points. The mean function was taken to be

$$\gamma(t) = \beta t^3$$

for selected values of β, and for each i, the elements of $\boldsymbol{\Sigma}_i$ were of the form $\sigma^2 \exp(-\phi(j-k)^2)$, $j, k = 1, \ldots, 5$.

		Order selection tests			Likelihood-ratio test
σ^2, ϕ	β	T_n^P	T_n^C	$T_n^{\max}$	LRT
	0	4.6	4.6	2.6	4.8
4, 0.1	0.0075	36.0	9.0	25.4	38.4
	0.01	57.4	15.4	46.8	61.6
	0	4.8	5.2	2.4	5.2
4, 0.5	0.0075	8.6	5.6	5.0	10.0
	0.01	11.2	5.4	8.0	12.6
	0	4.6	4.6	2.6	4.8
0.5, 0.1	0.0075	99.8	60.0	99.0	99.8
	0.01	100	83.6	99.8	100
	0	4.8	5.2	2.4	5.2
0.5, 0.5	0.0075	32.8	10.0	25.4	36.4
	0.01	55.6	15.8	45.6	57.6

Table 6.5: Empirical powers for three LIC-based tests and a likelihood ratio test

Taking β equal to 0, 0.0075 and 0.01, the null hypothesis of linearity

$$H_0 : \gamma(t) = \beta_0 + \beta_1 t$$

was tested at nominal significance level 5%. Let T_n^P and T_n^C be the versions of T_n (Section 2.3) that use polynomial and cosine bases, respectively. We then considered four test statistics: T_n^P, T_n^C, $\max(T_n^P, T_n^C)$ and a likelihood ratio in which the alternative hypothesis was

$$H_1 : \gamma(t) = \beta_0 + \beta_1 t + \beta_3 t^3.$$

Table 6.5.3 shows estimated rejection probabilities based on 500 samples. Maximum likelihood estimates were computed by the Oswald software in S-Plus (Smith, 1997). Since the true alternative is a polynomial, it is not surprising that the LIC test based on polynomials performs substantially better than the one based on cosines. In fact, the power of T_n^P is comparable to that of the parametric LRT.

In practice an optimal choice of basis is not a trivial problem, and hence the results for $T_n^{\max} = \max(T_n^P, T_n^C)$ are of interest. This statistic was compared with a critical value of 5.236, the 97.5(th) percentile of T_n's limit distribution. Bonferroni's inequality implies that the corresponding (asymptotic) level of the $T_n^{\max}$-based test is no more than 5%. Interestingly, the empirical power of the "Bonferroni" test is not much lower than that of T_n^P. Note also that the empirical level for $T_n^{\max}$ is close to 2.5%, suggesting that proper adjustment of the critical value will make the power of $T_n^{\max}$ close to that of T_n^P.

6.6 Data examples

Here we apply our proposed tests to some of the data sets introduced in Chapter 1.

6.6.1 The low-iron rat teratology data

Let $\pi(x)$ denote the expected proportion of dead foetuses for female rats whose hemoglobin level is x, and suppose we wish to test the following null hypothesis:

$$H_0 : \text{logit}(\pi(x)) = \theta_1 + \theta_2 x, \quad \text{for each } x. \tag{6.10}$$

We will consider two probability models for these data, both of which condition on the observed n_i's. A naive model is to assume that the $y_1, \ldots, y_{58}$ are independent

		Binomial		beta-binomial	
	Criterion	Value	P-value	Value	P-value
cosines	LIC	10.45	—	4.00	0.06
	SIC	15.56	—	8.23	0.004
	RIC	2.01	0.29	1.93	0.31
polynomials	LIC	10.12	—	6.74	0.01
	SIC	19.12	—	9.85	0.002
	RIC	1.77	0.37	3.18	0.10

Table 6.6: Test statistics for the low-iron rat teratology data

with $y_i \sim \text{Bin}(n_i, \pi(x_i))$, $i = 1, \ldots, 58$. It turns out that this model is untenable, as the data clearly exhibit more variability than would be expected under the binomial model. This overdispersion is undoubtedly due to correlation among foetuses belonging to the same litter; see Figure 2.1. Another model which allows for such correlation is the beta-binomial model. Using either of these models we may obtain a valid test of H_0 by using the robustified score test of Section 6.3.1. We will also compute the LIC- and SIC-based statistics to see how they differ between the two models.

Test statistics for hypothesis (6.10) are calculated using both polynomial and cosine alternatives. First we transform the design points to the interval $(0, 1)$. In the notation of Section 6.2.1 we then consider

$$u_j(x) = x^{j+1} \quad \text{and} \quad u_j(x) = \cos(\pi j x),$$

$j = 1, 2, \ldots$. In the beta-binomial model the correlation was modeled as a straight line function of x, a model which is, referring to the results of Chapter 2, found to be reasonable. The statistics are given in Table 6.6. Each one is of the form $\max_{1 \le r \le r_n}(V_r/r)$, where V_r is either $2(\mathcal{L}_r - \mathcal{L}_0)$, $\mathcal{S}_r$ or $\mathcal{R}_r$. The `glim()` function in S-Plus was used to fit the models required for calculation of the LIC-based statistic for the binomial likelihood. An upper bound of $r_n = 14$ was used in each statistic since the `glim()` function in S-Plus had difficulty fitting models with more than sixteen variables.

No P-values are reported for the LIC and SIC based test for the binomial likeli-

hood, since their asymptotic distribution differs from the one given in Theorem 6.1 and 6.3. In both likelihood models, neither the cosine nor the polynomial version of the robust score statistic exceeds 4.18, the asymptotic critical value for a .05 level test, and hence the data appear to be consistent with the hypothesis that $\pi(x)$ has the form $1/[1+\exp(-(\theta_1+\theta_2 x))]$. The fact that the *LIC* and *SIC* statistics are so large in the binomial model is not surprising since this model attributes too much precision to estimates of $\pi(x)$. For the beta-binomial model the values of the test statistics are in closer agreement.

6.6.2 The twins data

For the set of twins, we will test the null hypothesis of linearity in gestational age (in weeks) for the logit of the probability of mortality and morbidity within 28 days after birth,

$$H_0 : \text{logit}(\pi(x)) = \theta_1 + \theta_2 x, \quad \text{for each } x.$$

Because of the size of this data set (there are 113 twins) we model the correlation parameter, which is the nuisance parameter in a beta-binomial model, as a constant function of gestational age. The lack of information on correlation is already suggested in Figure 3.3 (see Section 3.5.3).

	Polynomial basis			Cosine basis		
Criterion	Value	P-value	r_n	Value	P-value	r_n
SIC	0.471	0.981	15	0.799	0.948	15
RIC	1.693	0.399	15	2.993	0.120	15

Table 6.7: Test statistics for the twins data.

We used both a polynomial and a cosine basis to extend the null model. Observed values of test statistics and approximate P-values based on the asymptotic distribution theory are given in Table 6.7. There are no results for the *LIC* based tests, because of convergence problems. This demonstrates clearly the advantage of using score-based tests instead of likelihood ratio based tests. "Naive" as well as robust score-based tests do not give evidence of a more complicated structure. Based on these data, and for a level of significance smaller than 11.9%, we cannot

reject the null hypothesis of linearity for the probability of mortality and morbidity within 28 days after birth.

6.6.3 The Wisconsin diabetes study

Here we apply our proposed tests to the data set concerning the effect of the systolic blood pressure on the occurrence of macular edema at the eyes of younger onset diabetic persons. Let y_i denote the number of eyes infected (0, 1 or 2), x_i the systolic blood pressure, $\pi(x)$ the expected proportion of infected eyes for a diabetic person whose systolic blood pressure is x, and let $\rho(x)$ denote the expected correlation between the outcomes of the left and right eye. We wish to test whether the correlation changes with systolic blood pressure, in which case the null hypothesis is

$$H_0 : \rho(x) = \theta, \quad \text{for each } x. \tag{6.11}$$

We will consider two models for these data, both accounting for the intra-person correlation. The first model is the full likelihood betabinomial model (see Section 2.6.1). The second model uses generalized estimating equations (GEE2, Zhao and Prentice, 1990) based on the first four moments of the Bahadur (1961) model. Test statistics for hypothesis (6.11) are calculated using both polynomial and cosine alternatives, their values are given in Table 6.8. Note that for the GEE2 model only the *RIC* criterion yields relevant test statistics and that the *LIC* criterion is not defined in this context. The parameters corresponding to $\pi(x)$ are, for present purposes, nuisance parameters. It turns out that for these data the test results are very similar for a linear or a quadratic form of $\pi(x)$. None of the "naive" score based tests is able to reject the null hypothesis. In the generalized estimating equations model the P-values are slightly smaller than in the full likelihood model, but in both cases, the *RIC* yields P-values which are not larger than 5.5%. For these data we might consider modeling the correlation as a non constant function of systolic blood pressure.

	Polynomial basis			Cosine basis		
	BETABINOMIAL LIKELIHOOD MODEL					
Criterion	Value	P-value	r_n	Value	P-value	r_n
	$\pi(x)$ linear					
SIC	2.096	0.262	15	2.347	0.207	15
RIC	4.350	0.045	15	4.628	0.037	15
LIC	3.522	0.080	4	3.453	0.084	4
	$\pi(x)$ quadratic					
SIC	2.105	0.260	15	2.267	0.223	15
RIC	4.045	0.055	15	4.315	0.046	15
LIC	3.405	0.087	4	3.340	0.091	4
	GEE2 (BAHADUR) MODEL					
Criterion	Value	P-value	r_n	Value	P-value	r_n
	$\pi(x)$ linear					
SIC	2.293	0.218	15	2.578	0.169	15
RIC	4.632	0.037	15	4.848	0.032	15
	$\pi(x)$ quadratic					
SIC	2.308	0.214	15	2.484	0.183	15
RIC	4.381	0.044	15	4.376	0.044	15

Table 6.8: Test statistics for the Wisconsin diabetes data

Chapter 7

Some Other Data-Driven Tests and Extensions to the Multiple Covariate Case

7.1 Introduction

This chapter has two aspects. Initially we consider several tests in the simple regression setting, most of which are analogs of previously proposed tests. Our aim is to provide some insight on the power of these tests. While no test is found to be either inadmissible or uniformly most powerful, we are able to concisely characterize the power of each test, thereby making a reasoned choice of tests possible. These tests are novel in that they employ penalized score statistics, as opposed to penalized likelihood or some other risk estimate. The score-based tests are advantageous in at least three ways. They are computationally simple, they can be easily robustified to protect against model misspecification, and they can be applied in logistic as well as continuous-response regression.

A second aspect of this chapter is to consider our score-based tests in the setting of multiple regression. In doing so, we address the issue of choosing an appropriate sequence of nested alternative models. This issue is relatively trivial when testing the fit of a function of one variable, but increases quickly in complexity when the number of variables increases. We propose and compare the power of several schemes for nesting models.

The remainder of the chapter proceeds as follows. In Section 7.2 we propose several test statistics in simple regression and describe their large sample distributions. Section 7.3 summarizes a simulation study comparing the power of the tests from Section 7.2. Section 7.4 proposes and analyzes new lack-of-fit tests in multiple regression, Section 7.5 studies these tests via simulation and Section 7.6 provides some data examples and concluding remarks.

Most results in this chapter are presented in Aerts, Claeskens and Hart (1998).

7.2 Tests in simple regression

We first give some notation, which for the major part overlaps with the notation introduced in Chapter 6.

Suppose we have independent observations $(x_1, Y_1), \ldots, (x_n, Y_n)$. The covariate values are assumed to be fixed, which could correspond either to a designed experiment or to conditioning on the values of a random covariate. The density (or probability mass function) of each Y_i has the form

$$Y_i \sim f(y; \gamma(x_i), \boldsymbol{\eta}),$$

where f is known up to the function of interest $\gamma(\cdot)$ and some k-dimensional nuisance parameter $\boldsymbol{\eta}$ ($k < \infty$). We wish to test a null hypothesis of the form

$$H_0 : \gamma(\cdot) \in \{\gamma(\,\cdot\,; \boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta\}. \tag{7.1}$$

In this section we propose five nonparametric tests of H_0 that all have two things in common: they are functions of score statistics based on different model dimensions, and the model dimension is chosen by a data-based rule. The large sample distribution of each test statistic will be obtained under general conditions. Comparisons of small sample behavior and power will be conducted via simulation in Section 7.3. One of these tests is the omnibus score-based test introduced in Chapter 6.

Consider an approximator $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r})$ of $\gamma(\cdot)$ with the property that $\gamma(\,\cdot\,; \boldsymbol{\theta}, \mathbf{0}_r^T) \equiv \gamma(\,\cdot\,; \boldsymbol{\theta})$, where $r \geq 1$ and $\mathbf{0}_r$ denotes a column vector of r 0's. The log-likelihood is denoted by

$$L_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r}) = \sum_{i=1}^{n} \log f(y_i; \gamma(x_i; \theta_1, \ldots, \theta_{p+r}), \boldsymbol{\eta}), \quad r = 0, 1, \ldots \ .$$

For each $r \geq 0$, define $\boldsymbol{U}_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ to be the (column) vector of first derivatives of L_r with respect to the parameters $\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r}$ and $\boldsymbol{J}_r(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+r})$ to be the expected Fisher information. Define, for $r \geq 1$, the score statistic

$$\mathcal{S}_r = \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})^T \left(\boldsymbol{J}_r(\hat{\boldsymbol{\delta}}_{r0})\right)^{-1} \boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0}), \tag{7.2}$$

where $\hat{\boldsymbol{\delta}}_{r0} = (\hat{\boldsymbol{\delta}}_{00}, \mathbf{0}_r^T)$ and $\hat{\boldsymbol{\delta}}_{00} = (\hat{\boldsymbol{\eta}}_0, \hat{\theta}_{10}, \ldots, \hat{\theta}_{p0})$ is the maximum likelihood estimate of model parameters assuming that H_0 is true. The null hypothesis is rejected in favor of the alternative

$$H_1 : \gamma(x) = \gamma(x; \theta_1, \ldots, \theta_{p+r})$$

if the value of $\mathcal{S}_r$ is sufficiently large. When H_0 is true and appropriate regularity conditions hold (Rao, 1973), $\mathcal{S}_r$ converges in distribution to a χ^2_r random variable as $n \to \infty$. One could also use the observed information matrix in place of $\boldsymbol{J}_r(\hat{\boldsymbol{\delta}}_{r0})$, but there is no general consensus as to which is better and often the expected information leads to simpler expressions.

Our interest is in tests which are sensitive to essentially *any* departure from H_0. The score test just described can be expected to have reasonably good power when γ is of the form in H_1, but for other alternatives it need not even be consistent. Suppose we consider a sequence $\{\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r}) : r = 1, 2, \ldots\}$ of approximators of γ with the property that $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r}) \equiv \gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r}, 0)$ for each $r = 0, 1, \ldots$ and all allowable parameter values $\theta_1, \ldots, \theta_{p+r}$. In other words, the models for γ are nested and become increasingly complex as r increases. We furthermore desire that functions of the form $\gamma(\,\cdot\,; \theta_1, \ldots, \theta_{p+r})$ come closer and closer to spanning the space of all functions of interest as $r \to \infty$. For example,

$$\gamma(\cdot; \theta_1, \ldots, \theta_{p+r}) = \gamma(\cdot; \theta_1, \ldots, \theta_p) + \sum_{j=1}^{r} \theta_{p+j} u_j(\cdot)$$

where $\{u_1(\cdot), u_2(\cdot), \ldots\}$ is complete for the class of functions that are continuous on the range of the design points. In the simulations we will use orthonormalized Legendre polynomials and a cosine system, see Section 7.3. Associated with each $r \geq 1$ is a score statistic $\mathcal{S}_r$ as defined in (7.2). One expects a test based on $\mathcal{S}_{s_n}$ with $s_n \to \infty$ to be consistent against any alternative of interest. However, we now have the problem of a multiplicity of tests. Which test should be used? One could arbitrarily choose a test, but there is no assurance that the chosen test will be

particularly good. This problem is precisely analogous to that of selecting the order of a Neyman smooth test (1937). An alternative to choosing one test is to perform a sequence of tests, with the critical value of each adjusted so as to maintain an overall level of significance. Such an approach was suggested in the goodness-of-fit context by Bickel and Ritov (1992).

Our method of choosing r uses the score-based information criterion,

$$SIC(r; C_n) = \mathcal{S}_r - C_n r, \quad r = 0, 1, \ldots, \tag{7.3}$$

where $\mathcal{S}_0$ is defined to be 0 and C_n is some constant larger than 1. We may choose r to maximize $SIC(r; C_n)$ over $r = 0, 1, \ldots, r_n$, with the upper bound r_n either fixed or tending to infinity with n. In contrast, a penalized log-likelihood criterion has the form

$$2L_r(\hat{\boldsymbol{\eta}}, \hat{\theta}_1, \ldots, \hat{\theta}_{p+r}) - C_n r, \quad r = 0, 1, \ldots, \tag{7.4}$$

where $(\hat{\boldsymbol{\eta}}, \hat{\theta}_1, \ldots, \hat{\theta}_{p+r})$ is a maximum likelihood estimate, i.e. the maximizer of L_r. Taking C_n equal to 2 and $\log n$ in (7.4) yields the well-known AIC (Akaike, 1974) and BIC (Schwarz, 1978), respectively. When the null hypothesis is true, the difference between $SIC(r; C_n)$ and (7.4) is negligible. For this reason we shall refer to $SIC(r; 2)$ and $SIC(r; \log n)$ as score analogs of AIC and BIC. When the alternative hypothesis is true, the maximizer of AIC is a reasonable estimator of the dimension of the model which minimizes Kullback-Leibler risk. The same cannot be said of $SIC(r; 2)$, but this is not necessarily relevant in a testing context where power is the relevant performance measure. The score criterion has the advantage of requiring MLEs of model parameters only under the null hypothesis.

Define the statistics $\hat{r}_a$ and $\hat{r}_b$ by

$$\hat{r}_a = \text{argmax}_{0 \le r \le r_n} SIC(r; 2)$$

and

$$\hat{r}_b = \text{argmax}_{1 \le r \le r_n} SIC(r; \log n).$$

We consider the following test statistics:

$$\begin{aligned} S_a &= \mathcal{S}_{\hat{r}_a}; \qquad S_b = \mathcal{S}_{\hat{r}_b} \\ T_a &= \frac{\mathcal{S}_{\hat{r}_a} - \hat{r}_a}{\max(1, \sqrt{\hat{r}_a})}; \qquad (7.5) \\ T_{os} &= \max_{1 \le r \le R_n} \frac{\mathcal{S}_r}{r}; \qquad (7.6) \\ T_{max} &= SIC(\hat{r}_a; 2). \qquad (7.7) \end{aligned}$$

Our first two test statistics are simply $\mathcal{S}_{\hat{r}_a}$ and $\mathcal{S}_{\hat{r}_b}$, i.e., the score statistic with number of alternative parameters chosen by the score analogs of AIC and BIC, respectively. Note that $\hat{r}_b$ maximizes $SIC(r; \log n)$ over $1, \ldots, r_n$ rather than $0, 1, \ldots, r_n$. This definition is used due to a consistency property of BIC-type order selection criteria. When maximization is conducted over the set $0, 1, \ldots, r_n$, the statistic $\hat{r}_b$ is a consistent estimator of 0 under the null hypothesis. This means that the limit distribution of $\mathcal{S}_{\hat{r}_b}$ is degenerate at 0, making the asymptotic level of a non-randomized test either 0 or 1. Maximizing SIC over the set $\{1, \ldots, r_n\}$ (as suggested by Ledwina, 1994) allows one to perform a non-randomized (large sample) test of any level.

Standardizing $\mathcal{S}_{\hat{r}_a}$ leads to the statistic (7.5) which is analogous to a statistic proposed by Eubank, Li and Wang (1997) and may be motivated as follows. Suppose we use a score statistic $\mathcal{S}_{s_n}$ with a nonstochastic dimension s_n. As mentioned above, if we desire a test that is consistent against virtually any alternative, we must let $s_n \to \infty$. The large sample distribution of $\mathcal{S}_{s_n}$ as $n \to \infty$ is thus relevant. Under general conditions, one can show that

$$\frac{\mathcal{S}_{s_n} - s_n}{\sqrt{2s_n}} \tag{7.8}$$

converges in distribution to a standard normal random variable as $n \to \infty$ and $s_n \to \infty$ at a sufficiently slow rate. It thus seems natural to use a test statistic of the form (7.8), and statistic (7.5) is essentially (7.8) with a data-driven choice for s_n. It turns out that standardizing S_a as above greatly stabilizes the null distribution of the statistic, which has a decided effect on the power for (7.5). One could similarly standardize S_b, but the null distribution of this statistic is already quite stable and the standardization has a negligible effect.

The null hypothesis is false if and only if the best model dimension, r, is at least 1. It would thus make sense to reject H_0 when $\hat{r}_a > 0$. A problem with this test is that it has a limiting size of 0.29 (see Section 6.2.2), but this can be overcome by taking $C_n > 2$ in $SIC(r; C_n)$. For example, if we reject H_0 if and only if the maximizer of $SIC(r; 4.18)$ exceeds 0, the result is a test with limiting size of 0.05. So, for the appropriate choice of $C_n = C_\alpha$, we define the order selection test by

$$\text{“reject } H_0 \text{ when } \hat{r}_{C_\alpha} > 0,\text{”}$$

where $\hat{r}_{C_\alpha}$ maximizes $SIC(r; C_\alpha)$. This test is a special case of one proposed in the previous chapter, and is analogous to that of Eubank and Hart (1992). It is

noteworthy that the order selection test is equivalent to a test that rejects H_0 for large values of T_{os}, as defined in (7.6).

Finally, we propose the test statistic (7.7) i.e., the AIC-type score criterion evaluated at its maximum. This type of statistic has an interesting history. Recently, Simonoff and Tsai (1999) propose to use the maximum of AIC as a lack-of-fit statistic. However, the test can actually be traced to a proposal of Parzen (1977) for testing the white noise hypothesis in time series. As he presented the test, it bears a stronger resemblance to the order selection test than to (7.7). However, Hart (1997, pp. 246–248) showed that Parzen's test is equivalent to rejecting the white noise hypothesis when the maximum of an estimated risk criterion is sufficiently large. The difference between the order selection and T_{max}-based tests is more than cosmetic. It turns out that the two tests have quite different power characteristics, as we will see in Section 7.3.

We now consider the large sample null distribution of each of the five statistics proposed above. To facilitate the statement of a theorem, we introduce the following notation. Let $Z_1, Z_2, \ldots$ be a sequence of independent and identically distributed standard normal random variables, and define

$$V_0 = 0, \qquad V_r = \sum_{j=1}^{r} Z_j^2, \quad r = 1, 2, \ldots,$$

and $\tilde{r}$ to be the value of r that maximizes $V_r - 2r$ over $r = 0, 1, \ldots$.

Theorem 7.1 *Let the probability model of Section 7.2 hold and assume that* $\boldsymbol{J}_1(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+1})$ *is positive definite for each* $(\boldsymbol{\eta}, \theta_1, \ldots, \theta_{p+1})$ *in the allowable set of parameter values. In addition we assume the following conditions.*

(B1) The matrix $\boldsymbol{J}_r(\hat{\boldsymbol{\delta}}_{r0})/n$ *converges in probability to a positive definite matrix* $\mathcal{J}_r$*, where* $\mathcal{J}_r$ *is the upper left* $(p+k+r)\times(p+k+r)$ *submatrix of* $\mathcal{J}_{r+1}$*,* $r = 1, 2, \ldots$.

(B2) The maximum likelihood estimate $\hat{\boldsymbol{\delta}}_{00}$ *is a solution of*

$$\boldsymbol{U}_0(\boldsymbol{\eta}, \theta_1, \ldots, \theta_p) = \mathbf{0}_{p+k}.$$

(B3) For each $r \geq 1$*,* $\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})/\sqrt{n}$ *converges in distribution to* $(\mathbf{0}_{p+k}^T, \mathcal{U}_r^T)^T$*, where* $\mathcal{U}_r$ *has an* r*-variate normal distribution with mean vector* $\mathbf{0}_r$ *and covariance matrix* $\tilde{\mathcal{J}}_r$*, with* $\tilde{\mathcal{J}}_r^{-1}$ *the lower right* $r \times r$ *submatrix of* $\mathcal{J}_r^{-1}$.

(B4) For any $\epsilon > 0$*, there exists a nonnegative function* g_ϵ *satisfying* $\lim_{y\to\infty} g_\epsilon(y) = 0$*, a summable sequence of nonnegative numbers* $\{p_{r,\epsilon}\}_{r\geq 1}$ *and a nonnegative sequence* $\{\delta_{n,\epsilon}\}$ *such that* $\lim_{n\to\infty} r_n\delta_{n,\epsilon} = 0$ *and for all* $a \geq \epsilon + 1$ *and* $r = 1, \ldots, r_n,$,

$$P(\mathcal{S}_r > ar) \leq g_\epsilon(a)p_{r,\epsilon} + \delta_{n,\epsilon}.$$

It follows that if the null hypothesis (7.1) is true and $n \to \infty$*, the statistics* S_a*,* T_a*,* T_{os} *and* T_{max} *converge in distribution to* $V_{\tilde{r}}$*,* $(V_{\tilde{r}} - \tilde{r})/\max(1, \sqrt{\tilde{r}})$*,* $\max_{r\geq 1}(V_r/r)$ *and* $V_{\tilde{r}} - 2\tilde{r}$*, respectively. Furthermore, as* $n \to \infty$*,* S_b *converges in distribution to a random variable having the* χ^2_1 *distribution.*

Proof. For the first four statistics, we will only prove the result for T_{max}; the other three proofs are very similar. Define $\gamma = \max_{r\geq 1}(V_r - 2r)$, $\gamma_R = \max_{1\leq r\leq R}(V_r - 2r)$, $\gamma_{R,n} = \max_{1\leq r\leq R}(\mathcal{S}_r - 2r)$ and $\beta_{R,n} = \max_{R<r\leq r_n}(\mathcal{S}_r - 2r)$. We have

$$P(T_{max} > x) = P(\gamma_{R,n} > x) + P(\beta_{R,n} > x) - P(\gamma_{R,n} > x \cap \beta_{R,n} > x),$$

and hence

$$\begin{aligned} &|P(T_{max} > x) - P(\gamma > x)| \leq \\ &\quad |P(\gamma_{R,n} > x) - P(\gamma_R > x)| + |P(\gamma_R > x) - P(\gamma > x)| + P(\beta_{R,n} > x). \end{aligned}$$

We must show that, given any $\epsilon > 0$, $|P(T_{max} > x) - P(\gamma > x)| < \epsilon$ for all n sufficiently large. By definition, $P(\gamma_R > x)$ increases to $P(\gamma > x)$ as $R \to \infty$, and so there exists $\bar{R}$ such that $|P(\gamma_R > x) - P(\gamma > x)| < \epsilon/3$ for all $R \geq \bar{R}$.

Now,

$$\begin{aligned} P(\beta_{R,n} > x) &= P\left(\bigcup_{r=R+1}^{r_n} \{\mathcal{S}_r - 2r > x\}\right) \\ &\leq \sum_{r=R+1}^{r_n} P(\mathcal{S}_r > 2r + x). \end{aligned}$$

If $R \geq 2|x|$, the very last quantity is bounded by $\sum_{r=R+1}^{r_n} P(\mathcal{S}_r > 1.5r)$, and by assumption (B4)

$$\sum_{r=R+1}^{r_n} P(\mathcal{S}_r > 1.5r) \leq g_\epsilon(1.5) \sum_{r=R+1}^{r_n} p_{r,\epsilon} + r_n\delta_{n,\epsilon}.$$

Again by (B4), there exists $\tilde{R}$ such that the last quantity is less than $\epsilon/3$ for all $R \geq \tilde{R}$ and all n sufficiently large. Defining $R_{max} = \max(\bar{R}, \tilde{R}, 2|x|)$ and combining previous steps we have

$$|P(\gamma_{R_{max}} > x) - P(\gamma > x)| + P(\beta_{R_{max},n} > x) < \frac{2\epsilon}{3}$$

for all n sufficiently large.

The only remaining task is to show that

$$|P(\gamma_{R_{max},n} > x) - P(\gamma_{R_{max}} > x)| < \epsilon/3$$

for all n sufficiently large. For any R, $\gamma_{R,n}$ is a continuous function of $\boldsymbol{U}_R(\hat{\boldsymbol{\delta}}_{R0})/\sqrt{n}$ and the elements of $\boldsymbol{J}_R(\hat{\boldsymbol{\delta}}_{R0})/n$. This is a consequence of the following facts:

- For $1 \leq r \leq R$, $\boldsymbol{J}_r(\hat{\boldsymbol{\delta}}_{r0})/n$ is a submatrix of $\boldsymbol{J}_R(\hat{\boldsymbol{\delta}}_{R0})/n$.
- $\boldsymbol{U}_r(\hat{\boldsymbol{\delta}}_{r0})/\sqrt{n}$ is a subvector of $\boldsymbol{U}_R(\hat{\boldsymbol{\delta}}_{R0})/\sqrt{n}$, $r = 1, \ldots, R$.
- The quantity $\max_{1 \leq r \leq R}(s_r - 2r)$ is a continuous function of $s_1, \ldots, s_R$.

By conditions (B1)–(B3), $(\boldsymbol{U}_R(\hat{\boldsymbol{\delta}}_{R0})/\sqrt{n}, \boldsymbol{J}_R(\hat{\boldsymbol{\delta}}_{R0})/n)$ converges in distribution to $((\mathbf{0}_{p+k}^T, \mathcal{U}_R^T)^T, \mathcal{J}_R)$ as $n \to \infty$. Since $\gamma_{R,n}$ is a continuous function of $(\boldsymbol{U}_R(\hat{\boldsymbol{\delta}}_{R0})/\sqrt{n}$, $\boldsymbol{J}_R(\hat{\boldsymbol{\delta}}_{R0})/n)$, it follows that $\gamma_{R,n}$ converges in distribution to the same function of $((\mathbf{0}_{p+k}^T, \mathcal{U}_R^T)^T, \mathcal{A}_R)$, which is equal in distribution to γ_R.

We now consider the distribution of S_b. Using assumptions (B1) and (B2) and the fact that

$$P(\mathcal{S}_{\hat{r}_b} > x) = P(\mathcal{S}_1 > x \cap \hat{r}_b = 1) + O\left\{P(1 < \hat{r}_b \leq r_n)\right\},$$

the result is proven by showing that $P(\hat{r}_b = 1) \to 1$ as $n \to \infty$. We have

$$\begin{aligned} P(\hat{r}_b > 1) &\leq \sum_{r=2}^{r_n} P\left(\mathcal{S}_r - \mathcal{S}_1 > \log n(r-1)\right) \\ &\leq \sum_{r=2}^{r_n} P\left(\mathcal{S}_r > \log n(r-1)\right), \end{aligned}$$

with the last inequality following from the fact that, by assumption, $\mathcal{S}_1$ is almost surely nonnegative. Applying (B4), the last quantity is bounded by

$$\max_{2 \leq r \leq r_n} g_\epsilon\left\{(\log n)/2\right\} \sum_{r=2}^{r_n} p_{r\epsilon} + r_n \delta_{n,\epsilon},$$

which tends to 0 by assumption (B4).

Some remarks are in order concerning the conditions in Theorem 7.1. Conditions (B1)–(B3) are standard in the asymptotic theory of maximum likelihood estimation. Condition (B4) is a sort of large deviation inequality for the score statistic $\mathcal{S}_r$. To motivate (B4), consider testing the null hypothesis

$$H_0^* : \gamma \equiv \text{constant}$$

in a model where $x_i = (i - 1/2)/n$, $i = 1, \ldots, n$. When testing H_0^*, if series approximators of the form

$$\gamma(x; \theta_1, \ldots, \theta_{r+1}) = \theta_1 + 2 \sum_{j=1}^{r} \theta_{j+1} \cos(\pi j x)$$

are used and $Y_1, \ldots, Y_n$ are modeled as independent $\mathcal{N}(\gamma(x_i), \sigma^2)$ random variables, then $\mathcal{S}_r$ takes the form

$$\mathcal{S}_r = \sum_{j=1}^{r} \frac{2n\hat{\theta}_{j+1}^2}{\hat{\sigma}^2}$$

in which

$$\hat{\theta}_{j+1} = \frac{1}{n} \sum_{i=1}^{n} Y_i \cos(\pi j x_i), \quad j = 0, 1, 2, \ldots ,$$

and $\hat{\sigma}^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2/n$. Let σ_0^2 denote the variance of Y_i under the null hypothesis. Then

$$P(\mathcal{S}_r > a_n r) \leq P\left(\sum_{j=1}^{r} \frac{2n\hat{\theta}_{j+1}^2}{\sigma_0^2} > (1-b) a_n r \right) + P(|\hat{\sigma}^2/\sigma_0^2 - 1| \geq b),$$

where b is a constant between 0 and 1 and such that $a_n(1-b) > 1$ for all n sufficiently large. When $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}(0, \sigma_0^2)$, it follows that $2n\hat{\theta}_2^2/\sigma_0^2, \ldots, 2n\hat{\theta}_n^2/\sigma_0^2$ are i.i.d. χ_1^2 random variables, implying that

$$P\left(\sum_{j=1}^{r} \frac{2n\hat{\theta}_{j+1}^2}{\sigma_0^2} > (1-b) a_n r \right) \leq (1-2t)^{-r/2} \exp(-t(1-b) a_n r)$$

for any number $t \in (0, 1/2)$. Taking $t = (1/2)\{1 - 1/[(1-b)a_n]\}$ the quantity on the right of the last inequality is

$$\begin{aligned} \exp\left\{-(r/2)\left[a_n(1-b) - 1 - \log\left(a_n(1-b)\right)\right]\right\} &= \exp(-c_n r) \\ &= \exp(-\beta r) \exp(-(c_n - \beta) r), \end{aligned}$$

where β is a constant such that $\beta \in (0, c_n)$ for all n sufficiently big (which is possible since $\{c_n\}$ is a nondecreasing sequence for $a_n(1-b) > 1$). In the notation of condition (B4), we may now take $p_r = \exp(-\beta r)$ and $g(a_n) = \exp(-(c_n - \beta))$. Finally, a large deviation formula for $\hat{\sigma}^2$ yields

$$P(|\hat{\sigma}^2/\sigma_0^2 - 1| \geq b) \leq \delta_n = O\left(\frac{1}{n^2}\right).$$

Notice that in the preceding example we may take $r_n = n$ and still obtain $r_n \delta_n \to 0$. Hence, the distribution theory in Theorem 7.1 holds in this case without placing an arbitrary upper bound on the number of Fourier coefficients used. One can similarly check condition (B4) in other settings by using large deviation formulae for $\hat{\boldsymbol{\delta}}_{00}$ and $\boldsymbol{U}_r(\boldsymbol{\delta}_{r0})^T \left(\boldsymbol{J}_r(\boldsymbol{\delta}_{r0})\right)^{-1} \boldsymbol{U}_r(\boldsymbol{\delta}_{r0})$. Note, however, that in general it may be necessary to take $r_n = o(n)$ in order to obtain the limit distributions of Theorem 7.1.

7.3 Simulations in simple regression models

To illustrate the different tests and learn about their power characteristics, we performed a simulation study using two members of the generalized linear models family. In general, the likelihood function for a generalized linear model (GLM) can be written as

$$f(y_i; \gamma(x_i), \boldsymbol{\eta}) = \exp\{(y_i \gamma(x_i) - b(\gamma(x_i)))/a(\boldsymbol{\eta}) + c(y_i, \boldsymbol{\eta})\}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot)$ are known functions, $\gamma(\cdot)$ is the unknown "natural" parameter of interest and $\boldsymbol{\eta}$ is an unknown dispersion parameter (see, e.g., McCullagh and Nelder, 1989). As approximators to $\gamma(\cdot)$ we take

$$\gamma(x; \theta_1, \ldots, \theta_{p+r}) = \sum_{j=1}^{p} \theta_j \gamma_j(x) + \sum_{j=1}^{r} \theta_{p+j} u_j(x), \quad r = 1, 2, \ldots .$$

It is assumed that the set of functions $\{u_1, u_2, \ldots\}$ is complete for the class of functions that are continuous on the range of the design points. It is understood that the functions $u_j(\cdot)$ are not linear combinations of any of the functions $\gamma_j(\cdot)$ $(j = 1, \ldots, p)$.

Let v and w be any two functions in $\mathcal{B}_r = \{\gamma_1, \ldots, \gamma_p, u_1, \ldots, u_r\}$, and suppose $\mathcal{B}_r$ is orthonormal in the sense that

$$\frac{1}{n} \sum_{i=1}^{n} b''\{\gamma(x_i)\} v(x_i) w(x_i) = \begin{cases} 1, & v \equiv w \\ 0, & v \not\equiv w. \end{cases}$$

Then the score statistic is given by

$$\mathcal{S}_r = \sum_{j=1}^{r} \frac{n\hat{\alpha}_j^2}{a(\hat{\boldsymbol{\eta}}_0)}, \tag{7.9}$$

where

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - b'(\gamma(x_i; \hat{\theta}_{10}, \ldots, \hat{\theta}_{p0}))\}\, u_j(x_i).$$

Expression (7.9) has essentially the same form as the statistic of Neyman's classic smooth test; see Hart (1997).

To compare the five tests based on S_b, S_a, T_a, T_{os} and $T_{\max}$, we generated continuous

$$Y_i \sim \mathcal{N}(\gamma(x_i), \boldsymbol{\eta})$$

and binary

$$Y_i \sim \text{Bern}\{1/(1+\exp\{-\gamma(x_i)\})\}$$

where, in both cases, $x_i = (i-1/2)/n$, $i = 1, \ldots, n$. We focus on testing for no effect, i.e., $\gamma(x)$ is constant, using the normalized Legendre polynomials $u_j(\cdot) \equiv \mathcal{L}_j(\cdot)$ on the interval $[1/(2n), 1-1/(2n)]$, or the cosine basis $u_j(\cdot) \equiv \sqrt{2}\cos(\pi j \cdot)$, $j = 1, \ldots, 20$. Unless otherwise stated, S-Plus is used for the calculations.

From a simulation based on 30000 replications, we obtained critical points (levels $\alpha = 0.01, 0.05, 0.10$) of the large sample distribution of each test statistic, except for S_b, which is asymptotically χ^2_1. The simulated critical values are shown in Table 7.1.

α:	0.10	0.05	0.01
S_b	2.7055	3.8415	6.6349
S_a	8.6055	13.8286	27.2336
T_a	3.4438	4.4387	6.6794
T_{os}	3.2208	4.1793	6.7442
$T_{\max}$	2.4557	4.0987	8.2920

Table 7.1: Critical points for single covariate models.

Table 7.2 shows simulated type I error probabilities for a simulation of size 5000, with $\gamma(x) \equiv 0$, sample size 100 and the Legendre polynomial basis. The results for

	Gaussian			Binomial		
Statistic	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
S_b	0.1226	0.0768	0.0266	0.1394	0.0886	0.0340
S_a	0.0880	0.0376	0.0038	0.1002	0.0376	0.0034
T_a	0.0806	0.0390	0.0076	0.0950	0.0444	0.0092
T_{os}	0.0878	0.0430	0.0090	0.1048	0.0506	0.0092
$T_{\max}$	0.0802	0.0410	0.0062	0.0940	0.0466	0.0070

Table 7.2: Simulated type I error probabilities for a Legendre polynomial basis.

Gaussian and binomial data are quite similar to each other. In each case, the true level of S_b is too high and, at the 1% level, the level of S_a is too low. The type I error probabilities are closer to their nominal values for binomial data, presumably as such data require no estimate of a nuisance parameter. The same conclusions hold when a cosine basis is used.

To examine power we consider two kinds of alternatives:

$$\gamma(x) = \cos(\pi m_0 x), \tag{7.10}$$

and

$$\gamma(x) = \mathcal{L}_{m_0}(x). \tag{7.11}$$

For the normal response Y we took $\eta = 1$ ($\eta = 0.1$) for the cosine (resp. polynomial) alternative. For $m_0 = 1, 2, \ldots, 10$, these alternative models are ordered from low to high frequency. In Figure 7.1 results are shown for 1000 data sets generated from a Gaussian model with, in panels (a) and (b), $\gamma(x)$ as in (7.10). In panels (c) and (d) $\gamma(x)$ is as in (7.11). Figure 7.2 shows the simulation results for a logistic regression model under the same set of alternatives. In both cases, the sample size n equals 100 and the level of significance is equal to 0.05. For all tests, critical points were calculated using 5000 simulated data sets under the null hypothesis. In this way, the power results for the various tests are directly comparable.

The order selection test T_{os} is very good at the lowest frequency, but going to higher frequencies, its power decreases rapidly. The BIC-based test S_b shows about the same behavior. It has good power at low frequency alternatives, but since it has a large penalty for m_0 large, it has almost no power at high frequencies. At low

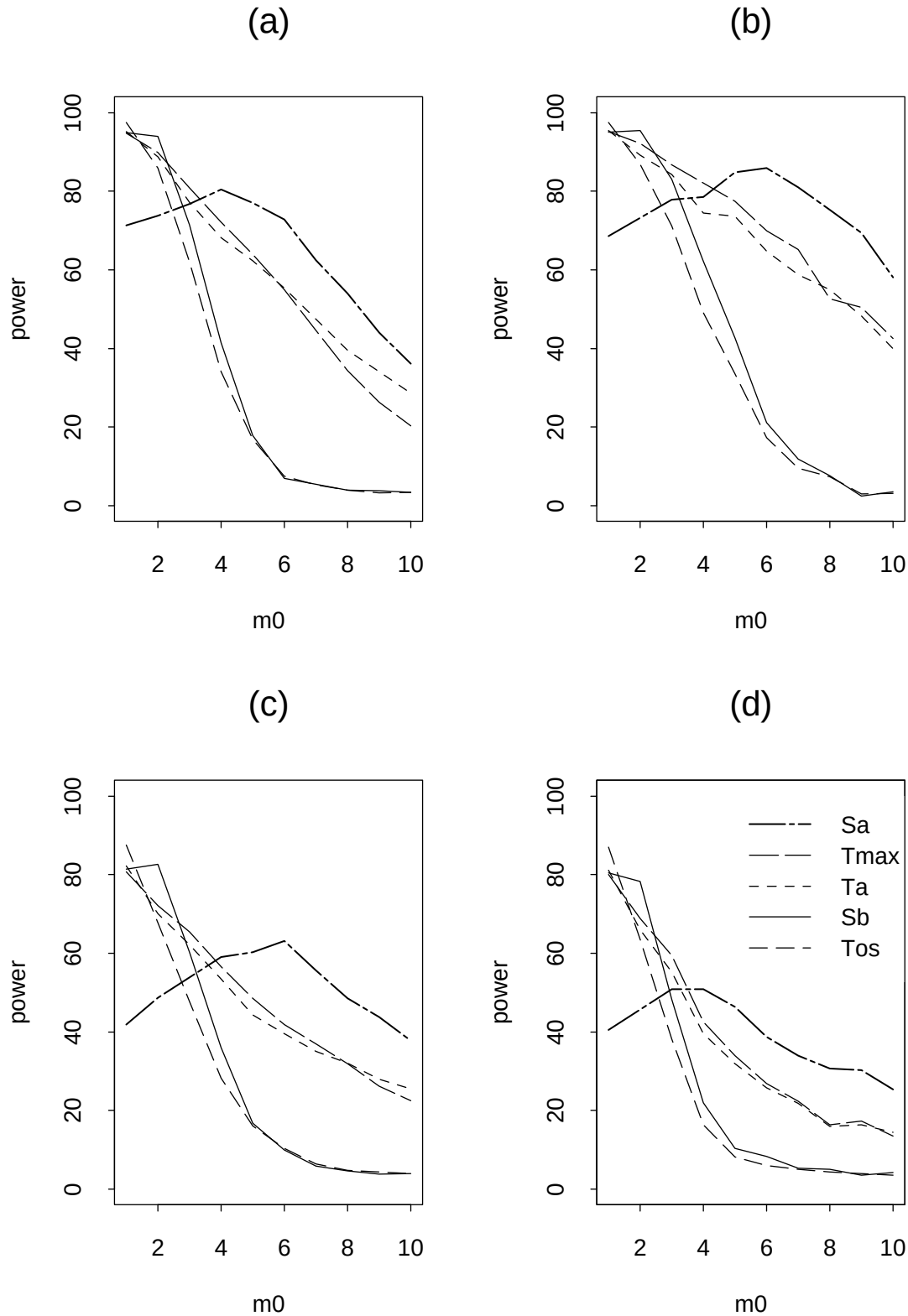


Figure 7.1: Simulated power curves for Gaussian model.

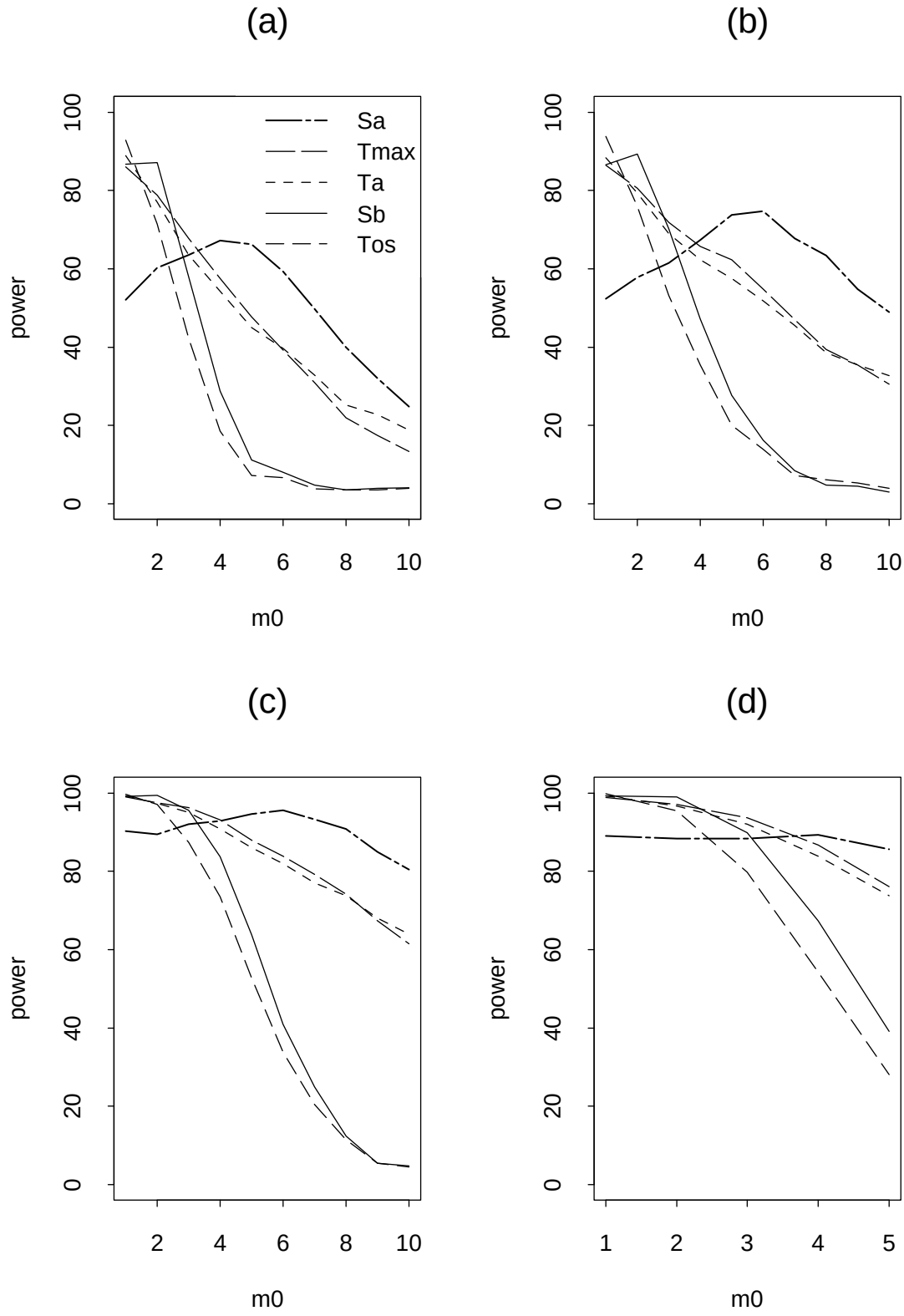


Figure 7.2: Simulated power curves for logistic regression model.

frequency alternatives, both T_a and $T_{\max}$ improve upon the unstandardized S_a test, but when m_0 is at least 4, S_a has the largest power. Since $T_{\max}$ and T_a are both doing very well for the low frequency cases and still have reasonable power at high frequencies, we recommend one of these two tests. If one suspects a higher frequency alternative, S_a is the best choice. For combinations of low and high frequency terms, the low frequencies will be dominant.

The same conclusions hold for each of the 8 situations (Gaussian or logistic, cosine or polynomial basis and both types of alternatives). For the alternatives in this simulation study, there is almost no loss in power when the "wrong" basis is used.

7.4 Multiple regression

Here we propose several ways in which the tests of Section 7.2 may be generalized to the case of more than one covariate. The first method, for models with two covariates, has the ability to detect virtually any departure from the null hypothesis, at least for a sufficiently large sample size. Tests in Sections 7.4.2 and 7.4.4 are more specialized, being designed to detect additive alternatives to H_0, interaction terms and single-index-type alternatives, respectively. The latter three tests have the advantage of being more powerful than the first against the specific type of alternative for which they are designed. Section 7.4.3 defines an extension to models with more than two covariates.

7.4.1 Omnibus tests in models with two covariates

To illustrate how the methods of Section 7.2 may be generalized to multiple regression, we consider the relatively simple case of two covariates. Let γ be an unknown function of the covariates x_1 and x_2. We wish to test the null hypothesis

$$H_0 : \gamma \in \{\gamma(\cdot, \cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}. \tag{7.12}$$

If we use a cosine series to represent γ, an alternative model may be expressed as

$$\gamma(x_1, x_2) = \gamma(x_1, x_2; \boldsymbol{\theta}) + \sum\sum_{(j,k)\in\Lambda} \alpha_{jk} \cos(\pi j x_1) \cos(\pi k x_2). \tag{7.13}$$

The definition of the index set Λ will, in general, depend on the specific model under the null hypothesis. For example, if we wish to test the hypothesis that $\gamma(x_1, x_2)$ has

the form $\theta_1 + \theta_2 \cos(\pi x_1) + \theta_2 \cos(\pi x_2)$, and we use a cosine basis, then clearly $(1,0)$ and $(0,1)$ should not be included in Λ. For ease of notation, we will now assume that the function $\gamma(x_1, x_2, \boldsymbol{\theta})$ is constant, but generalizations are straightforward. Under the no-effect null hypothesis, Λ is a subset of

$$\{(j,k) : 0 \leq j,k < n,\ j+k > 0\}.$$

In analogy to (7.2) and (7.3), we define a score statistic $\mathcal{S}_\Lambda$ and the criterion

$$SIC(\Lambda; C_n) = \mathcal{S}_\Lambda - C_n N(\Lambda),$$

respectively, where $N(\Lambda)$ denotes the number of elements in Λ.

To carry out a test as in Section 7.2 we maximize $SIC(\Lambda; C_n)$ over some collection of subsets $\Lambda_1, \Lambda_2, \ldots, \Lambda_{m_n}$. It is important that this collection corresponds to nested models, otherwise the distributions of the resultant test statistics will, in general, depend on parameters of the null model, even when $n \to \infty$. We thus insist that $\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_{m_n}$, and we call such a collection of sets a *model sequence*. The only problem now is deciding on how to choose a model sequence, since obviously there are many possibilities. One important consideration is whether a given sequence will lead to a consistent test. To ensure consistency against virtually any alternative to H_0, we ask that $N(\Lambda_{m_n}) \to \infty$ in such a way that, for each $(j,k) \neq (0,0)$ $(j,k \geq 0)$, (j,k) is in Λ_{m_n} for all n sufficiently large. The choice of a model sequence is further simplified if we consider only tests that place equal emphasis on the two covariates. In other words, we could insist that terms of the form $\cos(\pi j x_1)\cos(\pi k x_2)$ and $\cos(\pi k x_1)\cos(\pi j x_2)$ enter the model simultaneously.

We shall consider four different choices for the model sequence. Apparently, these sequences, and slight variations thereof, e.g. first main effects and then interactions in (d), are the only possibilities if one wishes to treat the covariates symmetrically. In Figure 7.3 these four sequences are graphically represented by the number of the step in which the basis elements enter the model. The numbers on the axes are the indices of the basis elements. We now give a concrete example of how to construct a model sequence according to scheme (b). As an illustration, we use cosine basis functions; it should be clear how to follow this scheme for other bases. Below we give explicitly the basis functions which are added to the previous model in the first three steps. For ease of notation, we assume that we want to test the null hypothesis of no effect.

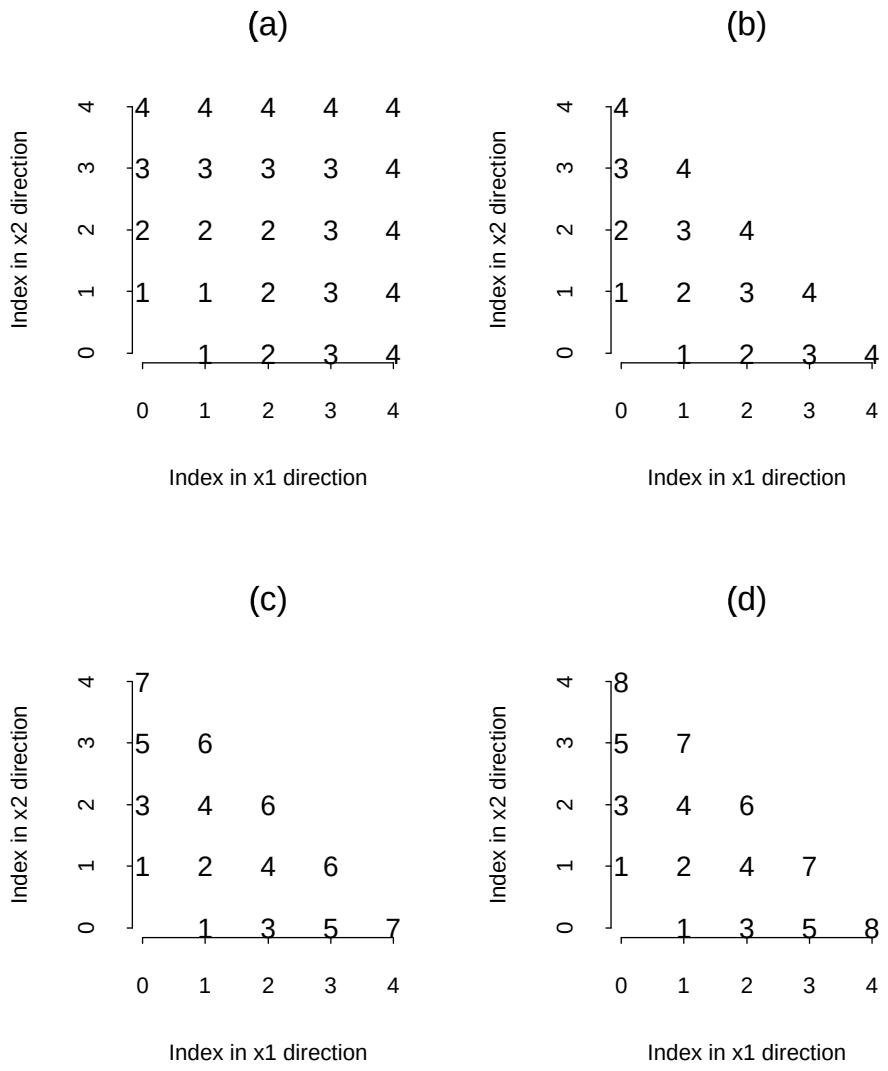


Figure 7.3: Model sequences in two dimensions

1. $\cos(\pi x_1)$ and $\cos(\pi x_2)$
2. $\cos(\pi 2x_1)$, $\cos(\pi x_1)\cos(\pi x_2)$ and $\cos(\pi 2x_2)$
3. $\cos(\pi 3x_1)$, $\cos(\pi x_1)\cos(\pi 2x_2)$, $\cos(\pi 2x_1)\cos(\pi x_2)$ and $\cos(\pi 3x_2)$
4. ...

For model sequence (a) the number of model parameters grows quickly, with $2j+1$ terms added to the previous model at step j. This entails that tests based on sequence (a) will, in general, have poor power properties. This problem is lessened in model sequence (b), where only $j+1$ terms are added at the jth step. Even more parsimonious sequences are shown in (c) and (d). In scheme (c), main effects corresponding to frequency j enter the model at step $2j-1$ and j interaction terms enter at step $2j$. Sequence (d) has the overall smallest step sizes, with at most two new terms added at each step. Schematically, (d) starts at the bisector and moves in a symmetric way towards the axes. Intuitively, we have a slight preference for the third and fourth sequences over the other two since at odd numbered steps they add only two new terms to the model, whereas the first two add an ever-increasing number at *each* step. Other omnibus tests are certainly possible; Kallenberg and Ledwina (1999) and Bogdan (1999), for example, propose model sequences for tests of independence and goodness-of-fit, respectively. For the sake of simplicity, though, we will restrict our attention to sequences (a)-(d).

The large sample distribution of the statistics S_b, S_a, T_a, T_{os} and $T_{\max}$ as given in Theorem 1 can be generalized to the multiple covariate setting by changing the definition of V_r and $\tilde{r}$ as follows. First define $\Lambda_0 = \emptyset$, $N_j = N(\Lambda_j) - N(\Lambda_{j-1})$, $j = 1,2,\ldots$, and let Z_{jk}, $k = 1,\ldots,N_j$, $j = 1,2,\ldots$, be i.i.d. standard normal random variables. Now we define

$$V_0 = 0, \qquad V_r = \sum_{j=1}^{r}\sum_{k=1}^{N_j} Z_{jk}^2, \quad r = 1,2,\ldots,$$

and

$$\tilde{r} = \arg\max_{r=0,1,\ldots} \{V_r - 2N(\Lambda_r)\}.$$

In the case of three or more covariates, it is worthwhile to point out the price to be paid by an omnibus test as the number, d, of covariates increases. Regardless of what d is, the maximum number of parameters we should consider in a model is

$O(n)$, and for simplicity let us just say n. For an omnibus test that places the same emphasis on all d covariates, this entails, roughly speaking, that r_n not exceed $n^{1/d}$. Clearly then, the ability of an omnibus test to detect higher frequency alternatives quickly wanes as the dimension of the x-space increases. This is but another example of the curse of dimensionality rearing its ugly head.

7.4.2 Tests in additive models

For a high dimensional covariate vector, the omnibus tests based on the four model sequences of Section 7.4.1 become less attractive. However, if we can assume that an *additive* model fits the data well, the curse of dimensionality can be circumvented. Under this assumption, an alternative to the null model (7.12) can, for two covariates and the cosine basis, be written as

$$\gamma(x_1, x_2) = \gamma(x_1, x_2; \boldsymbol{\theta}) + \sum_{j \in \Lambda_a} a_j \cos(\pi j x_1) + \sum_{j \in \Lambda_b} b_j \cos(\pi j x_2)$$

which has the form of (7.13) with
$\Lambda \subseteq \Lambda_n^* = \{(j,0) : 1 \le j \le m_n\} \cup \{(0,k) : 1 \le k \le m_n\}$.

As before we define a score statistic S_Λ and the criterion $SIC(\Lambda; C_n) = S_\Lambda - C_n(N(\Lambda_a) + N(\Lambda_b))$. The test is carried out by maximizing $SIC(\Lambda; C_n)$ over a collection $\Lambda_1 \subseteq \Lambda_2 \subseteq \ldots \subseteq \Lambda_{r_n}$ of subsets of Λ_n^*. Again, there are a number of ways to construct such a sequence of nested models. One possibility is to use, at step r, a series estimate of the form

$$\sum_{j=1}^{k_r} a_j \cos(\pi j x_1) + \sum_{j=1}^{\ell_r} b_j \cos(\pi j x_2),$$

where $k_r \ge k_{r-1}$ and $\ell_r \ge \ell_{r-1}$ for $r = 2, 3, \ldots$. If we insist that $k_r = \ell_r$ and let k_r increase by 1 at each step, then

$$\Lambda_j = \tilde{\Lambda}_j = \{(1,0), (0,1), (2,0), (0,2) \ldots, (j,0), (0,j)\}.$$

We will refer to a test based on this model sequence as a *diagonal test*, since the "path" $\{(k_r, \ell_r) : r \ge 1\}$ corresponding to this test proceeds along the diagonal $\{(k_r, k_r) : r \ge 1\}$ (see Figure 7.4). At each step in a diagonal test two terms are added to the previous model. The only effect this has on the asymptotic distribution theory of Section 7.2 is that now $V_r = \sum_{j=1}^{r} (Z_{2j-1}^2 + Z_{2j}^2)$. This approach has an obvious extension to the case of more than two covariates.

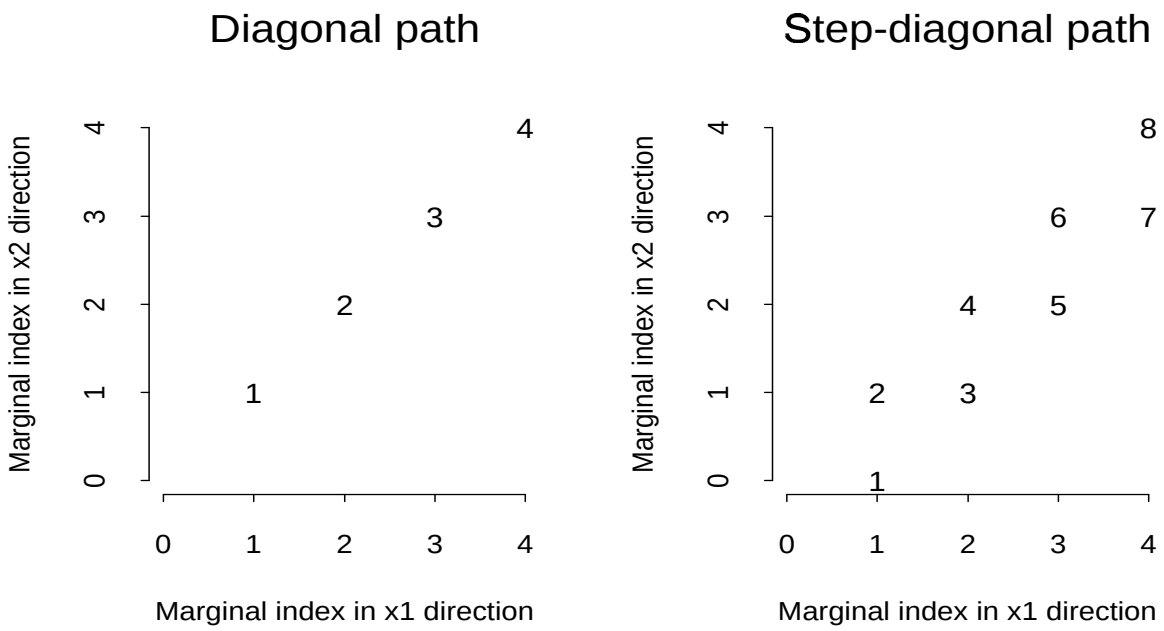


Figure 7.4: The diagonal and step-diagonal path in two dimensions

Another model sequence, for which the asymptotic distribution theory of Section 7.2 remains valid, is obtained by adding alternately a term in x_1, followed by the corresponding term in x_2, or vice versa. This may be referred to as a *step-diagonal* path, which has the disadvantage that the symmetry is lost. A comparison with the diagonal path with be made in the simulation study in Section 7.5.

7.4.3 The "max" tests in models with any number of covariates

First we explain the idea in two-covariate models. For example, consider as an alternative model

$$\gamma(x_1, x_2) = \gamma(x_1, x_2; \boldsymbol{\theta}) + \sum_{j \in \Lambda} a_j \cos(\pi j x_k), \quad k = 1 \text{ or } 2,$$

where only one of the covariates is used to distinguish from the null model. Of course, other basis functions may be used. Unless one has prior belief that only x_k would cause the lack of fit from the null model, this approach is not recommended. Instead we could take as our test statistic the maximum of the values obtained by

looking at each covariate "direction" separately; we refer to this as the *max test*. The level of this test can be controlled by application of Bonferroni's inequality. Both the diagonal and max tests are used in the simulation study of the next section.

This same idea can be used to extend the domain of application to models with more than two covariates. For a model with d covariates we consider, for each pair (r, s) separately, alternatives

$$\gamma(x_1, \ldots, x_d) = \gamma(x_1, \ldots, x_d; \boldsymbol{\theta}) + \sum \sum_{(j,k)\in\Lambda} \alpha_{jk} \cos(\pi j x_r) \cos(\pi k x_s). \qquad (7.14)$$

For a particular choice of (r, s), we may perform any of the tests for two-covariate models, e.g. one of the omnibus tests following path (d). Next, we take the maximum of all $d(d-1)/2$ test statistics. If the number of covariates d is large, using Bonferroni's inequality will result in a very conservative test, instead a bootstrap procedure might be applied. It seldom occurs that relevant hypotheses contain more than four or five covariates, usually a model selection stage is passed before testing a specific hypothesis.

7.4.4 Tests for more specific alternatives

The test in section 7.4.2 is not necessarily consistent unless the alternative is additive. This additivity assumption can be tested by a modification of the omnibus tests. Consider the null hypothesis

$$H_0 : \gamma(x_1, x_2) = \gamma_1(x_1; \boldsymbol{\theta}_1) + \gamma_2(x_2; \boldsymbol{\theta}_2).$$

A general additive model can be estimated by a Fourier series, but now the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is infinite dimensional. A possible approach to this problem is to first estimate, by use of a model selection criterion, optimal orders k_1 and k_2 for series estimates of $\gamma_1(\,\cdot\,; \boldsymbol{\theta}_1)$ and $\gamma_2(\,\cdot\,; \boldsymbol{\theta}_2)$. Then, the null model based on these estimated orders is extended with interaction terms according to one of the paths (a)-(d). It is not immediately clear how the additional model selection step will affect the testing procedure. This is an interesting question that will be addressed in future research.

Alternatively, one could perform tests of additivity for a large, fixed value of $k = k_1 = k_2$. A sensitivity analysis on k might show to what extent the choice of k_1 and k_2 influences the final conclusion. In Section 7.6.3 this method will be used on heart-attack data that were used by Hastie and Tibshirani (1990) to illustrate techniques on generalized additive models.

Finally, we also mention a *goodness of link* or *single-index* test. In this case the hypothesized model (7.12) is contrasted with alternative models of the following form:

$$\gamma(x_1, x_2) = \gamma(x_1, x_2; \boldsymbol{\theta}) + \sum_{j \in \Lambda} a_j u_j(\gamma(x_1, x_2; \boldsymbol{\theta})).$$

For generalized linear models, this provides a way of testing the adequacy of the link function. It is an alternative to methods described by Collett (1991, Section 5.3) and Brown (1982).

7.5 Simulations in multiple regression models

In a limited simulation study we compared the five tests S_b, S_a, T_a, T_{os} and $T_{\max}$. We generated the covariates X_{1i} and X_{2i} i.i.d. from a uniform distribution on (0,1), $i = 1, \ldots, 100$. Conditional on these values, independent normal response data $Y_i \sim \mathcal{N}(\gamma(x_{1i}, x_{2i}), \eta)$ were generated. We test the no-effect null hypothesis using the normalized Legendre polynomials $u_j(\cdot) = \mathcal{L}_j(\cdot)$. As before, we computed critical points of the tests based on 30000 replications. These can be found in Table 7.3. Note that the critical values for S_b are those of a χ^2_2 distribution, since for each of the model sequences S_b converges in distribution to χ^2_2 under H_0. In Table 7.4, based on 5000 simulated datasets under the null hypothesis, type I error probabilities are estimated. As in Table 7.1, the levels for the S_b tests are somewhat too large, and at the nominal level of 1%, the level of S_a is too small. There is no substantial difference for model sequences (c) and (d).

To study the power characteristics of the different tests and model sequences, we first considered the alternative models

$$\gamma(x_1, x_2) = 2\mathcal{L}_{m_0}(x_1) + 2\mathcal{L}_{n_0}(x_2) \quad \text{and} \quad \eta = 1. \tag{7.15}$$

A larger value of m_0 or n_0 corresponds to a higher frequency for the corresponding covariate. The various tests were performed for several choices of m_0 and n_0. In the first column of Figure 7.5 we used the "additive" paths from Section 7.4.2 and in the second column the "interaction" paths (c) and (d) (Figure 7.3), although there were no interactions present in the alternative models.

As before, critical points were obtained by simulation, based on 5000 data sets generated under the null hypothesis. The level of significance is equal to 0.05. The

	Diagonal		
α:	0.10	0.05	0.01
S_b	4.6052	5.9915	9.2103
S_a	8.3112	13.2128	26.5964
T_a	2.9439	3.9560	6.0710
T_{os}	2.5584	3.1534	4.6517
$T_{\max}$	1.6998	3.4562	7.7223

	Model sequence (c)			Model sequence (d)		
α:	0.10	0.05	0.01	0.10	0.05	0.01
S_b	4.6052	5.9915	9.2103	4.6052	5.9915	9.2103
S_a	8.0361	13.4481	26.4062	8.0240	13.5353	27.0476
T_a	3.0740	4.1006	6.1019	3.0740	4.1006	6.1019
T_{os}	2.6232	3.2142	4.7687	2.6169	3.2521	4.7259
$T_{\max}$	1.9300	3.6530	7.7622	1.9071	3.6275	7.7331

Table 7.3: Critical points for the multiple regression case

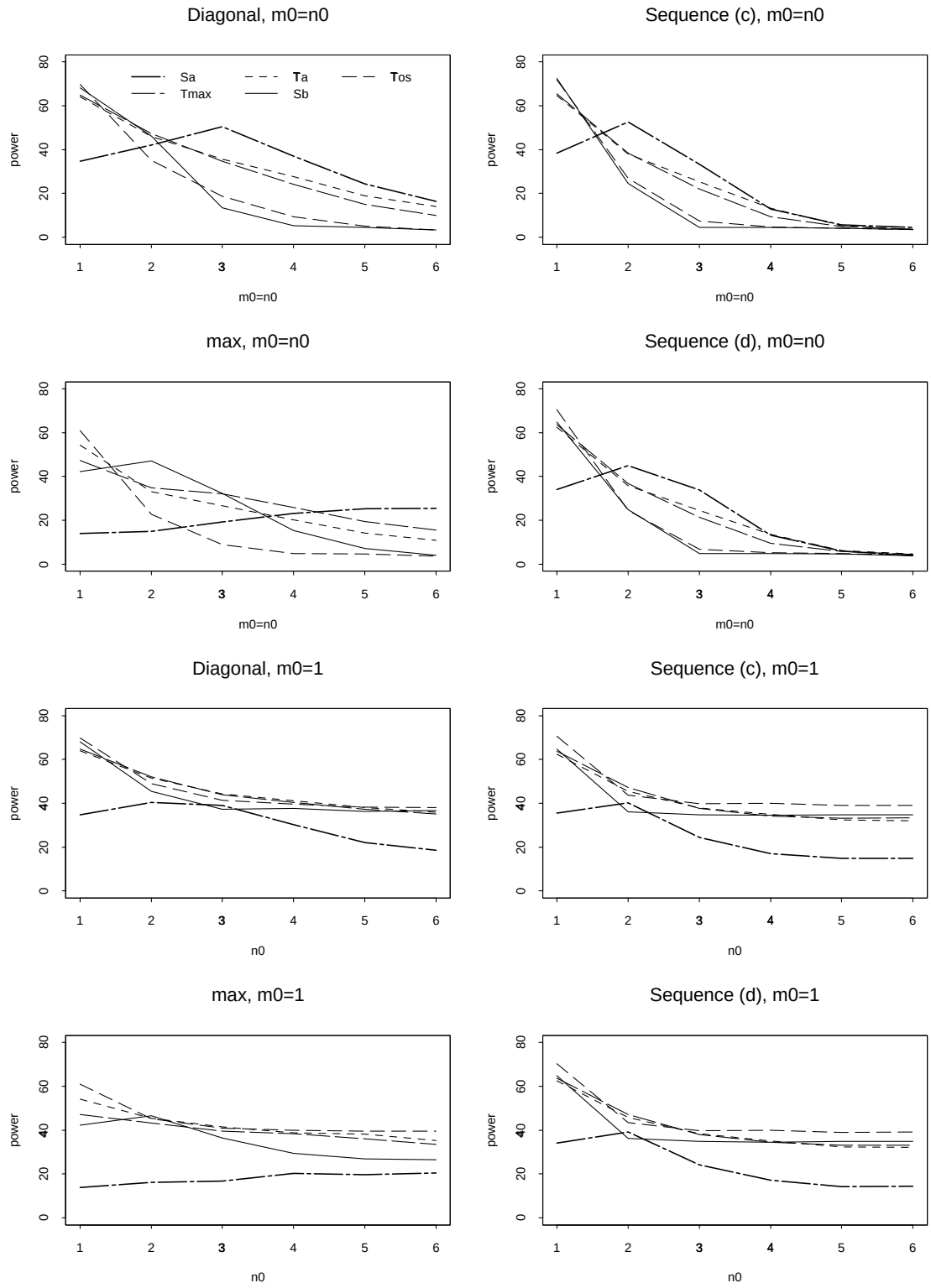


Figure 7.5: Simulated power curves when true model is additive.

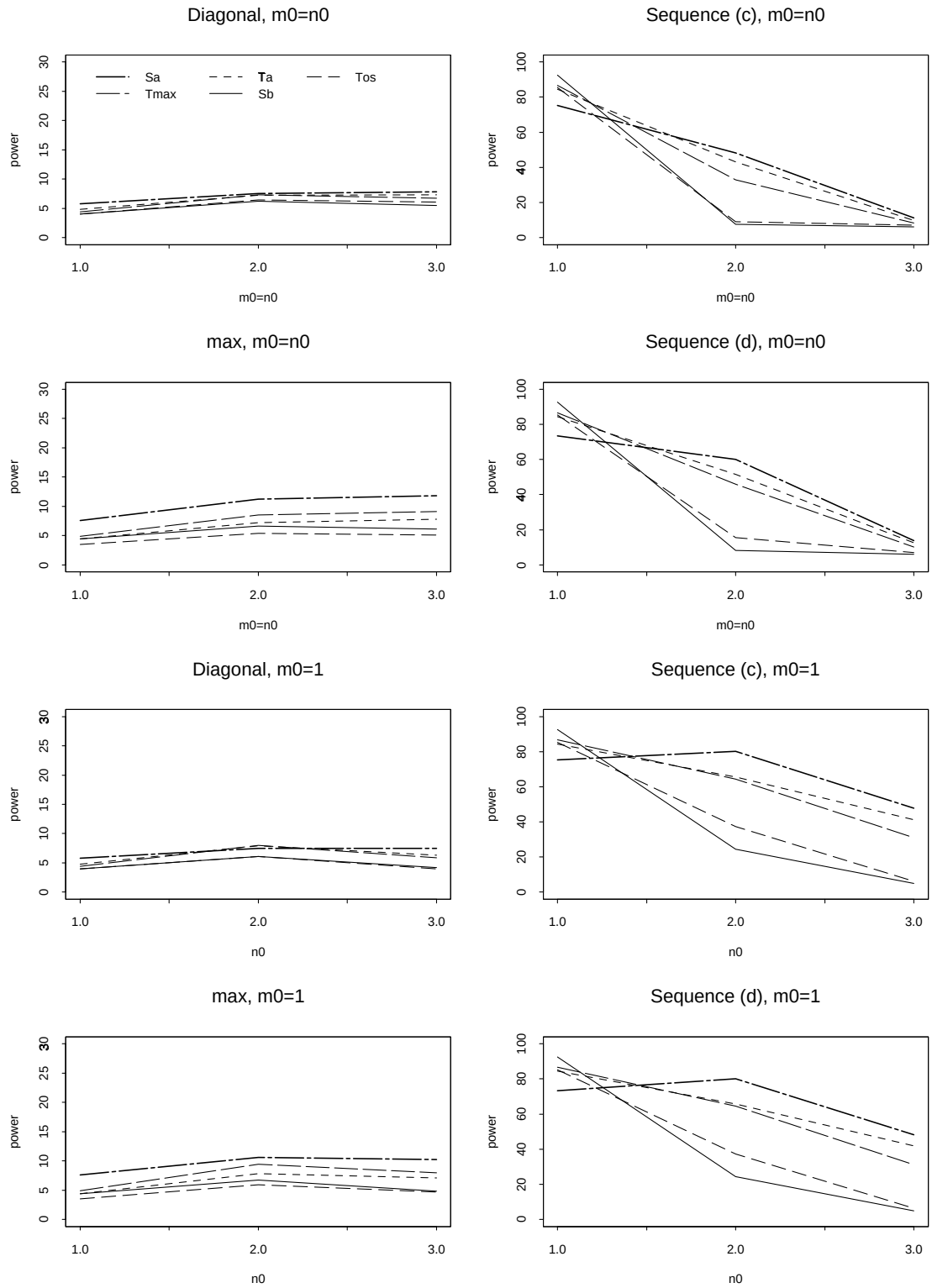


Figure 7.6: Simulated power curves when the true model has interaction structure.

	Diagonal			Max		
Statistic	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
S_b	0.1076	0.0582	0.0192	0.1650	0.1130	0.0420
S_a	0.0948	0.0420	0.0016	0.0728	0.0256	0.0002
T_a	0.0942	0.0414	0.0044	0.0848	0.0410	0.0064
T_{os}	0.0950	0.0480	0.0076	0.0970	0.0514	0.0074
$T_{\max}$	0.0952	0.0434	0.0044	0.0864	0.0402	0.0036
	Sequence (c)			Sequence (d)		
Statistic	α=0.10	α=0.05	α=0.01	α=0.10	α=0.05	α=0.01
S_b	0.1292	0.0700	0.0182	0.1292	0.0700	0.0182
S_a	0.0936	0.0444	0.0046	0.0950	0.0454	0.0042
T_a	0.0868	0.0416	0.0060	0.0868	0.0426	0.0062
T_{os}	0.0946	0.0442	0.0082	0.0960	0.0426	0.0086
$T_{\max}$	0.0872	0.0448	0.0062	0.0880	0.0460	0.0066

Table 7.4: Simulated Type I error probabilities

simulated powers were based on 1000 replications. In the four upper graphs, we chose $m_0 = n_0$, in which case both covariates have the same frequency. The diagonal path is designed to be best for this kind of alternative. Remarkably, except for the max test, the curves are ordered in the same way as in the simulation results for the univariate case. Whereas the power curves for S_a and S_b cross at $m_0 = n_0 = 2$ for the diagonal and omnibus test, this crossing is at about $m_0 = n_0 = 4$ for the max test. This is probably related to the fact that the max test is based on two univariate tests, for which the alternative model sequences increase by only one term at a time. A comparison of the left- and right-hand graphs shows that there is some loss in power when the interaction paths are used.

In the four lower graphs, the frequency of the first covariate is fixed at 1, and the frequency of the second covariate varies between 1 and 6. In this setting, the loss in power of the omnibus tests is usually negligible. For all power curves there seems to be a change point at frequency $n_0 = 2$, at which point the power decreases very slowly. The explanation for this phenomenon might be the dominance of the lowest frequency term ($m_0 = 1$) in all alternative models ($n_0 = 1, \ldots, 6$). The S_a test has remarkably low power. Here, standardizing has a beneficial effect on S_a for all alternatives. Finally, for all cases we observe that there is no substantial difference between model sequences (c) and (d).

In Figure 7.6 simulation results are shown for the interaction type alternative models

$$\gamma(x_1, x_2) = 4\mathcal{L}_{m_0}(x_1)\mathcal{L}_{n_0}(x_2) \quad \text{and} \quad \eta = 0.1. \tag{7.16}$$

A larger value of m_0 or n_0 corresponds to a higher order interaction. Note that some of the cases considered are quite extreme, e.g. $m_0 = n_0 = 3$ corresponds to a product of two 3rd degree, orthogonalized polynomials. As expected, for this kind of alternative, the additive type tests have almost no power. For small values of m_0 and n_0 (we are testing for no effect) the power of the interaction tests is relatively high. As before, but now even more pronounced, there are two groups of tests, with the AIC-based tests being preferred. Again, path (c) and (d) tests are almost indistinguishable.

In Table 7.5 we show simulation results obtained under alternative (7.15) when using the Legendre polynomial basis for a diagonal and a step-diagonal path. The simulated powers are shown as percentages, the corresponding level of significance equals 0.05. For each of the five test statistics, the one with the highest power (per

		Diagonal					Step-diagonal				
m_0	n_0	S_b	S_a	T_a	T_{os}	$T_{\max}$	S_b	S_a	T_a	T_{os}	$T_{\max}$
1	1	68.1	34.7	64.0	**69.7**	64.8	**63.5**	34.2	61.1	63.1	60.8
2	2	46.1	42.0	46.0	35.1	**47.3**	28.4	41.7	38.4	18.7	**42.7**
3	3	13.4	**50.4**	35.6	18.8	34.7	8.5	**49.1**	29.5	6.4	31.1
4	4	5.3	**37.1**	27.7	9.3	24.2	3.9	**38.8**	22.0	3.9	21.8
5	5	4.4	**24.4**	4.3	5.0	15.1	4.6	**24.6**	14.2	3.8	13.5
6	6	3.4	**16.3**	14.1	3.3	9.9	4.1	**16.7**	10.7	3.0	7.9
1	2	45.4	40.3	51.5	49.0	**52.0**	40.1	37.1	**51.8**	50.9	50.5
1	3	37.2	39.1	**44.3**	41.3	43.9	37.5	34.4	45.4	**48.4**	44.3
1	4	37.6	30.2	**41.1**	39.7	40.4	37.2	25.8	44.0	**49.0**	41.6
1	5	36.3	22.0	38.0	**38.2**	37.2	37.1	19.8	41.7	**48.6**	38.0
1	6	36.7	18.5	36.0	**38.1**	35.2	37.6	15.1	39.7	**49.5**	37.1
2	1	44.5	38.8	49.5	46.5	**50.7**	**53.9**	37.5	47.6	34.7	50.7
2	3	21.7	**47.2**	34.9	20.6	34.8	18.0	**42.5**	28.9	11.5	32.5
2	4	19.5	**37.6**	31.8	15.4	29.9	17.8	**34.4**	26.1	10.2	27.7
2	5	18.0	**25.6**	24.5	14.3	23.1	18.0	**23.7**	21.8	10.5	22.9
2	6	18.7	**22.9**	21.9	14.3	21.0	18.2	21.0	20.1	10.7	**21.6**
3	1	35.1	40.0	**43.2**	39.3	43.0	41.2	**43.1**	38.6	23.9	42.0
3	2	22.1	**48.1**	37.2	21.4	37.2	15.2	**48.8**	33.8	10.6	37.4
3	4	6.1	**43.1**	29.9	9.9	27.0	6.0	**41.8**	24.2	4.3	24.8
3	5	5.6	**34.1**	23.8	7.3	19.9	6.4	**32.7**	18.6	4.6	19.7
3	6	4.8	**29.7**	19.4	6.0	16.1	5.8	**27.8**	15.6	4.3	16.5

Table 7.5: Simulated powers (as %) when true model is additive, significance level 0.05.

combination of m_0 and n_0 and per path) is printed bold-faced. The step-diagonal path is constructed such that a basis element in x_1 enters first, next a basis element in x_2, ... Especially for the order selection test T_{os} and for the standardized test T_a, this ordering seems to be important, as the powers of these tests for the situation where m_0 is strictly smaller than n_0 (e.g., m_0=1 and n_0=3) are considerable higher than when these orders are reversed (m_0=3 and n_0=1). This clearly demonstrates the usefulness of a path which is symmetric in the covariates, unless of course, one has prior knowledge about the orders of the functional relationships in which the covariates enter the model. In case $m_0 = n_0$, the power of the tests using the step-diagonal path is at most settings smaller than the power of the corresponding tests using the diagonal path. This is already suggested by Figure 7.4 where it is clear that a step-diagonal test needs more steps to arrive at a place on the "diagonal", than a diagonal test, where two terms are added at each step.

We now illustrate the tests of Section 7.4.3 on a regression model with 4 covariates. Data are generated from the model $Y_i = 2x_1 + x_2 - 5x_3 + x_4 + 0.2\varepsilon_i$, where the errors ε_i are independent from a standard normal distribution. The covariates are all from a uniform distribution on (0, 1). The same design is used in all simulated data sets, the sample size equals 50. Test statistics, using a polynomial basis, and P-values based on a parametric bootstrap, of size 1000, are calculated using the Gauss software.

In the upper part of Table 7.6 we show the results of the "max test", which looks in all possible two-dimensional directions, using path (d) in each direction. These results should be compared to those of the lower part of Table 7.6 which only look in the directions of those variables in the alternative model. The last ones are referred to as "oracle tests", since they use information that normally only an oracle can know. The alternative models add the following terms to the null model under H_0, A_1: $0.05x_1^4$ is a deviation only in one specific direction; B_2: $0.8x_1x_2$; B_3: $0.8x_1x_2 + 5\log(x_2+1)$ are determined by two covariates, C_4: $0.8x_1x_2 + 5\log(x_4+1)$; C_5: $0.5x_1x_2x_3$ live in a three dimensional covariate space and D_6: $0.05x_1^4 + 0.25x_1x_2 + 0.25x_1x_2x_3 + 0.5\log(x_4+1)$ contains all four covariates. The test procedure in this simulation study adds twelve additional terms to the null model in each direction.

Results for $T_{\max}$ are nearly identical with those of T_a, except for alternative A_1 where its power at the 5% level is 0.333 (resp. 0.157 at 1%). The simulated power of S_a is everywhere the lowest, also except for A_1 where the obtained value was

		Max test						
		H_0	A_1	B_2	B_3	C_4	C_5	D_6
$\alpha = 0.05$	S_b	0.048	0.179	0.928	0.854	0.915	1.000	0.660
	S_a	0.045	0.577	0.955	0.504	0.628	0.682	0.615
	T_a	0.050	0.153	0.927	0.850	0.917	1.000	0.680
	T_{os}	0.053	0.559	0.734	0.684	0.730	0.940	0.808
	$T_{\max}$	0.049	0.333	0.925	0.853	0.912	1.000	0.782
$\alpha = 0.01$	S_b	0.011	0.159	0.783	0.684	0.766	0.983	0.302
	S_a	0.013	0.161	0.131	0.081	0.132	0.023	0.218
	T_a	0.011	0.016	0.787	0.677	0.769	0.986	0.262
	T_{os}	0.013	0.317	0.518	0.463	0.488	0.581	0.552
	$T_{\max}$	0.012	0.157	0.782	0.680	0.767	0.981	0.402
			Oracle test					
$\alpha = 0.05$	S_b		0.902	0.982	0.959	0.936	1.000	
	S_a		0.898	0.978	0.955	0.924	0.994	
	T_a		0.901	0.980	0.957	0.943	1.000	
	T_{os}		0.793	0.866	0.851	0.829	0.978	
	$T_{\max}$		0.903	0.981	0.959	0.945	1.000	
$\alpha = 0.01$	S_b		0.733	0.936	0.868	0.804	0.977	
	S_a		0.603	0.620	0.501	0.272	0.087	
	T_a		0.731	0.935	0.865	0.789	0.997	
	T_{os}		0.578	0.722	0.678	0.571	0.726	
	$T_{\max}$		0.730	0.931	0.863	0.804	0.996	

Table 7.6: Simulated powers of tests in a Gaussian regression model with 4 covariates.

0.577 (resp. 0.161). Especially at the 1% level, powers for S_a are extremely low, even going to 0.023 for C_5 at 1%. If the deviation is only in one specific direction, as expected, there will be some loss in power when looking in all six two-dimensional directions. This picture already changes quite a bit at alternative model B_2 where the deviation from the null model is in two-dimensions. At the 5% level, there is almost no loss in power for test statistics S_b, T_a and $T_{\max}$, while the loss is rather big for S_a. For alternatives C_4 and C_5, the poor behavior of S_a is remarkable, and results for the other max tests are comparable to those of the oracle tests.

7.6 Examples

Here we apply our proposed tests to three data sets. Two examples concern binary response data with two covariates. For the second example, where there are three covariates, we use a Gaussian model.

7.6.1 The POPS Data

The Project On Preterm and Small-for-gestational age infants (POPS) collected information on 1338 infants born in the Netherlands in 1983 and having gestational age (x_1) less than 32 weeks and/or birthweight (x_2) less than $1500g$; see Verloove, et al., 1986 for more details. The outcome of interest here concerns the situation after 2 years. The binary variable Y is 1 if an infant has died within 2 years after birth or survived with a major handicap, and 0 otherwise. After deleting observations with missing data, 1310 infants remain in the dataset.

Le Cessie and van Houwelingen (1991, 1993) examined these data to illustrate a lack-of-fit test based on a weighted sum of kernel smoothed standardized residuals. Their test failed to reject the null hypothesis of a logistic model having linear and quadratic terms in both covariates x_1 and x_2. Likewise, a likelihood ratio test showed that neither one of the third-order terms nor a first-order interaction term contributes significantly to the model.

Table 7.7 shows the values of the omnibus (sequence (d)), additive (diagonal and max) and single-index test statistics. The Legendre polynomials $\mathcal{L}_k(x_1)$ and $\mathcal{L}_\ell(x_2)$ were used to represent all models. With $\mathcal{L}_0(x) \equiv 1$, the null hypothesis can

	S_b	S_a	T_a	T_{os}	T_{max}	$\hat{r}_a$	$\hat{r}_b$
omnibus	0.42	10.64	4.41	3.55	4.64	2	1
P_∞	0.811	0.073	0.039	0.036	0.034		
P_B	0.532	0.034	0.031	0.071	0.016		
diagonal	3.03	0.00	0.00	1.52	0.00	0	1
$\max(x_1, x_2)$	2.33	2.33	1.33	2.33	0.33	1	1
single index	3.52	8.43	0.99	3.52	2.43	6	2

Table 7.7: POPS data: Results of testing H_0: "model is quadratic in x_1 and x_2".

be written as

$$H_0 : \text{logit}(E(Y)) = \sum_{k=0}^{2} \alpha_{k0}\mathcal{L}_k(x_1) + \sum_{\ell=1}^{2} \alpha_{0\ell}\mathcal{L}_\ell(x_2).$$

We considered alternative additive models extending this null model by extra terms $\mathcal{L}_k(x_1)$ and $\mathcal{L}_\ell(x_2)$ with $k, \ell = 3, \ldots, 15$. For the alternative models allowing interaction terms, we included the above main effects up to the sixth order together with all interaction terms $\mathcal{L}_k(x_1)\mathcal{L}_\ell(x_2)$ where $2 \leq k + \ell \leq 6$. For the omnibus tests, Table 7.7 also shows P-values based on the asymptotic distribution (P_∞) and on the bootstrap (P_B). Using a parametric bootstrap, 999 replications were generated under H_0 (using the null estimates). All omnibus tests except S_b indicate some evidence against H_0. The different behavior of S_b is a consequence of too large a penalty for large samples ($C_n = \log n \approx 7.18$). No additive or single index test is significant. These results suggest an extension of the null model by certain interactions. The value of $\hat{r}_a = 2$ corresponds to the interaction terms (Figure 7.3d) $\mathcal{L}_1(x_1)\mathcal{L}_1(x_2)$, $\mathcal{L}_1(x_1)\mathcal{L}_2(x_2)$ and/or $\mathcal{L}_2(x_1)\mathcal{L}_1(x_2)$.

Inspired by these findings we investigated numerous new models extending the null model and for each we computed the classical model selection criterion (7.4) with C_n=2 (AIC) and $C_n = \log n$ (BIC). The BIC criterion selected the null model, which, again, is a consequence of the large penalty. The model selected by AIC is

$$\text{logit}(E(Y)) = \sum_{k=0}^{5} \alpha_{k0}\mathcal{L}_k(x_1) + \sum_{\ell=1}^{2} \alpha_{0\ell}\mathcal{L}_\ell(x_2) + \alpha_{11}\mathcal{L}_1(x_1)\mathcal{L}_1(x_2) + \alpha_{12}\mathcal{L}_1(x_1)\mathcal{L}_2(x_2) \tag{7.17}$$

which can be rewritten in terms of $x_1, \ldots, x_1^5, x_2, x_2^2, x_1x_2, x_1x_2^2$. Both groups of higher order main effects x_1^3, x_1^4, x_1^5 and interactions $x_1x_2, x_1x_2^2$ are significant at the 5% level.

7.6.2 The peanuts data

These data come from an experiment concerning a device for automatically shelling peanuts (Dickens and Mason, 1962). The data consist of the grams of unshelled peanuts, out of one kilogram, which is the response variable Y, covariates are number of strokes per minutes (x_1), and length of stroke (x_2) and bar grid spacing (x_3) both in inches. There are 67 observations.

The null hypothesis of interest is a quadratic polynomial response surface, see Freund and Littell (1991, Sec. 5.6). Using the "max" approach as described in Section 7.4.3, we added 15 terms to this null model according to model sequence (d) and used polynomial basis functions. Adding more terms gave identical conclusions. Based on 1000 bootstrap replicates, the P-values of T_a, T_{os} and $T_{\max}$ were, respectively, 0.0839, 0.0210 and 0.0699. Since there are replicated observations, the classical F-test for lack of fit may also be applied. Using SAS procedure `RSREG`, with the lackfit option, one gets a P-value of 0.0997. This also indicates that there is some evidence against the nulhypothesis of a quadratic response surface.

7.6.3 The heart-attack data

The tests for additivity from Section 7.4.2 are illustrated on heart-attack data that are described in Hastie and Tibshirani (1990). The binary response variable Y indicates the presence of myocardial infarction at the time of survey for 463 white males between 15 and 64 years of age. We will use the same covariates as in Hastie and Tibshirani's additive model, namely, x_1, the systolic blood pressure and x_2, a cholesterol ratio. The hypothesized null model is

$$\text{logit}(E(Y)) = \beta_0 + \sum_{j=1}^{k_1} \beta_{1j} u_{1j}(x_1) + \sum_{j=1}^{k_2} \beta_{2j} u_{2j}(x_2) \tag{7.18}$$

where we take $u_{j1}(\cdot) = u_{j2}(\cdot) = \mathcal{L}_j(\cdot)$.

Instead of using a data driven choice of the unknown orders k_1 and k_2 we computed the values of all test statistics for $k = k_1 = k_2$ in $\{4, 5, 6, 7, 8\}$. This shows

how sensitive the test statistics are for different choices of k_1 and k_2. We adapted model sequence (d) from Figure 7.3 by excluding the terms on the axes. So, for example, the null model for $k = 5$ is sequentially extended by adding

$$x_1x_2 \;/\; x_1^2x_2,\; x_1x_2^2 \;/\; x_1^2x_2^2 \;/\; x_1^3x_2,\; x_1x_2^3 \;/\; x_1^3x_2^2,\; x_1^2x_2^3 \;/\; x_1^4x_2,\; x_1x_2^4.$$

Table 7.8 gives the results for Legendre and cosine basis functions. It also shows for each k, values of $\hat{r}_a$ and $\hat{r}_b$, and P-values obtained by the parametric bootstrap using 999 replications.

For the Legendre polynomials, most P-values indicate that the additivity hypothesis can be rejected. The values of the test statistics and their corresponding P-values are comparable for different values of k. At least for this example, the choice of the degree of the additive null model is not really crucial. A subject of ongoing research is to examine in more generality the importance of this choice. Turning to the cosine basis, T_{os} and the BIC test S_b fail to reject additivity. The other tests are usually significant except when $k = 7$ or 8. Table 7.8 indicates that the choice of basis can be important.

7.7 Discussion

Different statistics for testing lack of fit in multiple regression have been proposed. Based on simulations, we prefer T_a and $T_{\max}$ in combination with the omnibus path (d). All tests are based on a penalized score criterion, which only requires computation of *null* parameter estimates. Alternatively, a penalized log likelihood criterion could be used. A detailed study of the pros and cons of each criterion seems to be worthwhile. Besides being computationally simpler, the score criterion has the advantage that it can be robustified against likelihood misspecification. A robust version of T_{os} is studied in Chapter 6. Similarly, robust versions of the other statistics proposed in the current chapter can be constructed.

Basis		k	S_b	S_a	T_a	T_{os}	$T_{\max}$	$\hat{r}_a$	$\hat{r}_b$
Polynomial	Value	4	0.047	11.119	4.687	3.706	5.119	2	1
	P_B		0.831	0.040	0.034	0.077	0.022		
	Value	5	13.229	13.229	5.906	4.410	7.229	2	2
	P_B		0.002	0.042	0.017	0.058	0.012		
	Value	6	12.946	12.946	5.743	4.315	6.946	2	2
	P_B		0.007	0.074	0.028	0.067	0.025		
	Value	7	13.068	13.068	5.813	4.356	7.068	2	2
	P_B		0.002	0.053	0.015	0.057	0.011		
	Value	8	14.987	33.349	6.738	4.996	11.349	7	2
	P_B		0.002	0.009	0.011	0.045	0.003		
Cosine	Value	4	0.012	11.830	3.915	3.256	3.830	3	1
	P_B		0.923	0.030	0.045	0.093	0.033		
	Value	5	0.011	12.211	4.105	3.343	4.211	3	1
	P_B		0.927	0.057	0.055	0.093	0.033		
	Value	6	0.079	26.046	4.536	2.797	4.046	7	1
	P_B		0.801	0.013	0.046	0.188	0.054		
	Value	7	0.004	25.561	4.390	3.091	3.561	7	1
	P_B		0.952	0.017	0.076	0.149	0.101		
	Value	8	0.009	9.277	3.624	3.092	3.277	2	1
	P_B		0.926	0.150	0.139	0.160	0.108		

Table 7.8: Heart-attack data. Test results for the additivity hypothesis.

Part III

Bootstrap procedures

Chapter 8

A Parametric Bootstrap Procedure for Testing the Fit of a Parametric Function

8.1 Introduction

In this chapter we will study a bootstrap technique which is used in combination with *parametric methods* for testing hypotheses. After having studied the *omnibus* lack of fit tests in previous chapters, one might wonder why we now take a step backwards (or should we say forwards) and study tests of a null hypothesis versus a very specific alternative hypothesis.

This research topic can be motivated in many ways. First of all, one might simply be interested in a very specific parametric alternative model; for example, in contrasting a quadratic with a linear effect. A more practical aspect arises from the typical design of a toxicity study (see, e.g., the THEO data). Such a study typically includes only a very small number of different dose levels. This corresponds to a fixed-design covariate, which takes a small (e.g., 4 or 5) number of values. We might now call in question the applicability of smoothing-based tests. The latter usually implicitly assume that the number of different covariate values increases with the sample size. For example, if there are only four dose levels (as in the THEO data), it does not make much sense to fit polynomial models of degree 25. With regard to the

tests of Chapter 6, this obviously restricts the choice of "r_n", the length of the model sequence. A reason for studying alternatives to classical test procedures, is a matter of caution. Can we safely "trust" the P-values of classical likelihood ratio, Wald and score tests, as provided by standard software packages, which mostly use asymptotic distribution theory? This question motivated us to look after alternatives, such as bootstrap methods.

As already mentioned in Section 1.3, bootstrap methods are not often used in combination with classical test statistics in parametric models. The methodology and theory for bootstrap hypothesis testing is still not well developed. The main difficulty is the generation of bootstrap data reflecting the null hypothesis. Assuming the true likelihood of the data to be known, this can be achieved by the parametric bootstrap based on the null estimates. In this chapter we will shown that the parametric bootstrap test leads to a substantial improvement in comparison with tests based on classical asymptotic distribution theory.

In practice however the assumed probability model can be wrong, in which case the parametric bootstrap leads to incorrect results. For complex data structures, there is (at least by our knowledge) no general nonparametric or semiparametric method available and "theoretical and empirical studies are still called for" (citing Shao and Tu, 1995, p.189). In Chapter 9 we will propose a semiparametric bootstrap approach.

An important advantage of the parametric bootstrap is that it easily allows to generate bootstrap samples reflecting the data mechanism under any null hypothesis (in contrast with a nonparametric bootstrap method). In this chapter the parametric bootstrap technique is studied for hypothesis tests, and illustrated with examples on clustered binary data. We focus attention on pseudolikelihood estimation methods for conditionally specified models. Geys, Molenberghs and Ryan (1997, 1999) show that the use of pseudolikelihood estimators results in moderate efficiency loss. They also propose likelihood ratio tests in the pseudolikelihood framework. As a consequence of not working with a fully specified likelihood, the asymptotic distribution of the pseudolikelihood ratio test statistic is a weighted sum of independent χ^2_1 variables. The weights are unknown eigenvalues and have to be estimated. The estimators can be calculated under the null hypothesis but also under the alternative hypothesis. All this complicates the computation of critical points and P-values. In Theorem 8.4 it is shown theoretically that the parametric bootstrap leads to a

consistent estimator for the distribution of the pseudolikelihood ratio (PLR) test statistic. The bootstrap approach does not need any additional estimation of unknown eigenvalues. It automatically corrects for the "misspecification" of the joint distribution.

There are only a few papers in which bootstrap techniques are used for clustered binary data. They mainly use the nonparametric bootstrap method to estimate the variance of parameter estimators (Lockhart, Piegorsch and Bishop, 1992, Carr and Portier, 1993). Frangos and Schucany (1995) proposed a particular parametric bootstrap procedure to construct improved confidence intervals in certain toxicological experiments.

This chapter is organized as follows. Section 8.2 summarizes the basic asymptotic properties of the maximum pseudolikelihood estimators and tests. These results are reconsidered for the bootstrapped pseudolikelihood estimators in Section 8.3 (estimation) and 8.4 (hypothesis tests). Next to the PLR test we also include results for the robust Wald and the robust score test. The theoretical results in these sections are presented in the general framework of multiparameter pseudolikelihood when sampling from a finite number of associated populations. According to Bradley and Gart (1962), associated populations are distinct but related, in the sense that they may have some parameters in common. As indicated in Section 8.5, the analysis of clustered binary data from toxicity studies fits into this framework. This last section presents the results of a finite sample simulation study and some data examples.

The results of Chapter 8 can also be found in Aerts and Claeskens (1999).

8.2 Pseudolikelihood estimation and inference

The basic asymptotic properties of maximum pseudolikelihood estimators and hypothesis tests, when sampling from a finite number of associated populations, combine classical maximum likelihood theory as presented in, e.g., Serfling (1980) and Lehmann (1983), the extension to associated populations (Bradley and Gart, 1962) and pseudolikelihood estimation and testing (Geys, Molenberghs and Ryan, 1997, 1999). The proofs of the theorems stated in this section are omitted.

Let $f_i(\boldsymbol{y}, \boldsymbol{\theta})$ with $\boldsymbol{y} = (y_1, \ldots, y_{m_i}) \in R_i$ $(i = 1, \ldots, p)$ denote p joint density or discrete probability functions. The support R_i does not depend on the unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$. Let $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$ be independent random vari-

ables with common joint density function $f_i(\boldsymbol{y}, \boldsymbol{\theta})$. Define A_i as the set of all $2^{m_i} - 1$ vectors $\mathbf{a}$ of length m_i, consisting solely of zeros and ones, with each vector having at least one nonzero entry. Denote by $\boldsymbol{Y}_{ij}^{(\mathbf{a})}$ the subvector of $\boldsymbol{Y}_{ij}$ corresponding to the nonzero components of $\mathbf{a}$ with associated joint density function $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$. The logarithm of the parametric pseudolikelihood is defined as,

$$\log \mathrm{PL}_n(\boldsymbol{\theta}) = \sum_{i=1}^{p} \sum_{\mathbf{a} \in A_i} \gamma_{\mathbf{a}} \sum_{j=1}^{n_i} \log f_i^{(\mathbf{a})}(\boldsymbol{y}_{ij}^{(\mathbf{a})}, \boldsymbol{\theta}) \tag{8.1}$$

with $\{\gamma_{\mathbf{a}} | \mathbf{a} \in A_i\}$ a set of $2^{m_i} - 1$ real numbers, not all zero. We refer to expression (3.1) for a definition in a nonparametric context. Classical maximum likelihood corresponds to $\gamma_{\mathbf{a}} = 1$ for $\mathbf{a} = (1, \ldots, 1)$ and zero otherwise. Another typical choice is the full conditional log pseudolikelihood where $\gamma_{1_{m_i}} = m_i$ and $\gamma_{\mathbf{a}_\ell} = -1$ for $\ell = 1, \ldots, m_i$ where 1_{m_i} is a vector of ones and $\mathbf{a}_\ell$ consists of ones everywhere, except for the ℓth entry.

The number p of (possibly) different populations is considered as fixed whereas the numbers n_i of observations from the distinct populations become large as $n = \sum_{i=1}^{p} n_i$ tends to infinity, according to $n_i/n \to \lambda_i$ where $\sum_{i=1}^{p} \lambda_i = 1$ with all $\lambda_i > 0$. Before stating the main asymptotic properties of the maximum pseudolikelihood estimators, the Wald, the score and the pseudolikelihood ratio test, we first list the required regularity conditions on the density functions $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$ for all $\mathbf{a}$ in $A_i^\circ = \{\mathbf{a} \in A_i | \gamma_{\mathbf{a}} \neq 0\}$ and each $i = 1, \ldots, p$. Let μ denote the Lebesgue measure for continuous $\boldsymbol{Y}_{ij}^{(\mathbf{a})}$ and the counting measure if $\boldsymbol{Y}_{ij}^{(\mathbf{a})}$ is discrete. For $\boldsymbol{\theta}$ in an open set Ω containing the true value $\boldsymbol{\theta}_0$,

(R1) The densities $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$ are distinct for different values of the parameter $\boldsymbol{\theta}$; the supports $R_i^{(\mathbf{a})}$ of $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$ do not depend on $\boldsymbol{\theta}$.

(R2) Second order partial derivatives of $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ exist and may be passed under the integral sign in $\int f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta}) d\mu(\boldsymbol{y}^{(\mathbf{a})})$.

(R3) For each $k, \ell = 1, \ldots, r$,

$$E_\theta \left[\left| \frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell} \right| \right] < \infty$$

and the $r \times r$ matrix $\boldsymbol{J}(\boldsymbol{\theta}_0)$ as defined in Theorem 8.1 is positive definite.

(R4) Third order partial derivatives of $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ exist and there exist functions $K_1(\boldsymbol{y}^{(\mathbf{a})})$ and $K_2(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{y}^{(\mathbf{a}')})$ such that, for each $\mathbf{a}, \mathbf{a}'$ in A_i° and each

$k, \ell, m = 1, \ldots, r$,

$$\left|\frac{\partial^3 \log f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial\theta_k \partial\theta_\ell \partial\theta_m}\right| \leq K_1(\boldsymbol{y}^{(\mathbf{a})}),$$

$$\left|\frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial\theta_k \partial\theta_\ell} \frac{\partial \log f_i^{(\mathbf{a}')}(\boldsymbol{y}^{(\mathbf{a}')}, \boldsymbol{\theta})}{\partial\theta_m}\right| \leq K_2(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{y}^{(\mathbf{a}')})$$

and $E_\theta[K_1(\boldsymbol{Y}_{i1}^{(\mathbf{a})})]$ and $E_\theta[K_2(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{Y}_{i1}^{(\mathbf{a}')})]$ are, as functions of $\boldsymbol{\theta}$, uniformly bounded on Ω.

Theorem 8.1 guarantees the existence of at least one solution to the pseudolikelihood equations

$$\frac{\partial}{\partial\theta_\ell} \log \mathrm{PL}_n(\boldsymbol{\theta}) = 0 \quad \ell = 1, \ldots, r \tag{8.2}$$

which is strongly consistent and asymptotically normal.

Theorem 8.1 *Assume conditions (R1)-(R4). Then there exist solutions* $\widehat{\boldsymbol{\theta}}_n = (\widehat{\theta}_{n1}, \ldots, \widehat{\theta}_{nr})^T$ *to the pseudolikelihood equations (8.2) such that, as* $n \to \infty$ *and* $n_i/n \to \lambda_i$ *for all* $i = 1, \ldots, p$,

i) $\widehat{\boldsymbol{\theta}}_n$ *is strongly consistent for* $\boldsymbol{\theta}_0$

ii) $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *is asymptotically normal with mean vector* 0 *and covariance matrix*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1} \boldsymbol{K}(\boldsymbol{\theta}_0) \{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1}$$

with

$$[\boldsymbol{J}(\boldsymbol{\theta})]_{k\ell} = \sum_{i=1}^{p} \lambda_i \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} E_\theta \left[-\frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial\theta_k \partial\theta_\ell} \right]$$

and

$$[\boldsymbol{K}(\boldsymbol{\theta})]_{k\ell} = \sum_{i=1}^{p} \lambda_i \sum_{\mathbf{a} \in A_i^\circ} \sum_{\mathbf{a}' \in A_i^\circ} \gamma_{\mathbf{a}} \gamma_{\mathbf{a}'} E_\theta \left[\frac{\partial \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial\theta_k} \frac{\partial \log f_i^{(\mathbf{a}')}(\boldsymbol{Y}_{i1}^{(\mathbf{a}')}, \boldsymbol{\theta})}{\partial\theta_\ell} \right].$$

Next, consider the following hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \ \text{ versus } \ H_1 : \boldsymbol{\theta} \in \Theta \backslash \Theta_0$$

where Θ_0 is a $(r - t)$ dimensional subspace of the parameter space $\Theta \subset \mathbb{R}^r$ such that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$ belongs to Θ_0 if and only if $\theta_1 = \cdots = \theta_t = 0$, $1 \leq t \leq r$. More general situations, in which H_0 is of the form $H_0 : h_1(\boldsymbol{\theta}) = \cdots = h_t(\boldsymbol{\theta}) = 0$

for some smooth real-valued functions $h_1, \ldots, h_t$, can be put into this form by a reparametrization.

In maximum likelihood theory, the Wald, likelihood ratio and score tests are commonly used significance tests. Under appropriate regularity conditions, they are asymptotically equivalent in distribution, both under the null hypothesis and under local alternatives converging sufficiently fast (see, e.g., Serfling, 1980). Wald tests, however, are known not to be invariant to equivalent reparameterizations of nonlinear restrictions (see, e.g., Phillips and Park, 1988). Several authors studied and compared the three test procedures in more detail. See, e.g., Chandra and Joshi (1983), Chandra and Mukerjee (1984, 1985), Mukerjee (1990a, 1990b), Sutradhar and Bartlett (1993), and Cordeiro, Botter and Ferrari (1994). In the context of clustered binary data, several papers directly or indirectly compared these and other test procedures (see, e.g., Chapter 6 in Morgan, 1992). For pseudolikelihood, Geys, Molenberghs and Ryan (1999) propose, similar to the likelihood ratio test, a pseudolikelihood ratio (PLR) test.

The next theorem gives the asymptotic null distributions of the PLR, the robust Wald and the robust score test statistics. Let $\boldsymbol{C}^T = [\boldsymbol{I}_t \;\; \boldsymbol{0}_{t,r-t}]$ with $\boldsymbol{I}_t$ the $(t \times t)$ identity matrix and $\boldsymbol{0}_{t,r-t}$ the zero matrix of dimension $t \times (r-t)$. For an $r \times r$ matrix $\boldsymbol{A}$, define the partitioning

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{LL} & \boldsymbol{A}_{LR} \\ \boldsymbol{A}_{RL} & \boldsymbol{A}_{RR} \end{pmatrix}$$

where $\boldsymbol{A}_{LL} = \boldsymbol{C}^T\boldsymbol{A}\boldsymbol{C}$ and for any $r \times 1$ vector $\boldsymbol{v}$, let $\boldsymbol{v}_L$ (respectively $\boldsymbol{v}_R$) denote the subvector of the first t (respectively, last $r-t$) elements.

Theorem 8.2 *Assume conditions (R1)-(R4). Under the null hypothesis H_0, as $n \to \infty$ and $n_i/n \to \lambda_i$ for all $i = 1, \ldots, p$, we have*

i)

$$W_n = n(\widehat{\boldsymbol{\theta}}_{nL})^T\{\boldsymbol{\Sigma}_n(\widehat{\boldsymbol{\theta}}_n)_{LL}\}^{-1}\widehat{\boldsymbol{\theta}}_{nL}$$

converges in distribution to a χ^2_t random variable where

$$\boldsymbol{\Sigma}_n(\boldsymbol{\theta}) = \{\boldsymbol{J}_n(\boldsymbol{\theta})\}^{-1}\boldsymbol{K}_n(\boldsymbol{\theta})\{\boldsymbol{J}_n(\boldsymbol{\theta})\}^{-1}$$

with

$$[\boldsymbol{J}_n(\boldsymbol{\theta})]_{k\ell} = -\frac{1}{n}\sum_{i=1}^{p}\sum_{\mathbf{a}\in A_i^\circ}\gamma_{\mathbf{a}}\sum_{j=1}^{n_i}\frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial\theta_k\partial\theta_\ell}$$

and

$$[\boldsymbol{K}_n(\boldsymbol{\theta})]_{k\ell} = \frac{1}{n}\sum_{i=1}^{p}\sum_{\mathbf{a}\in A_i^\circ}\sum_{\mathbf{a}'\in A_i^\circ}\gamma_{\mathbf{a}}\gamma_{\mathbf{a}'}\sum_{j=1}^{n_i}\frac{\partial \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{(\mathbf{a})},\boldsymbol{\theta})}{\partial\theta_k}\frac{\partial \log f_i^{(\mathbf{a}')}(\boldsymbol{Y}_{ij}^{(\mathbf{a}')},\boldsymbol{\theta})}{\partial\theta_\ell}.$$

ii)

$$R_n = n\mathcal{H}_n(\widehat{\boldsymbol{\theta}}_n^\circ)_L^T(\{\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^\circ)\}^{-1})_{LL}\{\boldsymbol{\Sigma}_n(\widehat{\boldsymbol{\theta}}_n^\circ)_{LL}\}^{-1}(\{\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^\circ)\}^{-1})_{LL}\mathcal{H}_n(\widehat{\boldsymbol{\theta}}_n^\circ)_L$$

converges in distribution to a χ_t^2 *random variable where* $\widehat{\boldsymbol{\theta}}_n^\circ$ *is the maximum pseudolikelihood estimator over* Θ_0 *and*

$$\mathcal{H}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{a\in A_i^\circ}\gamma_a\sum_{j=1}^{n_i}\nabla_\theta \log f_i^{(a)}(Y_{ij}^{(a)},\boldsymbol{\theta})$$

with $\nabla_\theta \log f_i^{(a)}(y^{(a)},\theta)$ *the* $r\times 1$ *vector of partial derivatives of* $\log f_i^{(a)}(y^{(a)},\boldsymbol{\theta})$ *w.r.t.* $\boldsymbol{\theta}$.

iii) $-2\log\Lambda_n = 2(\log PL_n(\widehat{\boldsymbol{\theta}}_n) - \log PL_n(\widehat{\boldsymbol{\theta}}_n^\circ))$ *converges in distribution to* $\sum_{k=1}^{t}\alpha_k X_k$ *where* $\widehat{\boldsymbol{\theta}}_n$ *is the maximum pseudolikelihood estimator over* Θ, X_1, $\ldots$, X_t *are independent* χ_1^2 *random variables and* $\alpha_1 \geq \ldots \geq \alpha_t$ *are the eigenvalues of*

$$\{\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)_{LL}\}\{(\{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1})_{LL}\}^{-1}.$$

For the special case that $\gamma_{\mathbf{a}} = 1$ for $\mathbf{a} = (1,\ldots,1)$ and zero otherwise (classical maximum likelihood theory), the results of Theorem 8.1 and 8.2 simplify to the well-known results. Indeed, for this special case, $\boldsymbol{J}(\boldsymbol{\theta}_0) = \boldsymbol{K}(\boldsymbol{\theta}_0)$ such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \{\boldsymbol{J}(\boldsymbol{\theta}_0)\}^{-1}$. Hence all eigenvalues α_k are equal to 1 and the limiting distribution of $-2\log\Lambda_n$ reduces to a χ_t^2 distribution. Since for this special case the asymptotic null distribution does not depend on unknown parameters, the bootstrap test is expected to have a smaller asymptotic order of error in level. Beran (1988) shows that the bootstrap likelihood ratio test automatically accomplishes the Bartlett adjustment.

For all other choices, the asymptotic null distribution of the pseudolikelihood ratio test is rather complex (as a consequence of the "misspecification" of the joint distribution $f_i(\boldsymbol{y},\boldsymbol{\theta})$). The bootstrap can play an important role here. As shown in Section 8.4, the bootstrap estimator is a consistent estimator for the unknown distribution of $-2\log\Lambda_n$. No eigenvalues α_k have to be estimated.

8.3 A parametric bootstrap procedure

In this section we discuss the generation of the bootstrap samples and show how they can be used to estimate consistently the distribution of the maximum pseudolikelihood estimators $\widehat{\boldsymbol{\theta}}_n$. As before, let $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$ be a sample from the ith population. Based on these p samples, the maximum pseudolikelihood estimator $\widehat{\boldsymbol{\theta}}_n$ is computed. For each $i = 1, \ldots, p$, define the bootstrap sample $\boldsymbol{Y}^*_{i1}, \ldots, \boldsymbol{Y}^*_{in_i}$ as n_i independent random variables with common density function $f_i(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_n)$. Denote $\widehat{\boldsymbol{\theta}}^*_n$ the maximum pseudolikelihood estimator, which maximizes

$$\log \mathrm{PL}^*_n(\boldsymbol{\theta}) = \sum_{i=1}^{p} \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \sum_{j=1}^{n_i} \log f_i^{(\mathbf{a})}(\boldsymbol{y}_{ij}^{*(\mathbf{a})}, \boldsymbol{\theta}) \tag{8.3}$$

with $\boldsymbol{y}_{ij}^{*(\mathbf{a})}$ defined as before but now based on the bootstrap sample.

We need the following extra regularity conditions on the density functions $f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta})$, for all $\mathbf{a}$ in A_i° and each $i = 1, \ldots, p$.

(R5) For each $k, \ell = 1, \ldots, r$ and each $\mathbf{a}, \mathbf{a}'$ in A_i°,

$$E_\theta \left[\frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell} \right] \quad \text{and} \quad E_\theta \left[\frac{\partial \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \log f_i^{(\mathbf{a}')}(\boldsymbol{Y}_{i1}^{(\mathbf{a}')}, \boldsymbol{\theta})}{\partial \theta_\ell} \right]$$

are continuous as functions of $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and

$$E_\theta \left[\left(\frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell} \right)^2 \right]$$

is, as function of $\boldsymbol{\theta}$, uniformly bounded on Ω.

(R6) There exist a function $H^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})})$ and a $\delta > 0$ such that

$$\left| \left(\frac{\partial}{\partial \theta_k} \log f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{\theta}) \right)^{2+\delta} \right| \leq H^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})})$$

and $E_\theta[H^{(\mathbf{a})}(\boldsymbol{Y}_{i1}^{(\mathbf{a})})]$ is, as function of $\boldsymbol{\theta}$, uniformly bounded on Ω.

In what follows, P^*, E^*, Var^* stand for the bootstrap probability, expectation and variance, conditionally on $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$, $i = 1, \ldots, p$. The statements in the proofs hold, conditionally on $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$, $i = 1, \ldots, p$, for almost all sample paths $(\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \ldots)$, $i = 1, \ldots, p$.

Theorem 8.3 *Assume conditions (R1)-(R6). Then, for almost all sample paths* $(\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \ldots), i = 1, \ldots, p$, *there exist solutions* $\widehat{\boldsymbol{\theta}}^*_n$ *of the pseudolikelihood equations (8.3) such that, as* $n \to \infty$ *and* $n_i/n \to \lambda_i$ *for all* $i = 1, \ldots, p$,

i) $\widehat{\boldsymbol{\theta}}^*_n$ *converges in bootstrap probability to* $\boldsymbol{\theta}_0$

ii) $\sup_{\mathbf{t} \in I\!R^{r'}} \left| P^*\{g(n^{1/2}(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n)) \le \boldsymbol{t}\} - P\{g(n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)) \le \boldsymbol{t}\} \right| = o(1)$
with $g : I\!R^r \to I\!R^{r'}$ *a continuous function such that the distribution function of* $g(\boldsymbol{Z})$ *with* $\boldsymbol{Z}$ *a* r'*-dimensional normal distributed random variable, is continuous.*

Proof. Consider the expansion

$$H^*_k(\boldsymbol{\theta}) = H^*_k(\widehat{\boldsymbol{\theta}}_n) + \sum_{\ell=1}^{r} H^*_{k\ell}(\widehat{\boldsymbol{\theta}}_n)(\theta_\ell - \widehat{\theta}_{n\ell}) + \frac{1}{2} \sum_{\ell,m=1}^{r} H^*_{k\ell m}(\widetilde{\boldsymbol{\theta}}_n)(\theta_\ell - \widehat{\theta}_{n\ell})(\theta_m - \widehat{\theta}_{nm}) \tag{8.4}$$

where

$$H^*_k(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{p} \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \sum_{j=1}^{n_i} \frac{\partial \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{*(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k},$$

$$H^*_{k\ell}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{p} \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \sum_{j=1}^{n_i} \frac{\partial^2 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{*(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell}$$

and

$$H^*_{k\ell m}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{p} \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \sum_{j=1}^{n_i} \frac{\partial^3 \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{*(\mathbf{a})}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_\ell \partial \theta_m}$$

and with $\widetilde{\boldsymbol{\theta}}_n$ an interior point of the line segment joining $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_n$. The strong consistency of $\widehat{\boldsymbol{\theta}}_n$ (Theorem 8.1), (R2) and (R5) imply that as $n \to \infty$,

$$E^*[H^*_k(\widehat{\boldsymbol{\theta}}_n)], \quad \mathrm{Var}^*[H^*_k(\widehat{\boldsymbol{\theta}}_n)], \quad E^*[H^*_{k\ell}(\widehat{\boldsymbol{\theta}}_n)] + (\boldsymbol{J}(\boldsymbol{\theta}_0))_{k\ell} \quad \text{and} \quad \mathrm{Var}^*[H^*_{k\ell}(\widehat{\boldsymbol{\theta}}_n)]$$

all converge to zero from which it immediately follows that both $H^*_k(\widehat{\boldsymbol{\theta}}_n)$ and $H^*_{k\ell}(\widehat{\boldsymbol{\theta}}_n) + (\boldsymbol{J}(\boldsymbol{\theta}_0))_{k\ell}$ converge in bootstrap probability to zero as $n \to \infty$. Condition (R4) is sufficient to show that $H^*_{k\ell m}(\widetilde{\boldsymbol{\theta}}_n)$ is bounded in bootstrap probability. To complete the proof of (i), proceed as in Chanda (1954).

For $\boldsymbol{u} \in I\!R^r$ an arbitrary vector of unit norm, define for $j = 1, \ldots, n_i; i = 1, \ldots, p$

$$Z^*_{nij} = n_i^{-1/2} \boldsymbol{u}^T \left\{ \sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \nabla_\theta \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{*(\mathbf{a})}, \widehat{\boldsymbol{\theta}}_n) - E^* \left[\sum_{\mathbf{a} \in A_i^\circ} \gamma_{\mathbf{a}} \nabla_\theta \log f_i^{(\mathbf{a})}(\boldsymbol{Y}_{ij}^{*(\mathbf{a})}, \widehat{\boldsymbol{\theta}}_n) \right] \right\}.$$

$Z^*_{ni1}, \ldots, Z^*_{nin_i}$ are (conditionally) independent random variables with mean zero. With $\delta > 0$ from condition (R6), we have that, for n large enough

$$\left(E^*[|Z^*_{ni1}|^{2+\delta}]\right)^{1/(2+\delta)} \leq 2n_i^{-1/2} \sum_{k=1}^{r} \sum_{\mathbf{a} \in A_i^\circ} |\gamma_{\mathbf{a}}| \left\{E^*[H^{(\mathbf{a})}(\boldsymbol{Y}^{*(\mathbf{a})}_{i1})]\right\}^{1/(2+\delta)}$$

which is $O(n_i^{-1/2})$. Furthermore, using (R5), as $n \to \infty$,

$$n_i E^*[Z^{*2}_{ni1}] \to \sum_{\mathbf{a} \in A_i^\circ} \sum_{\mathbf{a}' \in A_i^\circ} \gamma_{\mathbf{a}} \gamma_{\mathbf{a}'} E_{\theta_0} \left[\frac{\partial \log f_i^{(\mathbf{a})}(\boldsymbol{Y}^{(\mathbf{a})}_{i1}, \boldsymbol{\theta}_0)}{\partial \theta_k} \frac{\partial \log f_i^{(\mathbf{a}')}(\boldsymbol{Y}^{(\mathbf{a}')}_{i1}, \boldsymbol{\theta}_0)}{\partial \theta_\ell} \right].$$

By an application of the Lyapunov form of the Central Limit Theorem for triangular arrays for each $i = 1, \ldots, p$, we get that

$$n^{1/2} \mathcal{H}^*_n(\widehat{\boldsymbol{\theta}}_n) \text{ converges weakly to } \mathcal{N}(\mathbf{0}, \boldsymbol{K}(\boldsymbol{\theta}_0)) \tag{8.5}$$

where $\mathcal{H}^*_n(\boldsymbol{\theta})$ is the $1 \times r$ vector with elements $H^*_k(\boldsymbol{\theta})$.
Evaluating (8.4) at $\hat{\boldsymbol{\theta}}^*_n$ gives

$$\mathcal{H}^*_n(\widehat{\boldsymbol{\theta}}_n) = \left(\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n) + \mathcal{G}^*(\widehat{\boldsymbol{\theta}}^*_n, \widehat{\boldsymbol{\theta}}_n)\right)(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n) \tag{8.6}$$

where $\boldsymbol{J}^*(\boldsymbol{\theta})$ is the $r \times r$ matrix with elements $-H^*_{k\ell}(\boldsymbol{\theta})$ and $\mathcal{G}^*(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ the $r \times r$ matrix with elements $\sum_{m=1}^{r} (\theta_m - \theta^*_m) H^*_{klm}(\widetilde{\boldsymbol{\theta}})/2$ ($\widetilde{\boldsymbol{\theta}}$ is some interior point of the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$).
From the proof of (i) it follows that,

$$\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n) \to \boldsymbol{J}(\boldsymbol{\theta}_0) \text{ in bootstrap probability} \tag{8.7}$$

and

$$\mathcal{G}^*(\widehat{\boldsymbol{\theta}}^*_n, \widehat{\boldsymbol{\theta}}_n) \to \mathbf{0} \text{ in bootstrap probability.} \tag{8.8}$$

The matrix $J_\delta^{-1}(\boldsymbol{\theta}_0)$ exists by condition (R3) such that $n^{1/2}(\widehat{\boldsymbol{\theta}}^*_n - \widehat{\boldsymbol{\theta}}_n)$ converges weakly to $\mathcal{N}(\mathbf{0}, \Sigma_\delta(\boldsymbol{\theta}_0))$. An application of well-known properties of transformed sequences and Pólya's theorem (see, e.g., Theorems 1.7 and 1.5.3 in Serfling, 1980) completes the proof.

Theorem 8.3 can be used to construct approximate confidence regions for $\boldsymbol{\theta}_0$. Indeed, choosing e.g. $g(u_1, \ldots, u_r) = \|\boldsymbol{u}\| = (\sum_{\ell=1}^{r} u_\ell^2)^{1/2}$, Theorem 8.3 implies that the confidence region

$$\mathcal{R}^\alpha_n = \left\{\boldsymbol{\theta} \in I\!R^r \;:\; \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \leq n^{-1/2} c^*_{n,\alpha}\right\}$$

where $c^*_{n,\alpha}$ denotes the $(1-\alpha)$th quantile of the (simulated) bootstrap approximation, has a coverage probability converging to $1-\alpha$. Also "studentized" versions of Theorem 8.3 can be formulated and used to construct confidence regions (percentile t-regions). For the coverage probability of such regions, one would expect a convergence rate at least as high as that of the confidence region based on the normal approximation.

8.4 Bootstrap pseudolikelihood tests

Recall the testing problem $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta \backslash \Theta_0$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ belongs to Θ_0 if and only if $\theta_1 = \cdots = \theta_t = 0$. Critical points and P-values are calculated under the null hypothesis H_0. Hence, the bootstrap samples should reflect the data generating mechanism under H_0. Therefore, the sampling mechanism used in the previous section is modified as follows: define $\boldsymbol{Y}^*_{i1}, \ldots, \boldsymbol{Y}^*_{in_i}$ to be n_i independent random variables according to the density function $f_i(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^\circ_n)$ where, as in Theorem 8.2, $\widehat{\boldsymbol{\theta}}^\circ_n$ is the maximum pseudolikelihood estimator over Θ_0. Based on this bootstrap sample, $\widehat{\boldsymbol{\theta}}^*_n$ denotes the maximum pseudolikelihood estimator, $\widehat{\boldsymbol{\theta}}^*_{nL}$ denotes the first t components of this maximum pseudolikelihood estimator and $\widehat{\boldsymbol{\theta}}^{*\circ}_n$ represents the maximum pseudolikelihood estimator over Θ_0. Further, denote W^*_n, R^*_n and $-2\log\Lambda^*_n$ the robust Wald, robust score and the pseudolikelihood ratio test statistic based on the bootstrap sample:

$$\begin{aligned} W^*_n &= n(\widehat{\boldsymbol{\theta}}^*_{nL})^T(\boldsymbol{\Sigma}^*(\widehat{\boldsymbol{\theta}}^*_n)_{LL})^{-1}\widehat{\boldsymbol{\theta}}^*_{nL}, \\ R^*_n &= n\mathcal{H}^*_n(\widehat{\boldsymbol{\theta}}^{*\circ}_n)^T_L(\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}^{*\circ}_n)^{-1})_{LL}(\boldsymbol{\Sigma}^*(\widehat{\boldsymbol{\theta}}^{*\circ}_n)_{LL})^{-1}(\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}^{*\circ}_n)^{-1})_{LL}\mathcal{H}^*_n(\widehat{\boldsymbol{\theta}}^{*\circ}_n)_L \\ -2\log\Lambda^*_n &= 2(\log \mathrm{PL}^*_n(\widehat{\boldsymbol{\theta}}^*_n) - \log \mathrm{PL}^*_n(\widehat{\boldsymbol{\theta}}^{*\circ}_n)). \end{aligned}$$

Here $\boldsymbol{\Sigma}^*(\boldsymbol{\theta})$, $\boldsymbol{J}^*(\boldsymbol{\theta})$ and $\boldsymbol{K}^*(\boldsymbol{\theta})$ are equal to $\boldsymbol{\Sigma}_n(\boldsymbol{\theta})$, $\boldsymbol{J}_n(\boldsymbol{\theta})$ and $\boldsymbol{K}_n(\boldsymbol{\theta})$ respectively, but with the original sample $\{\boldsymbol{Y}_{ij}\}$ replaced by the bootstrap sample $\{\boldsymbol{Y}^*_{ij}\}$.

The next theorem states that the bootstrap procedure is consistent in estimating the null distribution of W_n, R_n and $-2\log\Lambda_n$. Hereafter P_{H_0} will denote the probability under the null hypothesis H_0.

Theorem 8.4 *Assume conditions (R1)-(R6). Then, for almost all sample paths* $(\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2}, \ldots), i = 1, \ldots, p$, *we have, as* $n \to \infty$ *and* $n_i/n \to \lambda_i$ *for all* $i = 1, \ldots, p$,

i) $\sup_{t \in \mathbb{R}} |P^*\{W^*_n \le t\} - P_{H_0}\{W_n \le t\}| = o(1)$,

ii) $\sup_{t\in\mathbb{R}} |P^*\{R_n^* \le t\} - P_{H_0}\{R_n \le t\}| = o(1)$,
iii) $\sup_{t\in\mathbb{R}} |P^*\{-2\log\Lambda_n^* \le t\} - P_{H_0}\{-2\log\Lambda_n \le t\}| = o(1)$.

Proof. From the proof of Theorem 8.3 we have that, under the null hypothesis H_0, $n^{1/2}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ)$ converges weakly to $\mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$, where now the first t elements of the true parameter $\boldsymbol{\theta}_0$ equal 0. A one-term expansion shows that

$$[\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^*)]_{k\ell} = [\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^\circ)]_{k\ell} + R_{n1}^* ||\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ||$$

and

$$[\boldsymbol{K}^*(\widehat{\boldsymbol{\theta}}_n^*)]_{k\ell} = [\boldsymbol{K}^*(\widehat{\boldsymbol{\theta}}_n^\circ)]_{k\ell} + R_{n2}^* ||\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ||.$$

Using (R4) it is easy to see that

$$\begin{aligned} E^*[|R_{n1}^*|] &\le r\sum_{i=1}^{p} \frac{n_i}{n} \sum_{\mathbf{a}\in A_i^\circ} |\gamma_{\mathbf{a}}| E^*[K_1(\boldsymbol{Y}_{i1}^{*(\mathbf{a})})] \,, \\ E^*[|R_{n2}^*|] &\le 2r\sum_{i=1}^{p} \frac{n_i}{n} \sum_{\mathbf{a}\in A_i^\circ} \sum_{\mathbf{a}'\in A_i^\circ} |\gamma_{\mathbf{a}}|\,|\gamma_{\mathbf{a}'}| E^*[K_2(\boldsymbol{Y}_{i1}^{*(\mathbf{a})}, \boldsymbol{Y}_{i1}^{*(\mathbf{a}')})] \,. \end{aligned}$$

Similar arguments as in the proof of Theorem 8.3 lead to the asymptotic χ_t^2 distribution of W_n^*.

The proof of (ii) and (iii) is based on the same arguments as the proof of Theorem 3 in Rotnitzky and Jewell (1990). Assume H_0 holds. Then, both $\widehat{\boldsymbol{\theta}}_n^*$ and $\widehat{\boldsymbol{\theta}}_n^{*\circ}$ are consistent estimators which have asymptotically a normal distribution (by Theorem 8.3). Using the notation introduced in the proof of Theorem 8.3, an expansion of $\log \mathrm{PL}_n^*(.)$ leads to,

$$\begin{aligned} &\log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^*) - \log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^\circ) = \\ &\quad n(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ)^T \mathcal{H}_n^*(\widehat{\boldsymbol{\theta}}_n^\circ) - \frac{n}{2}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ)^T J^*(\widehat{\boldsymbol{\theta}}_n^\circ)(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ) + o^*(1). \end{aligned} \tag{8.9}$$

Here and in the remainder of the proof, $o^*(a_n)$ denotes a sequence of random variables of appropriate dimension tending to zero faster than a_n in bootstrap probability.

From (8.6) and (8.8) it follows that

$$\mathcal{H}_n^*(\widehat{\boldsymbol{\theta}}_n^\circ) = \boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^\circ)(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ) + o^*(n^{-1/2}). \tag{8.10}$$

Inserting (8.10) in (8.9) and using (8.7), we get

$$2\left(\log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^*) - \log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^\circ)\right) = n(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ)^T \boldsymbol{J}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^\circ) + o^*(1). \tag{8.11}$$

Identity (8.6) and results (8.7) and (8.8), applied on $\widehat{\boldsymbol{\theta}}_n^{*\circ}$ and $\widehat{\boldsymbol{\theta}}_n^{\circ}$, lead to

$$\mathcal{H}_{nR}^*(\widehat{\boldsymbol{\theta}}_n^{\circ}) = \boldsymbol{J}_{RR}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{nR}^{*\circ} - \widehat{\boldsymbol{\theta}}_{nR}^{\circ}) + o^*(n^{-1/2}) \tag{8.12}$$

and (8.10) reformulated for the last $r-t$ elements becomes

$$\mathcal{H}_{nR}^*(\widehat{\boldsymbol{\theta}}_n^{\circ}) = \boldsymbol{J}_{RL}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{nL}^{*} - \widehat{\boldsymbol{\theta}}_{nL}^{\circ}) + \boldsymbol{J}_{RR}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{nR}^{*} - \widehat{\boldsymbol{\theta}}_{nR}^{\circ}) + o^*(n^{-1/2}). \tag{8.13}$$

The analogue of (8.11) for the estimator $\widehat{\boldsymbol{\theta}}_n^{*\circ}$ is

$$2\left(\log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^{*\circ}) - \log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^{\circ})\right) = n(\widehat{\boldsymbol{\theta}}_{nR}^{*\circ} - \widehat{\boldsymbol{\theta}}_{nR}^{\circ})^T \boldsymbol{J}_{RR}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{nR}^{*\circ} - \widehat{\boldsymbol{\theta}}_{nR}^{\circ}) + o^*(1). \tag{8.14}$$

Subtracting (8.14) from (8.11), isolating $(\widehat{\boldsymbol{\theta}}_{nR}^{*\circ} - \widehat{\boldsymbol{\theta}}_{nR}^{\circ})$ from expressions (8.12) and (8.13) and inserting this, leads to

$$2\left(\log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^{*}) - \log \mathrm{PL}_n^*(\widehat{\boldsymbol{\theta}}_n^{*\circ})\right) = n(\widehat{\boldsymbol{\theta}}_{nL}^{*} - \widehat{\boldsymbol{\theta}}_{nL}^{\circ})^T \boldsymbol{J}_{LL}^{-1}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_{nL}^{*} - \widehat{\boldsymbol{\theta}}_{nL}^{\circ}) + o^*(1).$$

Theorem 8.3 and classical properties of quadratic forms yield the result (iii).

From (8.10) we obtain that

$$\widehat{\boldsymbol{\theta}}_{nL}^{*} - \widehat{\boldsymbol{\theta}}_{nL}^{\circ} = [\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^{\circ})^{-1}]_{LL}\mathcal{H}_{nL}^*(\widehat{\boldsymbol{\theta}}_n^{\circ}) + [\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^{\circ})^{-1}]_{LR}\mathcal{H}_{nR}^*(\widehat{\boldsymbol{\theta}}_n^{\circ}) + o^*(n^{-1/2}). \tag{8.15}$$

Theorem 8.3 implies that $\sqrt{n}\{\boldsymbol{\Sigma}^*(\widehat{\boldsymbol{\theta}}_n^{\circ})\}_{LL}^{-1/2}(\widehat{\boldsymbol{\theta}}_{nL}^{*} - \widehat{\boldsymbol{\theta}}_{nL}^{\circ})$ converges in distribution to a t-variate standard normal random variable. The result for the robust score statistic now follows by replacing $\widehat{\boldsymbol{\theta}}_n^{\circ}$ in $\mathcal{H}_n^*(\widehat{\boldsymbol{\theta}}_n^{\circ})$, $\boldsymbol{J}^*(\widehat{\boldsymbol{\theta}}_n^{\circ})$ and $\boldsymbol{\Sigma}^*(\widehat{\boldsymbol{\theta}}_n^{\circ})$ by its $\sqrt{n}$ consistent bootstrap estimator $\widehat{\boldsymbol{\theta}}_n^{*\circ}$.

In Section 8.5 a limited simulation study illustrates the finite sample behavior of the bootstrap tests for clustered binary data.

8.5 Simulations and data examples

The simulation study illustrates the estimation problem of the unknown eigenvalues and shows that the χ^2 type PLR test, based on these estimators, has severe problems in achieving the nominal level. The bootstrap PLR test, using bootstrap critical points, seems to nicely correct this towards the nominal size. Similar problems can be remedied by the bootstrap for the Wald and, less pronounced, for the score test. Data examples are given in Section 8.5.2 and 8.5.3.

8.5.1 Simulation study

The simulations will be constructed in a typical setting of developmental toxicity experiments, which are designed to assess the potential adverse effects of drugs or other exposures on developing foetuses of pregnant rodents (dams). A typical study includes a control group and some dosed groups. The exposure usually occurs early in gestation, the dams are sacrificed prior to term and next, the foetuses are examined for malformations. Denote $m_\alpha, \alpha = 1, \ldots, a$ the possible litter sizes and $d_\beta, \beta = 1, \ldots, b$ the possible dose levels. This leads to $p = a \times b$ different "associated" populations. For simplicity, we use a single index $i = 1, \ldots, p$ to enumerate these different populations. A litter (cluster) j from population i has a specific size m_i and was given a certain dose d_i. The number of population i litters is n_i and the total number of litters is n. Next to the malformation probability, clustered binary data models all include one or more parameters to describe the association between the outcomes $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijm_i})$ of litter j in population i, where Y_{ijk} indicates whether the kth fetus in litter j is abnormal.

The conditional model of Molenberghs and Ryan (1999) (abbreviated as MR) assumes $Y_{ijk} = 1$ when the kth fetus in litter j is abnormal and -1 otherwise. This coding provides a parameterization which more naturally leads to desirable properties when the role of success and failure is reversed and when cluster size is variable (Cox and Wermuth, 1994). Let $Y_{ij\bullet}$ denote the total number of malformed foetuses in litter j. The MR probability function of $\boldsymbol{Y}_{ij}$ is, by (3.2),

$$f(\boldsymbol{y}_{ij}; \beta_{1i}, \beta_{2i}) = \exp\left\{\beta_{1i} y_{ij\bullet} - \beta_{2i} y_{ij\bullet}(m_i - y_{ij\bullet}) - A\right\},$$

with A the normalizing constant and $(\beta_{1i}, \beta_{2i})^T = \boldsymbol{X}_i \boldsymbol{\theta}$ where $\boldsymbol{X}_i$ is a design matrix based on dose d_i associated with cluster j and $\boldsymbol{\theta}$ the coefficient vector. The parameter β_{1i} can be interpreted as a main effect and β_{2i} as a parameter measuring the intra-litter association.

Instead of the joint probability $f(y_{ij}; \beta_{1i}, \beta_{2i})$, Geys, Molenberghs and Ryan (1997, 1999) (abbreviated as GMR) consider the following product of the m_i conditional probabilities:

$$\prod_{k=1}^{m_i} f(y_{ijk} | y_{ij\ell}, \ell \neq k; \beta_{1i}, \beta_{2i}) = \prod_{k=1}^{m_i} p_{ijs}^{y_{ijk}} p_{ijf}^{1-y_{ijk}}$$

where p_{ijs} is the conditional probability of an additional success, i.e.

$$\text{logit}(p_{ijs}) \quad = \quad \text{logit}\{P(y_{ijk} = 1 | y_{ij\bullet} - 1 \text{ successes } \& \; m_i - y_{ij\bullet} \text{ failures})\}$$

$$= \quad (1\ ,\ (m_i - 2y_{ij\bullet} + 1))\boldsymbol{X}_i\boldsymbol{\theta}$$

and p_{ijf} the conditional probability of an additional failure, i.e.

$$\begin{aligned}\text{logit}(p_{ijf}) &= \text{logit}\{P(y_{ijk} = -1 | y_{ij\bullet} \text{ successes } \& \ m_i - y_{ij\bullet} - 1 \text{ failures})\} \\ &= (-1\ ,\ (m_i - 2y_{ij\bullet} - 1))\boldsymbol{X}_i\boldsymbol{\theta}.\end{aligned}$$

The contribution of the jth cluster ($j = 1, \ldots, n_i$) to the log pseudolikelihood is then given by $y_{ij\bullet}\log(p_{ijs}) + (m_i - y_{ij\bullet})\log(p_{ijf})$.

We performed a limited simulation study in order to illustrate the finite sample behavior of the parametric bootstrap procedure. We used one control group (dose 0) and three active groups (doses 0.25, 0.5 and 1). We experimented with an equal number of $NC = 5$, 15 or 30 clusters assigned to each dose group. The number m_i of foetuses per litter is assumed to follow a local linear smoothed version of the relative frequency distribution given in Table 1 of Kupper *et al.* (1986), which is considered representative of that encountered in actual experimental situations. Realistic parameter values $\boldsymbol{\theta}$ were used. We consider two cases. In case 1 data are generated and fitted with

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & d_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta}^T = \begin{pmatrix} \theta_{10} & \theta_{11} & \theta_{20} \end{pmatrix} \tag{8.16}$$

and the hypothesis of interest is $H_0 : \theta_{11} = 0$ (no dose effect on malformation probability). In case 2 we consider

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & d_i & 0 & 0 \\ 0 & 0 & 1 & d_i \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta}^T = \begin{pmatrix} \theta_{10} & \theta_{11} & \theta_{20} & \theta_{21} \end{pmatrix},$$

and $H_0 : \theta_{11} = \theta_{21} = 0$ (constant malformation probability and constant intra-litter association).

The pseudolikelihood ratio test can be modified such that it has an approximate χ^2_t distribution. Similarly to GMR (1999), we used the modified test $-2\log\Lambda_n/\bar{\alpha}$ with $\bar{\alpha}$ the mean of the eigenvalues α_k (see also Rotnitzky and Jewell, 1990). In analogy with the Wald test where the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = (\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{K}(\boldsymbol{\theta}_0)(\boldsymbol{J}(\boldsymbol{\theta}_0))^{-1}$ can be estimated under H_0 (using $\boldsymbol{\Sigma}_n(\widehat{\boldsymbol{\theta}}^{\circ}_n)$) or under H_1 (using $\boldsymbol{\Sigma}_n(\widehat{\boldsymbol{\theta}}_n)$), the eigenvalues α_k can be estimated under H_0 or under H_1. This leads to five classical χ^2 tests: W_0, W_1, R, PLR_0 and PLR_1. There are four bootstrap analogies. All tests have similar first order asymptotic behavior under H_0.

		Bootstrap tests				χ^2 tests				
θ_{20}	NC	PLR	W_1	W_0	R	PLR_1	W_1	PLR_0	W_0	R
0.1	5	6.0	4.2	5.1	4.8	13.6*	9.6*	5.4	11.1*	3.5*
	30	4.9	5.1	4.7	4.9	6.6*	5.9	5.2	5.5	4.9
0.25	15	5.6	4.0	5.9	4.7	18.5*	14.0*	5.8	12.7*	3.0*
	30	6.5*	4.9	6.3	6.5*	11.5*	9.3*	7.5*	10.1*	6.6*

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 8.1: GMR model. Case 1 with $\theta_{10} = -2.5$. Simulated type I error probabilities (as %), significance level 0.05.

For each situation, 1000 datasets were generated and on each dataset the P-values of the different test statistics were computed based on their limiting χ^2 distribution and on the simulated bootstrap distribution, using 1000 bootstrap samples.

Tables 8.1 and 8.2 show results for case 1 using the GMR approach. A global inspection of Table 8.1 clearly shows the superiority of the bootstrap tests in attaining the 5% level. Many χ^2 tests show inflated type I error probabilities which are nicely corrected by their corresponding bootstrap alternatives. Table 8.2 shows observed and size adjusted power estimates for two alternative values of θ_{11}. Apparently, the bootstrap tests have, compared with their classical counterparts, comparable power characteristics. As a global conclusion, the bootstrap PLR and W_1 and the χ^2-score tests seem to be the best choices, whereas the χ^2-W_0 and PLR_1 tests are the least favorable.

For case 2, Table 8.3 exhibits the same patterns, but even more pronounced. Also the χ^2-PLR_0 and score tests have size problems while the other χ^2 tests take unacceptably high type I error probabilities, especially in the case where sample information is lowest: a high intra-litter association $\theta_{20} = 0.25$ and only 15 clusters per dose level. We also compared the power estimates and found similar conclusions as in case 1.

		Bootstrap tests				χ^2 tests				
θ_{11}	NC	PLR	W_1	W_0	R	PLR_1	W_1	PLR_0	W_0	R
0.5	5	8.4	7.8	5.3	7.5	15.6	13.1	7.0	10.2	7.2
		(7.9)	(8.5)	(5.2)	(4.9)	(5.8)	(7.8)	(6.8)	(4.0)	(8.2)
1.0		22.9	19.8	11.1	18.5	16.2	28.7	19.7	20.9	19.6
		(21.1)	(21.9)	(11.1)	(19.0)	(12.5)	(18.9)	(18.9)	(8.8)	(22.4)
0.5	30	26.5	27.6	20.1	25.2	29.8	30.0	25.2	23.3	26.1
		(26.4)	(27.1)	(21.5)	(25.2)	(27.4)	(28.6)	(24.5)	(21.2)	(26.3)
1.0		88.5	88.3	82.8	85.8	89.4	89.3	86.0	85.0	86.5
		(88.6)	(87.7)	(84.4)	(85.9)	(87.6)	(88.5)	(85.6)	(83.6)	(86.7)

Table 8.2: GMR model. Case 1 with $\theta_{10} = -2.5, \theta_{20} = 0.1$. Simulated power (as %), significance level 0.05. Size adjusted values between brackets.

		Bootstrap tests				χ^2 tests				
θ_{20}	NC	PLR	W_1	W_0	R	PLR_1	W_1	PLR_0	W_0	R
0.1	5	5.3	4.8	5.2	4.9	8.1*	12.7*	2.3*	15.9*	1.3*
	15	3.1*	5.3	5.1	5.9	2.8*	11.7*	0.7*	10.1*	4.9
	30	5.2	3.5*	4.7	5.8	1.3*	9.6*	0.4*	7.5*	5.4
0.25	15	6.1	5.3	6.2	3.7	24.4*	28.6*	8.2*	17.5*	0.5*
	30	5.9	4.9	4.2	6.2	13.1*	19.5*	3.1*	11.5*	2.1*

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 8.3: GMR model. Case 2 with $\theta_{10} = -2.5$. Simulated type I error probabilities (as %), significance level 0.05.

		Bootstrap tests				χ^2 tests			
θ_{20}	NC	LR	W_1	W_0	R	LR	W_1	W_0	R
0.1	5	5.5	5.1	4.0	4.5	5.2	10.1*	10.2*	3.9
	15	5.6	5.8	5.4	5.4	6.1	6.9*	7.3*	5.3
	30	3.8	4.3	4.2	4.2	3.9	4.6	4.8	4.5
0.25	15	6.1	6.6*	7.8*	3.3*	4.2	11.0*	11.4*	4.3
	30	5.6	5.1	4.5	4.8	5.8	9.2*	8.3*	4.6

$*$ denotes the proportion of significant tests which differs significantly from 5%

Table 8.4: MR model. Case 1 with $\theta_{10} = -2.5$. Simulated type I error probabilities (as %), significance level 0.05.

Finally, for case 1, Table 8.4 shows some results for the full likelihood MR model. Recall that, since the joint distribution is correctly specified, no eigenvalues have to be estimated and hence there is only one likelihood ratio test. All test statistics except the robust χ^2 Wald tests, have very comparable simulated type I error probabilities. By using the bootstrap, the poor behavior of the Wald tests is nicely corrected. We also noticed no substantial differences in power for the corresponding bootstrap and χ^2 tests.

The main conclusion of this limited simulation study is that the bootstrap tests automatically correct for the likelihood misspecification and they seem to be superior to their classical counterparts. In fact, some improvements in size are quite spectacular. Comparing the χ^2 tests, the robust score test is clearly the preferable one.

8.5.2 The theophylline data

In this section we will apply the bootstrap tests to the toxicity and teratology study on theophylline administered to pregnant mice. As an illustration, we will consider all tests which are discussed in this chapter in both the MR and the GMR model. For each situation, a linear/constant model (case 1) was fit to the data and the null

hypothesis tested is the hypothesis of linearity:

$$H_0 : \theta_{11} = 0.$$

	Bootstrap tests					χ^2 tests			
	LR	W_1	W_0	R	(B)	LR	W_1	W_0	R
External	4.50	11.40	15.10	17.80	(1000)	4.43	9.76	12.02	16.98
Visceral	6.86	24.53	8.03	21.90	(685)	9.93	3.55	4.97	18.26
Skeletal	11.31	13.48	19.13	68.83	(831)	39.36	2.19	24.55	18.82
Collapsed	1.20	4.10	4.00	6.80	(1000)	1.30	3.40	3.58	7.54

Table 8.5: Theophylline data. Tests of $H_0 : \theta_{11} = 0$ in a linear/constant MR model. P-values are shown as %, (B) denotes the number of bootstrap replicates.

The results for the MR model are shown in Table 8.5. Recall that all Wald and score tests are robustified. The results of the likelihood ratio test are, for the χ^2 tests, based on the χ^2_1 distribution. This aspect of robustification might explain the difference between the likelihood ratio and the other tests for external malformations. There probably will not be a significant effect of theophylline on this type of malformation. For visceral and skeletal malformations, the difference between the P-values of the two robust Wald statistics (estimated covariance matrix under H_1) is remarkable. This is consistent with results from the simulation study (although the setting is different of course), where much too large simulated type I error probabilities were obtained. Only tests for the collapsed outcome give significant P-values. Based on these results, there might be an effect of theophylline on fetal development in mice.

The results of the tests in the pseudolikelihood model (GMR) are shown in Table 8.6. For external malformation, only the bootstrap PLR test gives a borderline significant result. All other tests do not indicate any evidence against the null hypothesis. The picture is less clear for visceral and skeletal malformations. There is not only a very large difference between bootstrap and χ^2 results for W_1 (as also noted for the MR model), but also for PLR_1. For the collapsed outcomes, the conclusion about the significance of the P-values is consistent with the conclusion of the MR model.

	Bootstrap tests				χ^2 tests				
	PLR	W_1	W_0	R	PLR_1	W_1	PLR_0	W_0	R
Ext.	5.00	14.60	14.20	17.80	9.83	12.26	15.60	12.14	17.06
Visc.	5.78	24.92	8.36	22.19	0.41	3.47	15.19	5.42	18.56
Skel.	12.31	8.55	0.57	56.10	0.02	0.64	10.45	1.46	16.08
Coll.	1.10	4.60	4.20	5.90	2.58	4.11	5.70	3.34	6.83

Table 8.6: Theophylline data. Tests of $H_0 : \theta_{11} = 0$ in a linear/constant PL model. P-values are shown as %. The number of bootstrap replicates was 1000, 658, 877 and 1000, respectively.

8.5.3 Study of herbicides on mice

As a second illustration we consider the data from the study of the influence of herbicides on mice. In the beta-binomial model with a linear dose effect on Fisher's z transform of the correlation between littermates, we will test the following null hypothesis:

$$H_0 : \text{logit}(\pi(d)) = \theta_{10} + \theta_{11} d$$

versus the alternative hypothesis

$$H_0 : \text{logit}(\pi(d)) = \theta_{10} + \theta_{11} d + \theta_{12} d^2.$$

The P- values of the score test were, respectively, 0.0106 and 0.0150 for the χ^2 and the bootstrap approximation, while for the likelihood ratio test we obtained 0.013 and 0.017. Based on these values, we can reject the null hypothesis of linearity in favor of the alternative hypothesis H_1.

Although the omnibus tests from Chapter 6 are not really developed for this kind of data, it does make sense to apply them with a suitable choice of the series truncation point r_n. With $r_n = 5$, we find for the test based on the robust score criterion an observed value of 6.60. The corresponding P-value, based on asymptotic distribution theory, is 0.0108. For the likelihood based test we observe a test value of 6.14 and a P-value of 0.0144. For this particular data set, the values are remarkably close to each other.

8.6 Discussion

Simulations were restricted here to the case of clustered univariate binary data. It would be interesting to see to which extent the χ^2 type PLR test fails in reaching the prescribed significance level and how the parametric bootstrap succeeds in correcting this for the multivariate case of several malformation indicators.

In the next chapter we define a semiparametric bootstrap method reflecting a specific null hypothesis. One might expect such a method to be more robust to distributional assumptions.

Chapter 9

A One-Step Semiparametric Bootstrap Procedure

9.1 Introduction

In this chapter we focus attention on the likelihood and pseudolikelihood approach, where we investigate the performance of the robust Wald and score test statistic based on the chi-squared approximation and based on a bootstrap estimator for their distribution.

An important difference with the previous chapter is that here we do not assume the "true" likelihood model of the data to be known. If the probability model is misspecified, it is well-known that classical test statistics (such as the Wald, score or likelihood ratio statistic) do *not* have an asymptotically chi-squared distribution anymore (see, e.g., White, 1982). By means of asymptotic calculations, small sample simulations and analyses of NTP data, Molenberghs, Declerck and Aerts (1998) investigated the behavior of the parameter estimates and the classical Wald and likelihood ratio test under model misspecification, while still using the incorrect asymptotic chi-squared distribution. It is desirable, however, to use modified test statistics, for *robustifying* the standard inference methods. For full likelihood models, robust Wald and score tests have been described by, e.g., Kent (1982), Viraswami and Reid (1996). The modified tests again have an asymptotic chi-squared distribution, even when the assumed model is not correct. We are not aware of

a modified likelihood ratio test with asymptotic chi-squared distribution. Robust test statistics are also used in the context of generalized estimating equations (only specifying the mean and variance structure, see Liang and Zeger, 1986, Rotnitzky and Jewell, 1990) and in the pseudolikelihood approach, see Geys, Molenberghs and Ryan (1999).

If a parametric bootstrap is applied to a wrong probability model, bootstrap data will be generated from this erroneous distribution. As a consequence, the parametric bootstrap might possibly lead to incorrect results. Another reason to search for alternatives to a parametric bootstrap procedure arises from generalized estimating equations models (GEE). There, no assumptions are made about the specific form of the likelihood of the data, and hence, there is not even a likelihood model available to generate data from.

As a possible approach, we propose a semiparametric bootstrap method. It remains valid when the assumed model is incorrect and no bootstrap data are generated such that no iterative fitting is required. Instead bootstrap parameter estimates reflecting the null hypothesis are generated directly. Our approach is closely related to the one-step bootstrap, an approximate method to simplify the bootstrap for estimators which have to be computed iteratively (see Schucany and Wang, 1991 and Section 5.4.7 in Shao and Tu, 1995).

The method is described in Section 9.3.2, including a second order improvement. Simulations indicate that in cases where the χ^2 type tests fail in reaching the prescribed significance level, the proposed bootstrap test succeeds in correcting this towards the nominal level. The results of a simulation study and a data example are shown in Section 9.4. In particular, the developmental toxicity data on theophylline are analyzed on possible dose effect. It should be stressed that although we restrict attention to (pseudo)likelihood methods for clustered binary data, the technique can be applied to generalized estimating equations. For the GEE method however, no probability distribution has to be specified and hence, by nature of the method, it is expected to behave more robust against misspecification. For illustration we also included results of the GEE method in the analysis of the THEO data.

Aerts and Claeskens (1998a) contains most results given in this chapter.

9.2 Pseudolikelihood and misspecification

In this section we describe the models we will be working with. Most of the notation has already been defined in previous chapters, although not in this context.

Let $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{in_i}$ be independent identically distributed random variables of length m with common (unknown) joint density or discrete probability function (pdf) $g_i(\boldsymbol{y})$, $\boldsymbol{y} = (y_1, \ldots, y_m) \in R_i$, $i = 1, \ldots, p$. The number p of possibly different (associated) populations is considered as fixed whereas the number n_i of observations from the distinct populations become large as $n = \sum_{i=1}^{p} n_i$ tends to infinity, according to $n_i/n \to \lambda_i$ where $\sum_{i=1}^{p} \lambda_i = 1$ with $\lambda_i > 0$.

In the context of clustered binary data from toxicological experiments, the different populations correspond to different dose levels d_i, n_i is the number of litters exposed to dose d_i, m is the litter size and y_{ijk} indicates whether the kth fetus of litter j in population i is malformed or not. For ease of notation, we restrict to a fixed littersize m. In real situations, the size m varies among the litters.

In general, parametric inference for associated populations is based on r dimensional vector functions $\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{t})$, the *score* functions, where the "true" parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ is defined as the solution $\boldsymbol{t}$ to

$$\sum_{i=1}^{p} \lambda_i E[\boldsymbol{\psi}_i(\boldsymbol{Y}_{i1}; \boldsymbol{t})] = \boldsymbol{0}, \tag{9.1}$$

where all expectations are w.r.t. the true pdf $g_i(\boldsymbol{y})$.

Solving the system of equations

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} \boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}; \boldsymbol{t}) = \boldsymbol{0}, \tag{9.2}$$

leads to the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.

For the pseudolikelihood method, the score functions $\boldsymbol{\psi}_i$ are the partial derivatives

$$\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{t}) = \sum_{\mathbf{a} \in A_i} \gamma_{\mathbf{a}} \frac{\partial}{\partial \boldsymbol{t}} \log f_i^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}, \boldsymbol{t}).$$

The log of the pseudolikelihood is defined as in (8.1).

There can be different sources of "misspecification": an incorrect pdf and/or pseudolikelihood instead of full likelihood. Within the classical maximum likelihood approach where $\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{t}) = (\partial/\partial \boldsymbol{t}) \log f_i(\boldsymbol{y}, \boldsymbol{t})$, the assumed pdf $f_i(\boldsymbol{y}, \boldsymbol{t})$ might

not contain the true structure $g_i(\boldsymbol{y})$. In this case, assuming interchangeability of integration and derivation, the parameter $\boldsymbol{\theta}$ defined by (9.1) is that value in the parameter space $\Theta \subset I\!R^r$ which brings $f_i(\boldsymbol{y}, \boldsymbol{t})$ as close as possible to $g_i(\boldsymbol{y})$; that is, $\boldsymbol{\theta}$ minimizes the Kullback-Leibler information criterion:

$$\sum_{i=1}^{p} \lambda_i E\left[\log\left(\frac{g_i(\boldsymbol{Y}_{i1})}{f_i(\boldsymbol{Y}_{i1}, \boldsymbol{t})}\right)\right].$$

Under severe model misspecification, this might lead to parameters $\boldsymbol{\theta}$ without direct biologically meaningful interpretation. Therefore, great care should be taken in the formulation of the null hypothesis to be tested.

9.3 Testing hypotheses

Consider the hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta \backslash \Theta_0$ where Θ_0 is a $(r-t)$ dimensional subspace of the parameter space Θ such that the parameter of interest $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ belongs to Θ_0 if and only if $\theta_1 = \ldots = \theta_t = 0$, $1 \leq t \leq r$. More general situations, in which H_0 is of the form $H_0 : h_1(\boldsymbol{\theta}) = \ldots = h_t(\boldsymbol{\theta}) = 0$ for some smooth real-valued functions $h_1, \ldots, h_t$, can be put into this form by a reparametrization. Note that the hypothesis is stated in terms of the parameter which minimizes the Kullback-Leibler information criterion. The proposed method is presented in full generality but, in general, when misspecification might occur, one always has to decide whether the stated hypothesis is meaningful.

9.3.1 Robustified test statistics

If model misspecification arises, classical Wald, score and likelihood ratio tests do not have an asymptotic chi-squared distribution anymore. This is due to the fact that the Bartlett identities are no longer valid, e.g.,

$$\boldsymbol{K}_i(\boldsymbol{\theta}) \equiv E\left[\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{\theta})^T\right] \neq -E\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}_i(\boldsymbol{y}, \boldsymbol{\theta})\right] \equiv \boldsymbol{J}_i(\boldsymbol{\theta}). \tag{9.3}$$

Wald and score tests can be robustified by using the so-called sandwich variance estimator. Properties of these robust test statistics are studied by Huber (1967), White (1982), Rotnitzky and Jewell (1990), Boos (1992) and Viraswami and Reid (1996) among others. There is no robustified version of the likelihood ratio test,

whose asymptotic distribution is, under misspecification, a weighted sum of independent chi-squared random variables with one degree of freedom, where the weights are unknown and have to be estimated from the data. Properties of this test are studied by, e.g., Foutz and Srivastava (1977) and Kent (1982). We will not consider the likelihood ratio test.

The following matrices $\boldsymbol{J}_n(\boldsymbol{\theta})$ and $\boldsymbol{K}_n(\boldsymbol{\theta})$ will be used in the construction of the covariance matrix of the estimator $\widehat{\boldsymbol{\theta}}_n$,

$$\begin{aligned}\boldsymbol{J}_n(\boldsymbol{\theta}) &= -\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij},\boldsymbol{\theta}),\\ \boldsymbol{K}_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij},\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij},\boldsymbol{\theta})^T.\end{aligned}$$

The robust Wald statistics (see Section 8.2 for the definition of a matrix $\boldsymbol{V}_{LL}$) is defined as

$$W_n = n(\widehat{\boldsymbol{\theta}}_{nL})^T\left(\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\boldsymbol{K}_n(\widehat{\boldsymbol{\theta}}_n)\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}\right)^{-1}_{LL}\widehat{\boldsymbol{\theta}}_{nL}$$

and the score statistic $S_n =$

$$\frac{1}{n}\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i}\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij};\widehat{\boldsymbol{\theta}}_n^{(0)})_L\right)^T\left(\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}\right)_{LL}\left(\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}\boldsymbol{K}_n(\widehat{\boldsymbol{\theta}}_n^{(0)})\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}\right)^{-1}_{LL}$$

$$\times\left(\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}_n^{(0)})^{-1}\right)_{LL}\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i}\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij};\widehat{\boldsymbol{\theta}}_n^{(0)})_L\right).$$

Both test statistics have an asymptotic χ^2_t distribution (White, 1982).

The distributional properties of the robust Wald and score test rely on asymptotic approximations. Our goal is to examine how well these approximations work in the context of misspecified clustered binary data models and to study the performance of a semiparametric bootstrap test.

9.3.2 Bootstrap test statistics

The main difficulty when constructing bootstrap tests is the generation of bootstrap data reflecting the null hypothesis. Assuming that $g_i(\boldsymbol{y}) = f_i(\boldsymbol{y},\boldsymbol{\theta})$, bootstrap data $\{\boldsymbol{Y}^*_{ij}\}$ can be generated from the fitted model $f_i(\boldsymbol{y},\widehat{\boldsymbol{\theta}}_n^{(0)})$ where $\widehat{\boldsymbol{\theta}}_n^{(0)}$ is a $\sqrt{n}$- consistent estimator of $\boldsymbol{\theta}$ under the null hypothesis. This parametric bootstrap method allows to nicely reflect the null hypothesis and to reconstruct the original design.

That approach is the subject of Chapter 8, where the model is correctly specified, that is $g_i(\boldsymbol{y}) = f_i(\boldsymbol{y}, \boldsymbol{\theta})$, but the pseudolikelihood is being maximized instead of the true likelihood. A major drawback, however, is that it fully relies on the possibly misspecified pdf $f_i(\boldsymbol{y}, \boldsymbol{\theta})$. If the assumed model is not correct, the superior performance of the parametric bootstrap test might collapse (see, e.g., Lee, 1994 who studied a procedure for choosing between the parametric and nonparametric bootstrap).

By resampling the data (nonparametric bootstrap) or residuals (semiparametric bootstrap), no likelihood model has to be specified and therefore this bootstrap approach seems to be preferable in the context of misspecification. A main difficulty however is to construct a resampling scheme that reflects the null hypothesis. Resampling data blindly leads to tests with very low power (see e.g. Hall and Wilson, 1991). As a consequence, most bootstrap results focus on the construction of confidence sets (see e.g. Chapter 4 in Shao and Tu, 1995).

A nice application of a semiparametric bootstrap test is given in Mammen (1993). For linear models of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and the hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{h}$, he proved the asymptotic correctness of a bootstrap test based on the residual bootstrap. That is, new data are generated as $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_0 + \boldsymbol{\varepsilon}^*$ with $\hat{\boldsymbol{\beta}}_0$ the restricted least squares estimator under H_0 and $\boldsymbol{\varepsilon}^*$ a resample taken with replacement from the residuals (corresponding to the fitted unrestricted model). The residual bootstrap has also been successfully used for generalized linear models, see, e.g., Moulton and Zeger (1989, 1991). For multiparameter (pseudo-)likelihood models, it is not clear how to define residuals that can be used for resampling purposes. There is a need for a method avoiding the introduction of residuals.

Therefore we propose to resample the score and the differentiated score values. Based on a linear approximation, we define a bootstrap replicate of $\widehat{\boldsymbol{\theta}}_n$ under H_0 as

$$\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n^{(0)} - \left(\sum_{i=1}^{p} \sum_{j=1}^{n_i} \dot{\boldsymbol{\psi}}_{ij}^*(\widehat{\boldsymbol{\theta}}_n) \right)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \boldsymbol{\psi}_{ij}^*(\widehat{\boldsymbol{\theta}}_n) \tag{9.4}$$

where, for each $i = 1, \ldots, p$, $(\boldsymbol{\psi}_{ij}^*(\widehat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_{ij}^*(\widehat{\boldsymbol{\theta}}_n)), j = 1, \ldots, n_i$ is a sample with replacement from the set

$$\left\{ \left(\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n), (\partial/\partial\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n) \right), \quad j = 1, \ldots, n_i \right\}.$$

Note that $\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n)$ is a $r \times 1$ vector and $(\partial/\partial\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n)$ is a $r \times r$ matrix. A similar linearization idea is used in simulation approaches for the bootstrap, as

the linear bootstrap (Davison, Hinkley and Schechtman, 1986) and the one-step bootstrap (Schucany and Wang, 1991). For linear models $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the idea of resampling scores has also been proposed by Hu and Zidek (1995).

The rationale behind definition (9.4) is as follows. The first term at the right-hand side of (9.4) reflects the null hypothesis and the second term represents the random fluctuation of the bootstrap replicate $\widehat{\boldsymbol{\theta}}_n^*$ around the estimator $\widehat{\boldsymbol{\theta}}_n^{(0)}$. The score values are evaluated in the unrestricted estimator $\widehat{\boldsymbol{\theta}}_n$, because this term should catch the random mechanism properly, even if the null hypothesis is not true.

The bootstrap Wald and score test statistics based on $\widehat{\boldsymbol{\theta}}_n^*$ as defined in (9.4), coincide and are given by

$$W_n^* = S_n^* = n(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^{(0)})_L^T \left(\boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)^{-1} \boldsymbol{K}_n^*(\widehat{\boldsymbol{\theta}}_n) \boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)^{-1} \right)_{LL}^{-1} (\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}_n^{(0)})_L,$$

where $\boldsymbol{J}_n^*(\boldsymbol{\theta})$ and $\boldsymbol{K}_n^*(\boldsymbol{\theta})$ are defined the same way as $\boldsymbol{J}_n(\boldsymbol{\theta})$ and $\boldsymbol{K}_n(\boldsymbol{\theta})$, but using the bootstrap scores and derivatives of the scores $(\boldsymbol{\psi}_{ij}^*(\widehat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_{ij}^*(\widehat{\boldsymbol{\theta}}_n))$, $j = 1, \ldots, n_i$ instead of $(\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n)$, $(\partial/\partial\boldsymbol{\theta})\boldsymbol{\psi}_i(\boldsymbol{Y}_{ij}, \widehat{\boldsymbol{\theta}}_n))$, $j = 1, \ldots, n_i$.

The motivation for defining S_n^* equal to W_n^* follows from the classical arguments in proving the asymptotic normality of the score test statistic. It is well known that both test statistics are first order equivalent. A typical way of obtaining the asymptotic distribution of the score test statistic is by substituting $(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n^{(0)})_L$ in the Wald statistic by the first L components of the second term in (9.4). By definition (9.4) of the one-step linear estimator, this substitution is exact; for estimators in general, this is only approximate.

Definition (9.4) follows from a linear approximation of the score equations. One might improve on this by including quadratic and higher order terms. A possible approach is suggested by the second and third order efficient approximations as discussed in, e.g., Ghosh (1994). We focus attention on the following second order approximation (simplified to one population),

$$\begin{aligned} \mathbf{0} &= \sum_{j=1}^{n} \boldsymbol{\psi}(\boldsymbol{Y}_j, \boldsymbol{\theta}) + \sum_{k=1}^{r} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_k} \boldsymbol{\psi}(\boldsymbol{Y}_j, \boldsymbol{\theta})(\widehat{\theta}_{nk} - \theta_k) \\ &\quad + \sum_{k=1}^{r} \sum_{\ell=1}^{r} \sum_{j=1}^{n} \frac{\partial^2}{\partial \theta_k \partial \theta_\ell} \boldsymbol{\psi}(\boldsymbol{Y}_j, \boldsymbol{\theta})(\widehat{\theta}_{nk} - \theta_k)(\widehat{\theta}_{n\ell} - \theta_\ell) + O_P(n^{-1/2}). \quad (9.5) \end{aligned}$$

By calculations similar to those of Ghosh (1994), the expansion (9.5) suggests the

following one-step quadratic estimator

$$\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n^{(0)} + \boldsymbol{U}_n^* - \frac{1}{2}\left(\sum_{j=1}^{n} \dot{\boldsymbol{\psi}}_j^*(\widehat{\boldsymbol{\theta}}_n)\right)^{-1} \sum_{k=1}^{r}\sum_{\ell=1}^{r}\sum_{j=1}^{n} \ddot{\boldsymbol{\psi}}_j^*(\widehat{\boldsymbol{\theta}}_n)_{k,\ell} U_{nk}^* U_{n\ell}^* \tag{9.6}$$

with

$$\boldsymbol{U}_n^* = -\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} \dot{\boldsymbol{\psi}}_{ij}^*(\widehat{\boldsymbol{\theta}}_n)\right)^{-1} \sum_{i=1}^{p}\sum_{j=1}^{n_i} \boldsymbol{\psi}_{ij}^*(\widehat{\boldsymbol{\theta}}_n).$$

This bootstrap estimator is based on the values $(\boldsymbol{\psi}_j^*(\widehat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_j^*(\widehat{\boldsymbol{\theta}}_n), \ddot{\boldsymbol{\psi}}_j^*(\widehat{\boldsymbol{\theta}}_n))$, $j = 1, \ldots, n$ taken with replacement from the set

$$\left\{ \left(\boldsymbol{\psi}(\boldsymbol{Y}_j, \widehat{\boldsymbol{\theta}}_n), (\partial/\partial\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{Y}_j, \widehat{\boldsymbol{\theta}}_n), (\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)\boldsymbol{\psi}(\boldsymbol{Y}_j, \widehat{\boldsymbol{\theta}}_n) \right), \quad j = 1, \ldots, n \right\}.$$

It is expected that the last term at the right-hand side of (9.6) improves the representation of the random variation about the null estimate $\widehat{\boldsymbol{\theta}}_n^{(0)}$. This is confirmed in the simulation study.

Both bootstrap procedures (linear and quadratic) lead to consistent estimators for the null distribution of the robust Wald and score test statistics. We first list the assumed regularity conditions.

(A1) The parameter space Θ is an open subset of $\mathbb{R}^r$. First order partial derivatives of $\boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{t})$ w.r.t. $\boldsymbol{t}$ exist, are continuous in $\boldsymbol{y}$ and are integrable.

(A2) There exists a function H_1 such that $E[H_1(\boldsymbol{Y})^2] < \infty$ and for each $k, \ell = 1, \ldots, r$, $(\partial/\partial t_k)\boldsymbol{\psi}_\ell(\boldsymbol{y}; \boldsymbol{t})$ is bounded in absolute value by $H_1(\boldsymbol{y})$ uniformly in some neighborhood of $\boldsymbol{\theta}$.

(A3) The matrices $\boldsymbol{J}(\cdot)$ and $\boldsymbol{K}(\cdot)$ as defined in (9.3) exist, $\boldsymbol{K}(\cdot)$ is positive definite in $\boldsymbol{\theta}$.

(A4) There exists a $\delta > 0$ and a function H_2 such that $E[H_2(\boldsymbol{Y})] < \infty$, and for each k,
$|\boldsymbol{\psi}_k(\boldsymbol{y}; \boldsymbol{t})|^{2+\delta}$ is bounded by $H_2(\boldsymbol{y})$, uniformly in some neighborhood of $\boldsymbol{\theta}$.

(A5) For each k and ℓ,

$$E\left[\sup_{||h|| \le d} |\boldsymbol{\psi}_k(\boldsymbol{Y}; \boldsymbol{\theta} + \boldsymbol{h})\boldsymbol{\psi}_\ell(\boldsymbol{Y}; \boldsymbol{\theta} + \boldsymbol{h}) - \boldsymbol{\psi}_k(\boldsymbol{Y}; \boldsymbol{\theta})\boldsymbol{\psi}_\ell(\boldsymbol{Y}; \boldsymbol{\theta})|\right]$$

and

$$E\left[\sup_{||h|| \le d} \left| \frac{\partial}{\partial t_k}\boldsymbol{\psi}_\ell(\boldsymbol{Y}; \boldsymbol{\theta} + \boldsymbol{h}) - \frac{\partial}{\partial t_k}\boldsymbol{\psi}_\ell(\boldsymbol{Y}; \boldsymbol{\theta}) \right| \right]$$

both tend to zero as $d \to 0$.

Theorem 9.1 *If the above regularity conditions hold, then, for* $\widehat{\boldsymbol{\theta}}_n^*$ *defined by (9.4) or (9.6) and for almost all sample paths* $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots)$, *as* $n \to \infty$

(i) The bootstrap estimator $\widehat{\boldsymbol{\theta}}_n^*$ *converges in bootstrap probability to* $\boldsymbol{\theta}$.

(ii) For any continuous function $g : \mathbb{R}^r \to \mathbb{R}^{r'}$ $(r' \leq r)$, *such that the distribution function of* $g(\boldsymbol{Z})$, *with* $\boldsymbol{Z}$ *a* r'*-dimensional normal distributed random variable, is continuous, we have that*

$$\sup_{t \in \mathbb{R}^{r'}} \left| P^*\{g(n^{1/2}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)) \leq \boldsymbol{t}\} - P\{g(n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})) \leq \boldsymbol{t}\} \right| = o(1).$$

Proof. Statement (i) follows from (ii). To prove (ii), we need the strong consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$ as defined in (9.2). By extending classical maximum likelihood theory (as e.g. in Ferguson, 1996) to allow for misspecification and pseudolikelihood estimation, it can be shown that conditions (A1) and (A2) guarantee the strong consistency. Moreover, if in addition (A3) holds, we have that

$$\left(\boldsymbol{J}_n^{-1}(\boldsymbol{\theta}) \boldsymbol{K}_n(\boldsymbol{\theta}) \boldsymbol{J}_n^{-1}(\boldsymbol{\theta})\right)^{-1/2} n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}_r(\mathbf{0}, \boldsymbol{I}_\mathbf{r}). \tag{9.7}$$

The next statements concerning the bootstrap hold conditionally on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, for almost all sample paths $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and E^*, Var^* and $\mathcal{D}^*$ stand for the bootstrap expectation, variance and convergence in distribution, conditionally on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$.

Choosing $\delta > 0$ according to condition (A4), the strong consistency of $\widehat{\boldsymbol{\theta}}_n$ and (A4) imply that for any r-dimensional vector $\boldsymbol{v}$, having norm equal to 1,

$$\sum_{i=1}^{n} E^*[|n^{-1/2} \boldsymbol{v}^T \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n)|^{2+\delta}] = O_P(n^{-\delta/2}). \tag{9.8}$$

The semiparametric resampling scheme is such that $E^*[\boldsymbol{K}_n^*(\widehat{\boldsymbol{\theta}}_n)] = \boldsymbol{K}_n(\widehat{\boldsymbol{\theta}}_n)$. By an application of Theorem 2.9 of Iverson and Randles (1989) to each of the r^2 components of the matrix $\boldsymbol{K}_n(\widehat{\boldsymbol{\theta}}_n)$, conditions (A3), (A5) and the strong consistency of $\widehat{\boldsymbol{\theta}}_n$ imply that $E^*[\boldsymbol{K}_n^*(\widehat{\boldsymbol{\theta}}_n)]$ converges to $\boldsymbol{K}(\boldsymbol{\theta})$ almost surely as n tends to infinity. Together with (9.8) this implies (Liapunov's condition)

$$\frac{\sum_{i=1}^n E^*[|n^{-1/2} \boldsymbol{v}^T \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n)|^{2+\delta}]}{\left(\sum_{i=1}^n E^*[(n^{-1/2} \boldsymbol{v}^T \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n))^2]\right)^{1+\delta/2}} \to 0.$$

By the Cramér-Wold theorem it then follows that

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{\mathcal{D}^*} \mathcal{N}(\mathbf{0}, \boldsymbol{K}(\boldsymbol{\theta})).$$

Again by Theorem 2.9 of Iverson and Randles, it can be shown that $E^*[\boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)] \to \boldsymbol{J}(\boldsymbol{\theta})$ and $\text{Var}^*[\boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)] = O_P(n^{-1})$ such that $\boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{P^*} \boldsymbol{J}(\boldsymbol{\theta})$.

Combining the above results, an application of Slutsky's theorem leads to

$$n^{1/2}\boldsymbol{U}_n^* \xrightarrow{\mathcal{D}^*} \mathcal{N}(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\theta})^{-1}\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{J}(\boldsymbol{\theta})^{-1}). \tag{9.9}$$

Statement (ii) follows from the asymptotic normality results (9.7) and (9.9) and Pólya's theorem.

9.4 Simulations and the THEO data

This section illustrates the finite sample behavior of the robust chi-squared and the bootstrap Wald and score statistics. We apply the semiparametric bootstrap procedure to simulated data and to the data set on theophylline.

9.4.1 The simulation setting

Using the models described in Section 2.6.1, 3.2.2 and 8.2, we examine the performance of the different tests for the *no dose effect* null hypothesis. For this null hypothesis, the "nearest" member of an incorrectly specified parametric family still satisfies the null hypothesis. The interpretation of this statement should be clear: if there is no effect of the dose in one model, there also will be no dose effect in any other model, and this is independent of the interpretation of the model parameters.

Note that although models of quite different structure are being contrasted, the problem we are looking at really makes a lot of sense. It is exactly what happens in daily practice, since one (almost) never can be sure about the probability model which generated the set of data at hand.

In the simulations we will reconstruct some realistic situations from developmental toxicity studies. A typical toxicological experiment includes one control group and some active dose groups. For the simulations we selected dose levels 0, 0.25, 0.5 and 1. Several parameter settings were investigated, all of these might occur in practical experiments. For each of the selected models we will construct a (robustified) Wald and score statistic for testing the null hypothesis. The simulated levels and size adjusted powers of these tests are compared with those obtained by applying the bootstrap, as explained above. For the no dose effect null hypothesis we might consider two different resampling schemes. The first method is the one as

described before, where scores and differentiated scores are resampled for each dose level separately. We denote this by B_1/D for the linear one-step approximation, or by B_2/D when the quadratic approximation is being constructed. For the no dose effect null hypothesis, an alternative valid resampling scheme is to ignore the presence of the dose and to resample from the complete set of scores and differentiated scores. This resampling scheme is denoted by B_i/A $(i = 1, 2)$.

An equal number of 15 clusters was assigned to each dose group. First, the number of foetuses m is assumed to be fixed to the value $m = 12$. In a second setting the cluster sizes are random, in that case m is assumed to follow a local linear smoothed version of the relative frequency distribution given in Kupper et al. (1986) (see Table 1 in Molenberghs, Declerck and Aerts, 1998). Ideally, when several different litter sizes are observed, resampling should be done within each group of litters with identical dosage and identical litter size (being a separate population). Since litter sizes typically vary from 1 to about 20, this would lead to very small sample sizes for several of these populations and resampling would not be very effective. In our simulations with random litter sizes and also for the NTP data, we resampled scores and differentiated scores from all litters in the same dose group (having different sizes) or from the complete set (as explained before). In this way a bootstrap estimate $\widehat{\boldsymbol{\theta}}_n^*$ is not based on contributions of litters with exactly the same size distribution as in the original sample, but asymptotically it does reflect the (unknown) littersize distribution.

For each setting 500 datasets were generated using the build-in GAUSS routine RNDU. On each dataset the P-values of the robust Wald (W_n) and the robust score (R_n) statistics were computed based on their limiting χ^2 distribution and on the simulated bootstrap distribution, using 1000 bootstrap resamples.

9.4.2 Simulation results

As a first case, data were generated from the conditional MR model with intercept $\theta_{10} = -1.5, -1.0, -0.5$, dose effect $\theta_{11} = 0, 0.5$ and association parameter $\theta_{20} = 0.1, 0.2$. The beta-binomial model was used to fit the data and to test the no dose null hypothesis at the 5% level of significance. Table 9.1 shows simulated type I errors (as percentages) and Table 9.2 shows the simulated power of the tests.

In this setting we only considered the linear approximation bootstrap tests. In principle, one could define a quadratic test too, but the mathematical calculations

		$\theta_{20}=0.1$			$\theta_{20}=0.2$		
θ_{10}		χ^2	B_1/D	B_1/A	χ^2	B_1/D	B_1/A
-1.5	W_n	7.20*	4.60	3.40	7.75*	7.30*	7.51*
	R_n	6.40	3.40	2.60*	6.44	5.80	4.94
-1.0	W_n	7.21*	5.41	4.21	6.28	4.25	3.24
	R_n	8.02*	4.61	3.21	5.67	3.24	2.83*
-0.5	W_n	7.80*	4.00	3.00*	6.80	4.20	4.00
	R_n	7.80*	4.00	3.60	7.60*	3.80	3.20

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 9.1: Simulated type I errors (as %), significance level =0.05. Data are generated with the MR model and fitted using the beta-binomial model, clustersize=12. $H_0 : \theta_{11} = 0$.

		$\theta_{11}=0.5$					
		$\theta_{20}=0.1$			$\theta_{20}=0.2$		
θ_{10}		χ^2	B_1/D	B_1/A	χ^2	B_1/D	B_1/A
-1.5	W_n	39.7	36.1	42.1	16.0	13.2	12.5
	R_n	36.7	33.6	36.0	13.6	10.4	11.0
-1.0	W_n	62.3	58.2	61.1	36.3	32.5	32.4
	R_n	63.5	59.4	61.4	33.2	27.1	30.8
-0.5	W_n	87.2	85.8	87.2	95.4	94.9	95.4
	R_n	83.7	84.5	85.4	95.6	95.7	96.7

Table 9.2: Simulated power (as %), significance level =0.05. Data are generated with the MR model and fitted using the beta-binomial model, clustersize=12. $H_0 : \theta_{11} = 0$.

are getting rather cumbersome for the beta-binomial model.

First, we observe that the robust Wald and score χ^2 tests seem to behave very

similar with only slightly inflated type I errors. There is not much room for improvement by the bootstrap but the linear one-step bootstrap pulls the inflated type I errors of both robust χ^2-tests down. If the scores are resampled from the complete set of scores (ignoring the dose levels), the B_1/A bootstrap tests tend to be somewhat conservative. For the results on power characteristics to be comparable for the different tests, we show the size-adjusted rejection probabilities in Table 9.2. These results indicate that the loss in power for the bootstrap methods is almost negligible, especially for the B_1/A test. In summary, for this setting, only a small correction in the level of χ^2 tests is necessary, which is achieved by the one-step bootstrap.

In a second case, data were generated using a beta-binomial model with parameters θ_{10} = -4, -3.5, -3.0, -2.5, θ_{11} = 0, 1.0 , θ_{20} = 0.2, 0.3. Fitting and testing was based on the pseudolikelihood model (8.2). Now there are two sources of "misspecification": the assumed probability model is wrong and the pseudolikelihood technique has been used, instead of full likelihood estimation.

Type I errors and rejection probabilities were simulated for the χ^2 Wald and score tests, and the linear and quadratic bootstrap tests for both resampling schemes ($B_i/D, B_i/A$, $i = 1, 2$). As in the first setting, the type I errors are larger than the nominal level but now it is more pronounced, especially for smaller values of θ_{10} (determining the baseline malformation probability for zero dose) and for higher intra-litter association represented by θ_{20}. This latter situation corresponds to the case of the least expected number of events (malformations) and the least sample information (similar behavior of subjects within the same litter). Now, robust Wald and score χ^2 tests behave differently. In fact, the robust score χ^2 test is doing very well. The one-step linear bootstrap test is hardly any better than its χ^2 counterpart but the quadratic bootstrap Wald tests (both B_2/D and B_2/A) seem to nicely correct the level downwards. Also for the single setting in which the score test has a small problem, the bootstrap works. Note that the robust score statistic for the quadratic bootstrap test coincides, by definition of a score test, with the statistic of the linear bootstrap test. For this reason they are not shown in the tables.

As a comparison, type I errors were also considered for the extremely computer intensive, fully iterative bootstrap Wald test B_{it}/D. To obtain the simulation results of this test, we resampled for each of the 500 sets of original data, 1000 times, per dose level (with replacement, on cluster level, sample size 15). The maximum likelihood estimates $\hat{\boldsymbol{\theta}}_n^*$ of each of these 1000 resampled data sets were computed

		$\theta_{20} = 0.2$					
θ_{10}		χ^2	B_1/D	B_2/D	B_{it}/D	B_1/A	B_2/A
-4.0	W_n	10.55*	10.76*	6.96	10.76*	9.28*	6.54
	R_n	6.12	6.75	—	—	5.70	—
-3.5	W_n	5.65	6.45	4.23	5.85	5.44	3.43
	R_n	4.64	3.83	—	—	3.02	—
-3.0	W_n	6.40	5.80	5.20	7.00*	5.60	4.00
	R_n	6.80	5.40	—	—	4.60	—
-2.5	W_n	7.80*	6.60	5.80	8.20*	6.00	5.20
	R_n	7.40*	5.60	—	—	5.20	—
		$\theta_{20} = 0.3$					
θ_{10}		χ^2	B_1/D	B_2/D	B_{it}/D	B_1/A	B_2/A
-4.0	W_n	10.71*	11.65*	7.07*	11.35*	10.71*	5.78
	R_n	6.21	6.21	—	—	4.71	—
-3.5	W_n	7.06*	7.66*	5.04	7.46*	6.25	4.64
	R_n	6.86	5.04	—	—	4.44	—
-3.0	W_n	7.80*	7.20*	5.20	8.00*	6.60	5.20
	R_n	6.60	5.60	—	—	5.00	—
-2.5	W_n	6.40	6.00	5.80	6.40	5.80	5.40
	R_n	6.80	5.20	—	—	4.40	—

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 9.3: Simulated type I errors (as %), significance level =0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. Clustersize=12. $H_0 : \theta_{11} = 0$.

fully iteratively (requiring about 20 iterations) and used to get, for each of the original data sets, 1000 replicates of the Wald test statistic, now defined as

$$W_n^* = n(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)_L^T \left(\boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)^{-1} \boldsymbol{K}_n^*(\widehat{\boldsymbol{\theta}}_n) \boldsymbol{J}_n^*(\widehat{\boldsymbol{\theta}}_n)^{-1} \right)_{LL}^{-1} (\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}_n)_L.$$

This definition follows the guidelines of Hall and Wilson (1991). A well-known drawback of this bootstrap method is the fact that it nowhere reflects the null hypothesis. Related to that, one might expect a substantial loss in power. From Table 9.3 it follows that, for this application, the method has a really poor performance. Moreover, it is computationally extremely time consuming.

Table 9.4 gives some indication of loss of power for the one-step bootstrap tests. This loss is more pronounced for the linear bootstrap. The quadratic bootstrap tests seem to lift up the power close to that of the χ^2 tests. In conclusion, for the robust Wald test, we recommend the quadratic one-step bootstrap approach. Its level is close to the nominal level, and it has good power characteristics. There is no substantial difference between the two resampling schemes B_i/D and B_i/A. The robust score χ^2 test is behaving very well (also having larger power) and the bootstrap can only confirm this good performance.

For the more realistic situation that the cluster sizes vary, the picture remains more or less the same. This is illustrated in Table 9.5 and 9.6 where the same settings as in Table 9.3 and 9.4 were used. The level of the Wald test, using chi-squared critical points, is often too high and only the quadratic bootstrap test is able to correct this sufficiently. Also for random litter sizes, the robustified score test is the one to be preferred. We also examined a setting similar to the last one using the MR instead of the pseudolikelihood model. Previous conclusions are confirmed, except that also the quadratic one-step bootstrap tests seem to have some loss in power as compared to the χ^2 tests, see Table 9.7.

9.4.3 The theophylline data

Lindstrom et al. (1990) investigated the effect in mice of the chemical theophylline. In each of the three models, the beta-binomial (BB) model (Section 2.6.1), the conditional model (Section 3.2.2) of Molenberghs and Ryan (MR) and the pseudolikelihood (PL) model (Section 8.2), we used parameterization (8.16). We are interested in testing the no dose effect hypothesis on the main effect parameter ($H_0 : \theta_{11} = 0$).

		$\theta_{20} = 0.2$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	15.2	8.4	8.9	8.8	9.8
	R_n	19.0	13.2	—	14.5	—
-3.5	W_n	30.2	24.9	30.2	26.2	27.5
	R_n	33.8	30.6	—	30.8	—
-3.0	W_n	41.0	34.7	35.2	33.8	36.6
	R_n	43.7	39.8	—	42.8	—
-2.5	W_n	51.5	44.6	48.1	45.5	47.6
	R_n	57.3	54.3	—	53.9	—
		$\theta_{20} = 0.3$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	11.6	7.1	11.3	7.0	12.7
	R_n	15.9	10.1	—	13.5	—
-3.5	W_n	25.2	17.4	23.4	21.0	23.5
	R_n	25.4	25.0	—	22.0	—
-3.0	W_n	29.1	23.6	29.5	24.9	27.7
	R_n	37.1	31.3	—	34.6	—
-2.5	W_n	45.2	38.8	40.7	38.0	39.7
	R_n	49.4	48.9	—	50.0	—

Table 9.4: Simulated power (as %), significance level =0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. Clustersize=12, $\theta_{11} = 1$. $H_0 : \theta_{11} = 0$.

		$\theta_{20} = 0.2$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	9.84*	9.04*	6.02	9.64*	5.62
	R_n	5.62	4.62	—	3.41	—
-3.5	W_n	8.40*	7.80*	4.80	6.60	4.40
	R_n	5.80	5.80	—	4.20	—
-3.0	W_n	8.40*	7.40*	5.60	6.80	5.40
	R_n	6.40	5.80	—	5.00	—
-2.5	W_n	6.00	5.00	4.40	5.00	4.20
	R_n	6.00	4.60	—	4.00	—
		$\theta_{20} = 0.3$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	11.72*	10.71*	4.85	9.70*	4.65
	R_n	3.84	4.04	—	2.63*	—
-3.5	W_n	9.00*	8.60*	4.80	7.80*	4.80
	R_n	6.40	5.80	—	4.80	—
-3.0	W_n	8.40*	6.80	5.20	5.80	4.20
	R_n	6.40	4.60	—	4.60	—
-2.5	W_n	6.80	5.60	4.00	5.40	3.60
	R_n	6.00	4.40	—	3.80	—

$*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%

Table 9.5: Simulated type I errors (as %), significance level =0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. Random clustersizes. $H_0 : \theta_{11} = 0$.

		$\theta_{20} = 0.2$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	14.5	9.8	15.0	9.2	14.3
	R_n	18.4	15.7	—	17.2	—
-3.5	W_n	26.0	24.4	25.2	24.2	27.7
	R_n	27.5	23.4	—	23.8	—
-3.0	W_n	34.2	29.4	31.1	28.2	31.5
	R_n	36.2	34.6	—	34.8	—
-2.5	W_n	51.2	51.1	50.6	46.9	48.5
	R_n	54.4	52.4	—	51.6	—
		$\theta_{20} = 0.3$				
θ_{10}		χ^2	B_1/D	B_2/D	B_1/A	B_2/A
-4.0	W_n	12.8	9.8	13.4	9.4	13.9
	R_n	18.4	16.8	—	17.9	—
-3.5	W_n	18.9	16.6	20.0	19.1	19.3
	R_n	20.4	18.0	—	18.6	—
-3.0	W_n	28.6	25.9	26.2	27.2	28.0
	R_n	31.2	30.6	—	29.9	—
-2.5	W_n	44.9	43.6	43.7	39.9	42.4
	R_n	47.7	48.0	—	45.3	—

Table 9.6: Simulated power (as %), significance level =0.05. Data are generated with the beta-binomial model and fitted using the pseudolikelihood model. $\theta_{11} = 1$. Random clustersizes. $H_0 : \theta_{11} = 0$.

		$\theta_{20}=0.1$			$\theta_{20}=0.2$		
θ_{10}		χ^2	B_1/D	B_2/D	χ^2	B_1/D	B_2/D
				$\theta_{11}=0$			
-3.5	W_n	7.4*	8.2*	7.4*	9.2*	9.2*	9.2*
	R_n	4.2	4.0	—	3.8	4.2	—
-3.0	W_n	6.2	6.6	5.8	8.2*	8.8*	8.0*
	R_n	4.0	3.8	—	5.6	6.0	—
-2.5	W_n	6.4	6.4	5.8	6.4	7.0*	6.4
	R_n	4.2	4.2	—	4.0	4.4	—
				$\theta_{11}=1$			
-3.5	W_n	36.9	36.0	28.8	30.0	21.4	16.8
	R_n	36.0	27.1	—	26.6	25.0	—
-3.0	W_n	56.0	53.7	48.2	31.4	28.9	23.3
	R_n	56.1	58.2	—	34.0	33.2	—
-2.5	W_n	70.0	70.6	65.2	52.5	52.8	43.6
	R_n	72.6	69.4	—	55.7	53.1	—

Table 9.7: Simulated type I errors and power (as %), significance level =0.05. Data are generated with the beta-binomial model and fitted using the MR model Random clustersizes $H_0 : \theta_{11} = 0$.

		External		Visceral		Skeletal		Collapsed	
		W_n	R_n	W_n	R_n	W_n	R_n	W_n	R_n
BB	χ^2	22.85	23.54	4.50*	—	0.00*	31.98	4.58*	5.33
	B_1/A	22.78	23.25	0.50*	—	0.00*	13.10	6.16	6.96
	B_1/D	23.98	24.66	0.90*	—	0.00*	13.70	5.32	6.00
GEE2	χ^2	14.60	13.90	3.61*◇	17.27◇	0.00*	36.86	3.45*	4.20*
	B_1/A	15.70	14.40	1.40*◇	18.80◇	0.00*	19.43	4.05*	4.80*
	B_1/D	14.20	13.30	1.00*◇	20.20◇	0.00*	22.86	1.70*	2.20*
MR	χ^2	9.76	16.98	3.55*	18.26	2.19*	18.82	3.40*	7.54
	B_1/A	9.30	18.80	0.60*	21.60	0.30*	16.60	2.30*	7.20
	B_1/D	7.80	16.30	0.50*	19.70	0.20*	14.50	1.50*	5.20
	B_2/A	8.70	—	5.20	—	28.10	—	2.50*	—
	B_2/D	7.70	—	3.40*	—	27.40	—	1.70*	—
PL	χ^2	12.26	17.06	3.47*	18.56	0.64*	16.08	4.11*	6.83
	B_1/A	12.30	19.10	0.50*	21.70	0.00*	7.30	3.90*	6.70
	B_1/D	11.10	16.10	0.50*	19.70	0.00*	6.70	2.20*	5.20
	B_2/A	17.60	—	10.30	—	1.90*	—	8.20	—
	B_2/D	13.90	—	8.60	—	1.10*	—	5.40	—

A $*$ denotes rejection at the 5% level and a $\diamond$ indicates that a Moore-Penrose generalized inverse is used to obtain the results.

Table 9.8: Analysis of the NTP data on theophylline with $H_0 : \theta_{11} = 0$. P-values are shown as %.

The results are shown in Table 9.8. The table shows the P-values (as %) of the different test statistics discussed before. We also included the results from a GEE2 estimation method based on the first four order moments of the Bahadur model. The Bahadur model for clustered binary data has the same first- and second-order moments as the beta-binomial model. For more details, we refer to Bahadur (1961) and Kupper and Haseman (1978) for the Bahadur model and to Zhao and Prentice (1990) for the GEE2 estimation method.

For *external* malformations, the results of the different tests are almost the same for the beta-binomial model and close to each other for the other models. There seems to be no significant effect of theophylline on the external malformation probability.

For *visceral* and *skeletal* malformations, there is a striking discrepancy between the Wald and the score test. The significance of the Wald test should be interpreted with care (see also the inflated type I errors in Tables 9.1, 9.3 and 9.5). The quadratic bootstrap seems to correct the Wald test in the direction of the score test. Compared with the chi-squared tests, the bootstrap score test has higher P-values for visceral malformation and lower P-values for skeletal malformation. For the beta-binomial model there were convergence problems when fitting the null model. For this reason, the score statistics could not be obtained. Also for the GEE2 model some problems arose, but these could be avoided by using a Moore-Penrose generalized inverse of the matrix $\boldsymbol{J}$. For visceral malformations, the null hypothesis cannot be rejected. A conclusion is less clear for skeletal malformations. Except for the quadratic bootstrap for the MR model, all Wald tests indicate a significant dose effect, while all score tests indicate no effect (although the bootstrap score PL test is getting close to 5 %). Since the score tests showed to have a good behavior in our simulation study, we might believe the results of these test, though further investigation might be necessary to come to a conclusion. Finally, also note that, for all three types of malformation, the different Wald tests lead to highly variable P-values whereas the score tests are much more stable.

The *collapsed* version seems to indicate that theophylline might have an effect on the development of foetuses. Here, all score P-values are between 0.0220 and 0.0754 and the Wald P-values are between 0.0150 and 0.0820. This indicates that a separate analysis of each type of malformation can lead to misleading conclusions and that for this type of problems one has to consider all types jointly as a multivariate response

or at least, as we did here, a collapsed malformation indicator.

9.5 Discussion

In this chapter we examine the behavior of the robust Wald and score test for clustered binary data, based on (pseudo-)likelihood estimation and allowing for misspecification. A new semiparametric bootstrap test based on resampling scores and their derivatives is proposed and is contrasted with the classical chi-squared robust tests. Asymptotically the bootstrap method is consistent and finite sample simulations show substantial improvements in size, especially for the quadratic bootstrap Wald test. This quadratic one-step method is easily obtained for exponential family based distributions, but calculations can be more difficult for other distributions. Since our simulations showed the poor behavior of the chi-squared Wald test and suggested good results for the chi-squared score test, this latter test might be a good choice to use for statistical analysis. Unfortunately, most of the existing statistical software packages do not provide this score test automatically.

Although the technique is introduced in the context of (pseudo-)likelihood, it can also be applied for estimating equations in general, as is being illustrated for the data on theophylline. In Chapter 10 we investigate other domains of application of the quadratic bootstrap method, such as the construction of confidence intervals.

Chapter 10

An Application of the One-Step Quadratic Bootstrap to Bias Correction and the Construction of Confidence Intervals

10.1 Introduction

The use of the one-step linear bootstrap has been advocated in cases where the computation of the estimators requires iterations. It is a simple and attractive method based on a linear representation of the estimators and is very appealing when the asymptotic distribution is very complicated or even unknown. Instead of calculating thousands of bootstrap estimates iteratively, the one-step approach uses only one step of the iterative process. The original idea is due to Schucany and Wang (1991) and is presented in a more general setting in Section 9.3.2. Although the one-step linear method is asymptotically equivalent to the fully iterative one, our experience in the setting of binary response data showed to interpret its results with care (see Sections 9.4.2 and 10.3). Compared to the normal based confidence intervals, the linear one-step bootstrap tends to produce shorter confidence intervals but simulations show an equivalent decrease in coverage probability. Also, by definition, the

one-step approach is not able to detect the bias of the estimator. These shortcomings seem to be greatly eliminated by the one-step *quadratic* bootstrap, which allows the construction of a bias corrected estimator and improved confidence intervals.

The quadratic bootstrap is based on a quadratic approximation of the estimator and has been studied in Chapter 9 for hypotheses tests. This chapter focuses on the use of this improved one-step bootstrap approach to sharpen estimators in terms of bias and mean squared error and to improve interval estimation in the context of logistic regression. In a non-bootstrap context, a quadratic approximation to maximum likelihood estimators in logistic regression is studied by Jennings (1986).

The chapter is organized as follows. In Section 10.2, we show how the quadratic bootstrap can be used to define a bias corrected estimator, and how a double bootstrap algorithm can be implemented to estimate the variance of a bias corrected estimator. In Section 10.3 we construct improved confidence intervals, of which the finite sample performance is illustrated by a small simulation study. We end this chapter by a short discussion in Section 10.4.

These results can also be found in Aerts and Claeskens (1998b).

10.2 Improved estimators

Although the estimator $\widehat{\boldsymbol{\theta}}_n$ obtained from (9.2) is asymptotically unbiased, we will show how the bootstrap procedure using the quadratic one-step bootstrap can be used for finite sample bias correction.

10.2.1 Bias corrected estimation

Finite sample bias correction is often obtained by application of bootstrap methods. For some recent literature about this subject, we refer to Kim and Singh (1998) and MacKinnon and Smith (1998). In practical applications a large number, say B, resamples are taken, resulting in a set of B bootstrap estimators $\widehat{\boldsymbol{\theta}}_n^{*1}, \ldots, \widehat{\boldsymbol{\theta}}_n^{*B}$. From this set a bias corrected estimator is defined as

$$\widehat{\boldsymbol{\theta}}_n^{bc} = 2\widehat{\boldsymbol{\theta}}_n - \frac{1}{B}\sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_n^{*i}. \tag{10.1}$$

For bias estimation, the second order approximation turns out to be very useful, which is not completely unexpected since the bias is a second order aspect of the

asymptotic properties of the estimator. The intuition behind equation (10.1) should be clear. We subtract from the estimator $\widehat{\boldsymbol{\theta}}_n$, the estimated bias based on the B bootstrap replicates. The estimated bias is calculated as $\sum_{i=1}^{n} \widehat{\boldsymbol{\theta}}_n^{*i}/B - \widehat{\boldsymbol{\theta}}_n$.

Table 10.1 shows that the quadratic one-step bootstrap estimator is quite able to estimate the finite sample bias.

The settings in this simulation were as follows. We generated 2000 data sets of size $n = 10$ and $n = 25$, for each value of x, from a logistic regression model

$$\text{logit}\{P(Y = 1)\} = \beta_0 + \beta_1 x,$$

with (β_0, β_1) equal to (-1,-1), (-2.5, 1) or (-2.5, 2), and $x = 0, 0.25, 0.5$ and 1. For each of these 2000 data sets we constructed 1000 one-step quadratic bootstrap replicates, the latter were used to obtain the bias corrected estimates $(\hat{\beta}_0^{bc}, \hat{\beta}_1^{bc})$. There were some numerical problems with obtaining the original estimates $(\hat{\beta}_0, \hat{\beta}_1)$; the number of generated sets of data without these problems is indicated in Table 10.1 (conv).

An important observation is that the bias correction even *decreases* the variance, as the simulated standard deviation $\sigma(\hat{\beta}_0^{bc})$ and $\sigma(\hat{\beta}_1^{bc})$ are, for all settings in this study, smaller than the corresponding simulated values of $\sigma(\hat{\beta}_0)$ and $\sigma(\hat{\beta}_1)$, respectively. In fact, the reduction in Mean Squared Error (MSE) is quite spectacular. Note that this reduction is much larger for the smallest sample size (n=10) than for the setting where n=25. The explanation for this is the finite sample bias which is less severe for larger samples. In those samples, there is less need for bias correction.

10.2.2 Double bootstrap and variance estimation

Since the bias corrected estimator seems to have interesting properties, we also might study its distribution. This estimator is already based on a bootstrap resampling scheme, therefore we will need a second bootstrap stage to obtain this extra amount of information. The double bootstrap procedure reads as follows.

1. Using a nonparametric resampling scheme, that is, resample the data directly, we construct a set of bootstrap estimators. In the same way as described before, using the one-step quadratic bootstrap, perform a bias correction, and next, construct the bias corrected estimator.

2. To study the distribution of the resulting estimator, we construct from each of the resampled data new bootstrap estimators via the one-step quadratic

	$\beta_0 = -1$		$\beta_0 = -2.5$		$\beta_0 = -2.5$	
	$\beta_1 = -1$		$\beta_1 = 1$		$\beta_1 = 2$	
	$n = 10$	$n = 25$	$n = 10$	$n = 25$	$n = 10$	$n = 25$
conv.	1992	2000	1895	1997	1967	2000
$E(\hat{\beta}_0)$	-1.019	-1.020	-2.604	-2.582	-2.695	-2.599
$E(\hat{\beta}_0^{bc})$	-0.978	-1.006	-2.360	-2.486	-2.459	-2.513
$\sigma(\hat{\beta}_0)$	0.645	0.375	0.889	0.611	0.914	0.559
$\sigma(\hat{\beta}_0^{bc})$	0.588	0.362	0.724	0.554	0.758	0.519
$\frac{MSE(\hat{\beta}_0^{bc})}{MSE(\hat{\beta}_0)}$	0.833	0.933	0.679	0.807	0.660	0.835
$E(\hat{\beta}_1)$	-1.265	-1.070	0.889	1.027	2.152	2.096
$E(\hat{\beta}_1^{bc})$	-1.043	-0.988	0.852	1.001	1.959	2.022
$\sigma(\hat{\beta}_1)$	1.511	0.795	1.555	0.908	1.303	0.779
$\sigma(\hat{\beta}_1^{bc})$	1.314	0.747	1.297	0.836	1.119	0.735
$\frac{MSE(\hat{\beta}_1^{bc})}{MSE(\hat{\beta}_1)}$	0.734	0.875	0.701	0.846	0.728	0.878
$\frac{MSE(\hat{\boldsymbol{\beta}}^{bc})}{MSE(\hat{\boldsymbol{\beta}})}$	0.749	0.885	0.696	0.834	0.705	0.863

Table 10.1: Simulated mean, standard deviation and mean squared error values of original and bias corrected estimators.

bootstrap. These values can now be used to compute the estimator's variance, quantiles, etc.

Table 10.2 illustrates how a double bootstrap method can be used to get estimators for the variance of the bias corrected estimator. We restrict attention to the slope parameter and to samples of size 10. It seems that all variance estimators (classical robust and naive variance estimators and double bootstrap variance estimator) somewhat underestimate the true variability. The double bootstrap variance estimator however is clearly less variable which leads to a substantial reduction in mean squared error.

We now list some details on the results of the simulation study presented in Table 10.2. The design of the covariate x is the same as in the previous setting. The number of simulated data sets equals 500, and for each simulated data set we used 500 replicates for the outer and 250 replicates for the inner bootstrap loop. This table shows us something about the variability of the variance estimators. The simulated mean of all three estimators for the standard deviation is everywhere smaller than the simulated standard deviation $\sigma(\hat{\beta}_1)$. In this simulation we observe the following ordering:

$$\sigma(\hat{\beta}_1) \geq E[\hat{\sigma}_n(\hat{\beta}_1)] \geq E[\hat{\sigma}_r(\hat{\beta}_1)] \geq E[\hat{\sigma}(\hat{\beta}_1^{bc})],$$

except for the first case where $E[\hat{\sigma}(\hat{\beta}_1^{bc})] \geq E[\hat{\sigma}_r(\hat{\beta}_1)]$.

10.3 Bootstrap confidence intervals

10.3.1 Construction

Confidence intervals for the parameter $\boldsymbol{\theta}$ can be derived from the asymptotic normality result, by using the Wald statistic as a pivot. Next to this classical approach, the appropriate quantiles can be selected from the bootstrap approximation to the asymptotic distribution. In this section we construct bootstrap confidence intervals from the so-called hybrid bootstrap (see, e.g., Shao and Tu, 1995, Sections 4.1 and 4.2). A $100(1-\alpha)\%$ confidence interval for the parameter $\boldsymbol{\theta}$ is defined as

$$\{\boldsymbol{\theta} : z_L^* \leq \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq z_R^*\}$$

		$\beta_0 = -1$	$\beta_0 = -2.5$	$\beta_0 = -2.5$
	$n = 10$	$\beta_1 = -1$	$\beta_1 = 1$	$\beta_1 = 2$
	conv.	499	480	493
	$\sigma(\hat{\beta}_1)$	1.511	1.555	1.303
Robust	$E(\hat{\sigma}_r(\hat{\beta}_1))$	1.239	1.358	1.207
	$\mathrm{Var}(\hat{\sigma}_r(\hat{\beta}_1))$	0.152	0.186	0.072
	$MSE(\hat{\sigma}_r(\hat{\beta}_1))$	0.226	0.225	0.081
Naive	$E(\hat{\sigma}_n(\hat{\beta}_1))$	1.320	1.514	1.258
	$\mathrm{Var}(\hat{\sigma}_n(\hat{\beta}_1))$	0.171	0.203	0.082
	$MSE(\hat{\sigma}_n(\hat{\beta}_1))$	0.207	0.205	0.084
Bias-corrected	$\sigma(\hat{\beta}_1^{bc})$	1.314	1.297	1.119
	$E(\hat{\sigma}(\hat{\beta}_1^{bc}))$	1.247	1.103	1.067
	$\mathrm{Var}(\hat{\sigma}(\hat{\beta}_1^{bc}))$	0.136	0.101	0.022
	$MSE(\hat{\sigma}(\hat{\beta}_1^{bc}))$	0.141	0.138	0.025

Table 10.2: Simulated mean, variance and mean squared error values of naive, robust and double bootstrap variance estimators.

where z_L^* is the $100\alpha/2\%$ and z_R^* is the $100(1-\alpha/2)\%$ quantile of the distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)$. In the following section, confidence intervals based on the quadratic one-step bootstrap estimator are compared to the normal-based intervals.

10.3.2 Simulation results

In linear logistic regression models we will illustrate the usefulness of quadratic one-step bootstrap estimators to build confidence intervals. Data are generated from the following logistic model:

$$\text{logit}\{P(Y=1)\} = \theta_0 + \theta_1 x,$$

where x belongs to $\{0, .25, .5, 1\}$, θ_0 is -1.5 or -2.5; θ_1 takes values in {1,2} and the sample size varies between 5 and 25 observations per dose level. For each simulated data set, 1000 bootstrap estimators were constructed.

For 1000 simulated data sets, Table 10.3 shows the mean lengths of the confidence intervals for both the normal and the bootstrap procedures. Intervals were constructed at the 90%, 95% and 99% confidence level. Since it turned out that the bootstrap intervals were everywhere shorter than their classical counterparts, the table also shows the percentage of reduction in length.

The most important observation is that the bootstrap intervals all have significantly smaller length for all the situations considered in this simulation setting. Moreover, as shown in Table 10.4, also their coverage probability is usually somewhat higher.

Except for the smallest sample size and $\theta_0 = -2.5$ where the coverage probabilities are too small for the quadratic bootstrap, the quadratic bootstrap performs extremely well. Note that in those cases the reduction of length of the confidence intervals is enormous: even up to 22%. We should interpret these results with care, since exactly for these extreme cases, the number of convergences is very small. The reason for this is not directly clear. Possibly the number of events was very small in those cases, which caused some extreme situations to lead to divergent estimators.

Table 10.5 shows the reduction in length of confidence intervals for the slope parameter in a linear logistic regression model when the linear one-step bootstrap is applied. These results should be compared with those of Table 10.3. It clearly demonstrates the need of the quadratic approximation, for which the results are everywhere superior to those of the linear approximation. There still is a reduction

		Normal approx.			Quadr. Bootstrap		
Obs/d	θ_1	90%	95%	99%	90%	95%	99%
					$\theta_0 = -2.5$		
10	0	4.77	5.69	7.48	4.24 (11.3%)	4.87 (14.3%)	5.81 (22.3%)
15		4.39	5.23	6.87	4.01 (8.51%)	4.68 (10.4%)	5.77 (15.9%)
25		3.60	4.29	5.64	3.38 (5.93%)	4.40 (6.65%)	5.61 (8.70%)
10	1	4.38	5.22	6.87	4.03 (8.09%)	4.70 (10.1%)	5.79 (15.6%)
15		3.76	4.48	5.88	3.53 (5.97%)	4.16 (6.99%)	5.30 (10.0%)
25		2.80	3.34	4.39	2.71 (3.51%)	3.22 (3.66%)	4.20 (4.50%)
10	2	3.97	4.73	6.21	3.76 (5.32%)	4.40 (6.92%)	5.61 (9.74%)
15		3.20	3.81	5.00	3.07 (3.98%)	3.64 (4.28%)	4.69 (6.18%)
25		2.45	2.92	3.84	2.39 (2.52%)	2.84 (2.80%)	3.70 (3.60%)
					$\theta_0 = -1.5$		
10	1	3.40	4.05	5.32	3.27 (3.67%)	3.87 (4.42%)	4.98 (6.38%)
15		2.70	3.21	4.23	2.63 (2.34%)	3.13 (2.60%)	4.07 (3.63%)
25		2.05	2.44	3.21	2.03 (1.22%)	2.41 (1.39%)	3.16 (1.66%)
10	2	3.40	4.05	5.32	3.25 (4.25%)	3.82 (5.70%)	4.82 (9.52%)
15		2.69	3.20	4.21	2.62 (2.62%)	3.09 (3.44%)	3.96 (5.75%)
25		2.04	2.44	3.20	2.01 (1.41%)	2.39 (1.96%)	3.10 (3.21%)

Table 10.3: Mean length (reduction in length) of confidence intervals for the slope parameter in a linear logistic regression model.

			Normal approx.			Quadr. Bootstrap		
Obs/d	θ_1	conv	90%	95%	99%	90%	95%	99%
				$\theta_0 = -2.5$				
10	0	(850)	81.76*	88.71*	93.65*	81.29*	86.82*	87.88*
15		(950)	83.47*	89.58*	96.00*	86.84*	90.63*	93.37*
25		(995)	84.73*	91.86*	99.10	88.64	95.78	98.89
10	1	(944)	84.32*	91.00*	97.24	87.92	91.10*	94.81*
15		(993)	88.82	94.26	98.79	92.15	96.07	98.59
25		(997)	88.16	93.58	98.60	91.57	96.19	99.20
10	2	(985)	91.47	96.14	99.29	93.81•	96.55	98.78
15		(997)	88.67	94.88	98.80	90.97	96.19	99.70
25		(1000)	90.60	95.20	99.30	92.20	96.00	99.70
				$\theta_0 = -1.5$				
10	1	(1000)	89.30	94.40	99.30	93.00•	97.10•	99.70
15		(1000)	88.20	94.00	98.90	91.10	95.90	99.60
25		(1000)	87.90	94.50	99.40	90.50	96.40	99.60
10	2	(1000)	89.10	94.70	99.50	93.00•	96.40	99.50
15		(1000)	91.20	95.90	99.50	93.00•	96.90•	99.50
25		(1000)	89.00	94.60	99.30	89.10	95.30	99.30

Table 10.4: Simulated coverage probabilities (as %) of confidence intervals for the slope parameter in linear logistic regression model. A $*$ indicates a too small, and $\bullet$ a too large number, according to the 0.01 level.

		Normal approx.			Linear Bootstrap		
Obs/d	θ_1	90%	95%	99%	90%	95%	99%
					$\theta_0 = -2.5$		
10	1	4.38	5.22	6.87	4.17 (4.97%)	5.10 (2.36%)	6.65 (3.10%)
15		3.76	4.48	5.88	3.67 (2.39%)	4.43 (9.86%)	5.84 (0.67%)
25		2.80	3.34	4.39	2.78 (0.88%)	3.32 (0.59%)	4.38 (0.12%)
10	2	3.97	4.73	6.21	3.87 (2.45%)	4.62 (2.16%)	6.06 (2.53%)
15		3.20	3.81	5.00	3.15 (1.47%)	3.76 (1.34%)	4.93 (1.42%)
25		2.45	2.92	3.84	2.43 (0.89%)	2.90 (0.77%)	3.81 (0.74%)
					$\theta_0 = -1.5$		
10	1	3.40	4.05	5.32	3.34 (1.85%)	3.97 (1.98%)	5.77 (2.78%)
15		2.70	3.21	4.23	2.67 (1.21%)	3.18 (1.24%)	4.15 (1.74%)
25		2.05	2.44	3.21	2.04 (0.58%)	2.43 (0.72%)	3.19 (0.78%)
10	2	3.40	4.05	5.32	3.34 (1.83%)	3.95 (2.43%)	5.13 (3.72%)
15		2.69	3.20	4.21	2.65 (1.31%)	3.15 (1.59%)	4.11 (2.16%)
25		2.04	2.44	3.20	2.03 (0.65%)	2.41 (0.86%)	3.16 (1.18%)

Table 10.5: Mean length (reduction in length) of confidence intervals for the slope parameter in a linear logistic regression model.

			Normal approx.			Linear Bootstrap		
Obs/d	θ_1	conv	90%	95%	99%	90%	95%	99%
				$\theta_0 = -2.5$				
10	1	(944)	84.32*	91.00*	97.24	80.08*	88.77*	93.96*
15		(993)	88.82	94.26	98.79	85.60*	90.74*	96.88*
25		(997)	88.16	93.58	98.60	86.46*	92.68*	97.89*
10	2	(985)	91.47	96.14	99.29	86.70*	92.49*	98.29
15		(997)	88.67	94.88	98.80	86.36*	93.48	97.89*
25		(1000)	90.60	95.20	99.30	89.20	94.10	98.50
				$\theta_0 = -1.5$				
10	1	(1000)	89.30	94.40	99.30	87.10*	92.50*	97.00*
15		(1000)	88.20	94.00	98.90	86.70*	92.70*	97.80*
25		(1000)	87.90	94.50	99.40	87.50*	93.50	98.80
10	2	(1000)	89.10	94.70	99.50	85.70*	91.20*	97.00*
15		(1000)	91.20	95.90	99.50	89.50	95.00	98.80
25		(1000)	89.00	94.60	99.30	88.60	94.40	98.50

Table 10.6: Simulated coverage probabilities (as %) of confidence intervals for the slope parameter in linear logistic regression model. A $*$ indicates a too small number, according to the 0.01 level.

in length of the confidence intervals when the linear one-step bootstrap is applied, but in Table 10.6 we see that the coverage probabilities decrease too! For most of the settings in this simulation study, the coverage probabilities were even much too small for the linear bootstrap, while they were within the range of allowable values (at 1%) for the normal approximation. As a summary, in these settings, the linear bootstrap perform worse than the much simpler normal approximation, this in contrast to the quadratic bootstrap, which outperforms both.

In Tables 10.7 and 10.8, data are generated from a beta-binomial model with clustersize two. Results of robust statistics are shown. The conclusions drawn from these tables are consistent with the previous ones. The quadratic bootstrap everywhere reduces the length of the confidence intervals. This reduction in length is usually more pronounced for the highest confidence levels. For all examined situations, the simulated coverage probabilities were within the range of allowable values. Although they were slightly higher for the quadratic bootstrap, the differences were minor and not statistically significant at the 1% level.

10.4 Discussion

It should be clear that the applications of the one-step bootstrap method are not restricted to those shown here. There are various situations in which this resampling technique can be used. Let us consider the estimation of the probability of success in a logistic model, e.g.,

$$\text{logit}\{\pi(x)\} = \theta_0 + \theta_1 x.$$

In this chapter we focused on bias correction and the construction of confidence intervals for the parameters θ_0 or θ_1, representing, respectively, the intercept term and the slope term of the linear logistic model. Once we have estimators available for these parameters, also $\pi(x)$ can be estimated. In toxicity studies, see e.g. the data on theophylline, there are usually only a small number of dose levels. It would be interesting to obtain confidence intervals for the parameter $\pi(x)$ at exactly these dose levels. To obtain simultaneous confidence intervals, we can, for example, use the rank based approach as explained in Section 3.5.1. If bootstrap replicates of each of these proportions $\pi(x)$ are obtained, the same construction will yield the desired simultaneous confidence intervals. A study of its performance in practice is a topic of future research.

		Normal approx.			Quadr. Bootstrap		
Obs/d	θ_1	90%	95%	99%	90%	95%	99%
				$F(\rho) = 0.1, \quad \theta_0 = -1.5$			
10	1	2.32	2.76	3.63	2.27 (2.34%)	2.69 (2.74%)	3.50 (3.64%)
15		1.89	2.25	2.95	1.85 (1.75%)	2.05 (1.91%)	2.88 (2.44%)
25		1.46	1.74	2.28	1.44 (1.00%)	1.72 (0.92%)	2.26 (1.18%)
10	2	2.32	2.77	2.64	2.26 (2.73%)	2.67 (3.49%)	3.44 (5.38%)
15		1.89	2.26	2.96	1.86 (1.80%)	2.20 (2.34%)	2.86 (3.68%)
25		1.45	1.73	2.28	1.44 (1.12%)	1.71 (1.40%)	2.23 (1.91%)
				$F(\rho) = 0.1, \quad \theta_0 = -2.5$			
10	1	3.18	3.79	4.98	3.02 (5.05%)	3.58 (5.47%)	4.61 (7.36%)
15		2.58	3.08	4.04	2.49 (3.47%)	2.97 (3.32%)	3.88 (4.04%)
25		1.98	2.36	3.10	1.94 (2.01%)	2.31 (1.98%)	3.04 (1.88%)
10	2	2.78	3.31	4.35	2.67 (3.98%)	3.16 (4.33%)	4.09 (5.94%)
15		2.25	2.68	3.52	2.19 (2.71%)	2.59 (3.05%)	3.38 (4.02%)
25		1.72	2.05	2.70	1.70 (1.57%)	2.02 (1.65%)	2.64 (2.06%)

Table 10.7: Mean length (reduction in length) of confidence intervals for the slope parameter. The size of each cluster is 2. Robust logistic regression.

			Normal approx.			Quadr. Bootstrap		
Obs/d	θ_1	conv	90%	95%	99%	90%	95%	99%
			$F(\rho) = 0.1, \quad \theta_0 = -1.5$					
10	1	(1000)	87.60	93.30	98.40	89.90	94.80	99.30
15		(1000)	89.90	95.10	98.80	90.90	95.70	98.90
25		(1000)	90.00	94.80	99.00	90.10	95.80	99.40
10	2	(1000)	90.10	94.50	99.40	90.80	95.40	99.20
15		(1000)	89.70	94.10	98.50	90.60	94.20	98.50
25		(1000)	89.40	94.20	99.10	89.30	94.00	99.20
			$F(\rho) = 0.1, \quad \theta_0 = -2.5$					
10	1	(995)	89.15	94.17	98.39	90.95	95.58	98.79
15		(1000)	87.60	93.60	99.00	89.80	95.60	99.40
25		(1000)	88.50	94.90	99.20	90.50	96.50	99.60
10	2	(1000)	88.10	94.20	99.10	90.40	95.40	99.10
15		(1000)	88.70	94.20	98.80	90.20	95.50	98.60
25		(1000)	99.70	95.80	99.00	90.50	96.60	99.40

Table 10.8: Simulated coverage probabilities (as %) of confidence intervals for the slope parameter. The size of each cluster is 2.

Another area of application is the determination of a benchmark dose or a virtually safe dose. Instead of basing the lower confidence limit on asymptotic approximations based on the Gaussian distribution, bootstrap approximations will provide an interesting alternative.

Reference List

Aerts, M. and Claeskens, G. (1997). "Local polynomial estimators in multiparameter likelihood models", *Journal of the American Statistical Association*, **92**, 1536–1545.

Aerts, M. and Claeskens, G. (1998a). "Bootstrap tests for misspecified models, with application to clustered binary data", Submitted.

Aerts, M. and Claeskens, G. (1998b). "A note on the quadratic bootstrap and improved estimation in logistic regression", Unpublished Manuscript, Limburgs Universitair Centrum, Diepenbeek.

Aerts, M. and Claeskens, G. (1999). "Bootstrapping pseudolikelihood models for clustered binary data", *Annals of the Institute of Statistical Mathematics*, **51**, to appear.

Aerts, M., Claeskens, G. and Hart, J.D. (1998). "Testing lack of fit in multiple regression", Submitted.

Aerts, M., Claeskens, G. and Hart, J.D. (1999). "Testing the fit of a parametric function", *Journal of the American Statistical Association*, **94**, to appear.

Aerts, M., Claeskens, G. and Wand, M.P. (1999). "Some theory for penalized spline additive models", Submitted.

Aerts, M. and Veraverbeke, N. (1995). "Bootstrapping a nonparametric polytomous regression model", *Mathematical Methods of Statistics*, **4**, 189-200.

Ahn, H. and Chen, J.J. (1997). "Tree-structured logistic model for over-dispersed binomial data with applications to modeling developmental effects", *Biometrics*, **53**, 435–455.

Akaike, H. (1974). "A new look at statistical model identification", *I.E.E.E. Transactions on Automatic Control,* **19**, 716–723.

Aragaki, A. and Altman, N. (1997). "Local polynomial regression for binary response", Unpublished Manuscript.

Arnold, B.C., Castillo, E. and Sarabia, J.-M. (1992). *Conditionally Specified Distributions.* Lecture Notes in Statistics 73, New York: Springer-Verlag.

Arnold, B.C. and Strauss, D. (1991). "Pseudolikelihood estimation: some examples", *Sankhyā B*, **53**, 233–243.

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

Bahadur, R.R. (1961). "A representation of the joint distribution of responses of n dichotomous items", In *Studies in item analysis and prediction*, H. Solomon (ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

Barry, D. (1993). "Testing for additivity of a regression function", *The Annals of Statistics*, **21**, 235–254.

Barry, D. and Hartigan, J.A. (1990). "An omnibus test for departures from constant mean", *The Annals of Statistics*, **18**, 1340–1357.

Bartlett, M.S. (1954). "A note on some multiplying factors for various χ^2 approximations", *Journal of the Royal Statistical Society, Series B*, **16**, 296–298.

Beck, N. and Jackman, S. (1998). "Beyond linearity by default: generalized additive models", *Americal Journal of Political Science*, **42**, 596–627.

Beran, R. (1988). "Prepivoting test statistics: a bootstrap view of asymptotic refinements", *Journal of the American Statistical Association*, **83**, 687–697.

Besag, J. (1975). "Statistical analysis of non-lattice data", *The Statistician*, **24**, 179–195.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). "Bayesian computation and stochastic systems", (with discussion), *Statistical Science*, **10**, 1–66.

Betensky, R.A. (1997). "Local estimation of smooth curves for longitudinal data", *Statistics in Medicine*, **16**, 2429–2445.

Bickel, P.J. and Ritov, Y. (1992). "Testing for goodness of fit: a new approach", *Nonparametric Statistics and Related Topics* (A. K. Md. E. Saleh, ed.), North-Holland, Amsterdam, pp.51–57.

Bogdan, M. (1999). "Data driven smooth tests for bivariate normality", *Journal of Multivariate Analysis*, **68**, 26–53.

Boos, D.D. (1992). "On generalized score tests", *The American Statistician,* **46**, 327–333.

Bradley, R.A. and Gart, J.J. (1962). "The asymptotic properties of ML estimators when sampling from associated populations", *Biometrika*, **49**, 205–214.

Breiman, L. and Friedman, J. (1985). "Estimating optimal transformation for multiple regression and correlation", (with discussion), *Journal of the American Statistical Association*, **80**, 580–619.

Brown, C.C. (1982). "On a goodness of fit test for the logistic model based on score statistics", *Communications in Statistics - Theory and Methods*, **11**, 1087–1105.

Bunke, O. (1997). "Bootstrapping in heteroscedastic regression situations", Unpublished manuscript, Sonderforschungsbereich 373, Humboldt University, Berlin.

Burke, M.D. and Gombay, E. (1991). "The bootstrapped maximum likelihood estimator with an application", *Statistics and Probability Letters*, **12**, 421–427.

Cao-Abad, R. and González-Manteiga, W. (1993). "Bootstrap methods in regression smoothing", *Journal of Nonparametric Statistics*, **2**, 379–388.

Carr, G.J. and Portier, C.J. (1993). "An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicology data", *Biometrics*, **49**, 779–791.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). "Generalized partially linear single-index models", *Journal of the American Statistical Association*, **92**, 477–489.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998). "Local estimating equations", *Journal of the American Statistical Association*, **93**, 214–227.

Chambers, J.M. and Hastie T.J. (1991). *Statistical Models in S.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Chanda, K.C. (1954). "A note on the consistency and maxima of the roots of likelihood equations", *Biometrika*, **41**, 56–61.

Chandra, T.K. and Joshi, S.N. (1983). "Comparison of the likelihood ratio, Rao's and Wald's tests and a conjecture of C.R. Rao", *Sankhyā A*, **45**, 226–246.

Chandra, T.K. and Mukerjee, R. (1984). "On the optimality of Rao's statistic", *Communications in Statistics*, **13**, 1507–1515.

Chandra, T.K. and Mukerjee, R. (1985). "Comparison of the likelihood ratio, Wald's and Rao's tests", *Sankhyā A*, **47**, 271–284.

Chaudhuri, P. and Dewanji, A. (1995). "On a likelihood-based approach in non-parametric smoothing and cross-validation", *Statistics and Probability Letters*, **22**, 7–15.

Chen, J.-C. (1994). "Testing goodness of fit of polynomial models via spline smoothing techniques", *Statistics and Probability Letters*, **19**, 65–76.

Claeskens, G. and Aerts, M. (1998). "Bootstrapping local polynomial estimators in likelihood-based models", Submitted.

Claeskens, G. and Aerts, M. (1999). "On local estimating equations in additive multiparameter models," Submitted.

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

Collet, D. (1991). *Modelling Binary Data.* London: Chapman & Hall.

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

Cordeiro, G.M., Botter, D.A. and Ferrari, S.L.P. (1994). "Nonnull asymptotic distributions of three classic criteria in generalized linear models", *Biometrika*, **81**, 709–720.

Cox, D., Koh, E., Wahba, G. and Yandell, B.S. (1988). "Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models", *The Annals of Statistics*, **16**, 113-119.

Cox, D. and O'Sullivan, F. (1990). "Asymptotic analysis of penalized likelihood and related estimators", *The Annals of Statistics*, **18**, 1676–1695.

Cox, D.R. (1972). "The analysis of multivariate binary data", *Applied Statistics*, **21**, 113–120.

Cox, D.R. and Wermuth, N. (1994). "A note on the quadratic exponential binary distribution", *Biometrika*, **81**, 403–408.

Cressie, N. (1991). *Statistics for Spatial Data.* New York: Wiley.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Davison, A.C., Hinkley, D.V. and Schechtman, E. (1986). "Efficient bootstrap simulations", *Biometrika*, **73**, 555–566.

Davison, A.C. and Ramesh, N.I. (1998). "Local likelihood smoothing of sample extremes", Unpublished manuscript, Swiss Federal Institute of Technology, Lausanne.

Davison, A.C. and Tsai, C.-L. (1992). "Regression model diagnostics", *International Statistical Review*, **60**, 337–353.

Dette, H. and Munk, A. (1998). "Testing heteroscedasticity in nonparametric regression", *Journal of the Royal Statistical Society, Series B*, **60**, 693–708.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Efromovich, S. (1996). "On nonparametric regression for iid observations in a general setting", *The Annals of Statistics*, **24**, 1126–1144.

Efron, B. (1979). "Bootstrap methods: another look at the jacknife", *The Annals of Statistics*, **7**, 1–26.

Efron, B. and Hinkley, D.V. (1978). "Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information", *Biometrika*, **65**, 457–487.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Eilers, P.H.C. and Marx, B.D. (1996). "Flexible smoothing with B-splines and penalties", (with discussion), *Statistical Science*, **89**, 89–121.

Engle, R.R. (1984). "Wald, likelihood ratio and Lagrange multiplier tests in econometrics", in *Handbook of Econometrics* (Vol. II), eds. Z. Griliches and M. Intriligator. Amsterdam: North Holland.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Eubank, R.L. (1997). "On testing for no effect in nonparametric regression," Unpublished manuscript, Texas A&M University.

Eubank, R.L. and Hart, J.D. (1992). "Testing goodness-of-fit in regression via order selection criteria", *The Annals of Statistics,* **20**, 1412–1425.

Eubank, R.L., Hart, J.D. and LaRiccia, V.N. (1993). "Testing goodness-of-fit via nonparametric function estimation techniques", *Communications in Statistics - Theory and Methods*, **22**, 3327–3354.

Eubank, R.L., Hart, J.D., Simpson, D.G. and Stefanski, L.A. (1995). "Testing for additivity in nonparametric regression", *The Annals of Statistics,* **23**, 1896–1920.

Eubank, R.L., Li, C.-S. and Wang, S. (1997). "Testing lack of fit of parametric regression models using nonparametric regression techniques", *Proceedings of the New York University Symposium on Recent Developments in Smoothing Methods, May 30, 1997*, 103-131.

Eubank, R.L. and Spiegelman, C.H. (1990). "Testing the goodness of fit of a linear model via nonparametric regression techniques", *Journal of the American Statistical Association*, **85**, 387–392.

Fan, J. (1992). "Design adaptive nonparametric regression", *Journal of the American Statistical Association*, **87**, 998–1545.

Fan, J. (1993). "Local linear regression smoothers and their minimax efficiencies", *The Annals of Statistics*, **21**, 196–216.

Fan, J. (1996). "Test of significance based on wavelet thresholding and Neyman's truncation", *Journal of the American Statistical Association*, **91**, 674–688.

Fan, J., Farmen, M. and Gijbels, I. (1998). "Local maximum likelihood estimation and inference", *Journal of the Royal Statistical Society, Series B,* **60**, 591–608.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman & Hall.

Fan, J., Härdle, W. and Mammen, E. (1998). "Direct estimation of low dimensional components in additive models", *The Annals of Statistics*, **26**, 943–971.

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

Ferguson, T.S. (1996). *A Course in Large Sample Theory.* London: Chapman & Hall.

Firth, D., Glosup, J. and Hinkley, D.V. (1991). "Model checking with nonparametric curves", **78**, 245–252.

Foutz, R.V. (1977). "On the unique consistent solution to the likelihood equations", *Journal of the American Statistical Association*, **72**, 147–148.

Foutz, R.V. and Srivastava, R.C. (1977). "The performance of the likelihood ratio test when the model is incorrect", *The Annals of Statistics*, **5**, 1183–1194.

Frangos, C.C. and Schucany, W.R. (1995). "Improved bootstrap confidence intervals in certain toxicological experiments", *Communications in Statistics*, **24**, 829–844.

Friedman, J.H. and Silverman, B.W. (1989). "Flexible parsimonious smoothing and additive modeling", (with discussion), *Technometrics*, **31**, 3–39.

Freund, R.J. and Littell, R.C. (1991). *SAS System for Regression*, 2nd edition, SAS Institute, Inc., Cary, NC.

Gasser, T. and Müller, H.-G. (1979). "Kernel estimation of regression functions", In *Smoothing techniques for curve estimation* (eds. T. Gasser and M. Rosenblatt). Springer-Verlag, Heidelberg, pp. 23–68.

Gasser, T. and Müller, H.-G. (1984). "Estimating regression functions and their derivatives by the kernel method", *Scandinavian Journal of Statistics*, **11**, 171–185.

Geys, H., Molenberghs, G. and Ryan, L. (1997). "Pseudolikelihood inference for clustered binary data", *Communications in Statistics – Theory and Methods*, **26**, 2743–2768.

Geys, H., Molenberghs, G. and Ryan, L. (1999). "Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology", *Journal of the American Statistical Association*, to appear.

Ghosh, J.K. (1994). *Higher Order Asymptotics.* NSF-CBMS regional conference series in probability and statistics, Vol. 4.

Gourieroux, C., Montfort, A. and Trognon, A. (1984). "Pseudo maximum likelihood methods: theory", *Econometrica,* **52**, 681–700.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman & Hall.

Hall, P. and Wilson, S.R. (1991). "Two guidelines for bootstrap hypothesis testing", *Biometrics*, **47**, 757–762.

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

Härdle, W., Mammen, E. and Müller, M. (1998). "Testing parametric versus semiparametric modeling in generalized linear models", *Journal of the American Statistical Association*, **93**, 1461–1474.

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

Härdle, W. and Marron, J.S. (1995). "Fast and simple scatterplot smoothing", *Journal of Computational Statistics and Data Analysis*, **20**, 1–17.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer-Verlag.

Hart, J.D. and Wehrly, T.E. (1992). "Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models", *Journal of the American Statistical Association*, **87**, 1018–1024.

Hastie, T.J. (1996). "Pseudosplines", *Journal of the Royal Statistical Society, Series B*, **58**, 379–396.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hu, F. and Zidek, J. (1995). "A bootstrap based on the estimation equations of the linear model", *Biometrika*, **82**, 263–275.

Huber, P.J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions", *Proceedings of the 5th Berkeley Symposium*, **1**, 221–233.

Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

Inglot, T., Kallenberg, W.C.M. and Ledwina, T. (1997). "Data driven smooth tests for composite hypotheses", *The Annals of Statistics* **25**, 1222–1250.

Iverson, H.K. and Randles, R.H. (1989). "The effects on convergence of substituting parameter estimates into U-statistics and other families of statistics", *Probability Theory and Related Fields*, **81**, 453–471.

Jennings, D.E. (1986). "Judging inference adequacy in logistic regression", *Journal of the American Statistical Association*, **81**, 471–476.

Kallenberg, W.C.M. and Ledwina, T. (1995). "Consistency and Monte Carlo simulation of a data-driven version of smooth goodness-of-fit tests", *The Annals of Statistics*, **23**, 1594–1608.

Kallenberg, W.C.M. and Ledwina, T. (1997). "Data-driven smooth tests when the hypothesis is composite", *Journal of the American Statistical Association,* **92**, 1094–1104.

Kallenberg, W.C.M. and Ledwina, T. (1999). "Data driven rank tests for independence", *Journal of the American Statistical Association*, **94**, 285–301.

Kauermann, G. and Tutz, G. (1996). "On model diagnostics and bootstrapping in varying coefficient models", Technical report, Technical University Berlin, Germany.

Kent, J.T. (1982). "Robust properties of likelihood ratio tests", *Biometrika*, **69**, 19–27.

Kim, J.-T. (1992). *Testing goodness-of-fit via order selection criteria.* Ph.D. dissertation, Department of Statistics, Texas A&M University.

Kim, Y. and Singh, K. (1998). "Sharpening estimators using resampling", *Journal of Statistical Planning and Inference*, **66**, 121–146.

Kim, W., Linton, O.B. and Hengartner, N.W. (1998). "A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals", Unpublished manuscript, Yale University.

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

Klein, R., Klein, B.E.K., Moss, S.E., Davis, M.D. and DeMets, D.L. (1984), "The Wisconsin epidemiologic study of diabetic retinopaty: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years", *Archives of Ophthalmology,* **102**, 520–526.

Kleinman, J.C. (1973). "Proportions with extraneous variance: single and independent samples", *Journal of the American Statistical Association*, **68**, 46–54.

Kuchibhatla, M. and Hart, J.D. (1996). "Smoothing-based lack-of-fit tests: variations on a theme", *Journal of Nonparametric Statistics*, **7**, 1–22.

Kupper, L.L. and Haseman, J.K. (1978). "The use of a correlated binomial model for the analysis of certain toxicological experiments", *Biometrics*, **34**, 69–76.

Kupper, L.L., Portier, C., Hogan, M.D. and Yamamoto, E. (1986). "The impact of litter effects on dose-response modeling in teratology", *Biometrics*, **42**, 85–98.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). "Graphical methods for assessing logistic regression models", (with discussion), *Journal of the American Statistical Association*, **79**, 61–83

le Cessie, S. and van Houwelingen, J.C. (1991). "A goodness-of-fit test for binary regression models, based on smoothing methods", *Biometrics*, **47**, 1267–1282.

le Cessie, S. and van Houwelingen, J.C. (1993). "Building logistic models by means of a non parametric goodness of fit test: a case study", *Statistica Neerlandica*, **47**, 97–109.

Ledwina, T. (1994). "Data-driven version of Neyman's smooth test of fit", *Journal of the American Statistical Association*, **89**, 1000–1005.

Lee, G.-H. and Hart, J.D. (1998). "An L_2 error test with order selection and thresholding", *Statistics and Probability Letters*, **39**, 61–72.

Lee, S. (1994). "Optimal choice between parametric and nonparametric bootstrap estimates", *Mathematical Proceedings of the Cambridge Philosophy Society*, **115**, 335–363.

Lehmann, E.L. (1983). *Theory of Point Estimation.* New York: Wiley.

Li, C.-S. (1998a). "Testing lack of fit of heteroscedastic regression models", Unpublished manuscript.

Li, C.-S. (1998b). "Testing for linearity in generalized linear models using kernel smoothing", Unpublished manuscript.

Liang, K.-Y. and Zeger, S.L. (1986). "Longitudinal data analysis using generalized linear models", *Biometrika*, **73**, 13–22.

Liaw, A. (1997). *An application of Fourier series smoothing to a diagnostic test of heteroscedasticity.* Ph.D. dissertation, Department of Statistics, Texas A&M University.

Lindstrom, P., Morrisey, R.E., George, J.D., Price, C.J., Marr, M.C., Kimmel, C.A. and Schwetz, B.A. (1990). "The developmental toxicity of orally administered theophylline in rats and mice", *Fundamental and Applied Toxicology*, **14**, 167–178.

Linton, O.B. (1997), Efficient estimation of additive nonparametric regression models, *Biometrika*, **84**, 469–473.

Linton, O.B. (1998). "Efficient estimation of generalized additive nonparametric regression models", Unpublished manuscript, Yale University.

Linton, O.B. and Härdle, W. (1996). "Estimation of additive regression models with known links", *Biometrika*, **83**, 529–540.

Linton, O. and Nielsen, J.P. (1995). "A kernel method of estimating structured nonparametric regression based on marginal integration", *Biometrika*, **82**, 93–100.

Lockhart, A-M. C., Piegorsch, W.W. and Bishop, J.B. (1992). "Assessing overdispersion and dose-response in the male dominant lethal assay", *Mutation Research*, **272**, 35–58.

Mack, Y.P. and Müller, H.-G. (1989). "Derivative estimation in nonparametric regression with random predictor variable", *Sankyā A*, **51**, 59–72.

MacKinnon, J.G. and Smith, A.A. (1998). "Approximate bias correction in econometrics", *Journal of Econometrics*, **85**, 205–230.

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

Mammen, E. (1993). "Bootstrap and wild bootstrap for high dimensional linear models", *The Annals of Statistics*, **21**, 255–285.

Marx, B.D. and Eilers, P.H.C. (1998). "Direct generalized additive modeling with penalized likelihood", *Journal of Computational Statistics and Data Analysis*, **28**, 193–209.

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

Molenberghs, G. and Ryan, L.M. (1999). "Likelihood inference for clustered multivariate binary data", *Environmetrics*, to appear.

Molenberghs, G., Declerck, L. and Aerts, M. (1998). "Misspecifying the likelihood for clustered binary data", *Computational Statistics and Data Analysis,* **26**, 327-349.

Moore, D.F. and Tsiatis, A. (1991). "Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation", *Biometrics*, **47**, 383–401.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data.* London: Chapman & Hall.

Moulton, L.H. and Zeger, S.L. (1989). "Analyzing repeated measures on generalized linear models via the bootstrap", *Biometrics*, **45**, 381–394.

Moulton, L.H. and Zeger, S.L. (1991). "Bootstrapping generalized linear models", *Computational Statistics and Data Analysis*, **11**, 53–63.

Mukerjee, R. (1990a). "Comparison of tests in the multiparameter case I: second-order power", *Journal of Multivariate Analysis*, **33**, 17–30.

Mukerjee, R. (1990b). "Comparison of tests in the multiparameter case II: a third-order optimality property of Rao's test", *Journal of Multivariate Analysis*, **33**, 31–48.

Müller, H.-G. (1992). "Goodness-of-fit diagnostics for regression models", *Scandinavian Journal of Statistics*, **19**, 157–172.

Müller, H.-G. and Schmitt, T. (1988). "Kernel and probit estimates in quantal bioassay", *Journal of the American Statistical Association*, **83**, 750–759.

Müller, H.-G. and Stadtmüller, U. (1987). "Estimation of heteroscedasticity in regression analysis", *The Annals of Statistics*, **15**, 610–625.

Neumann, M.H. and Polzehl, J. (1998). "Simultaneous bootstrap confidence bands in nonparametric regression", *Journal of Nonparametric Statistics*, **9**, 307–333.

Newton, H.J. (1988). *Timeslab: a time series analysis laboratory.* Belmont CA: Wadsworth & Brooks-Cole.

Neyman, J. (1937)." 'Smooth' test for goodness of fit", *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

Opsomer, J.D. (1995). *Optimal bandwidth selection for fitting an additive model by local polynomial regression.* Ph.D. thesis, Cornell University.

Opsomer, J.D. (1999). "Asymptotic properties of backfitting estimators", *Journal of Multivariate Analysis*, to appear.

Opsomer, J.D. and Ruppert, D. (1997a). "Fitting a bivariate additive model by local polynomial regression", *The Annals of Statistics,* **25**, 186–211.

Opsomer, J.D. and Ruppert, D. (1997b). "A root-n consistent backfitting estimator for semiparametric additive modeling", Unpublished Manuscript, Iowa State University.

Opsomer, J.D. and Ruppert, D. (1998). "A fully automated bandwidth selection method for fitting additive models", *Journal of the American Statistical Association*, **93**, 605–619.

Parzen, E. (1977). "Multiple time series: determining the order of approximating autoregressive schemes", *Multivariate Analysis- IV* (P. Khrishnaiah, ed.), North-Holland, Amsterdam, pp. 293–295.

Pendergast, J.F., Gange, S.J., Newton, M.A., Linstrom, M.J., Palta, M. and Fisher, M.R. (1996). "A survey of methods for analyzing clustered binary response data," *International Statistical Review*, **64**, 89–118.

Phillips, P.C.B. and Park, J.Y. (1988). "On the formulation of Wald tests of nonlinear restrictions", *Econometrica*, **56**, 1065–1083.

Priestley, M.B. and Chao, M.J. (1972). "Non-parametric function fitting", *Journal of the Royal Statistical Society, Series B*, **34**, 385–392.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: Wiley.

Raz, J. (1990). "Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach", *Journal of the American Statistical Association*, **85**, 132–138.

Rodríguez-Campos, M.C. and Cao-Abad, R. (1993). "Nonparametric bootstrap confidence intervals for discrete regression functions", *Journal of Econometrics*, **58**, 207–222.

Rodríguez-Campos, M.C., González-Manteiga, W. and Cao, R. (1998). "Testing the hypothesis of a generalized linear regression model using nonparametric regression estimation", *Journal of Statistical Planning and Inference*, **67**, 99–122.

Rotnitzky, A. and Jewell, N.P. (1990). "Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data", *Biometrika*, **77**, 485–497.

Ruppert, D. and Carroll, R.J. (1997). "Penalized regression splines", Unpublished Manuscript, Cornell University.

Ruppert, D. and Carroll, R.J. (1999). "Spatially-adaptive penalties for spline fitting", Unpublished Manuscript, Cornell University.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). "An effective bandwidth selector for local least squares regression", *Journal of the American Statistical Association*, **90**, 1257–1270.

Ruppert, D. and Wand, M.P. (1994). "Multivariate locally weighted least squares regression", *The Annals of Statistics*, **22**, 1346–1370.

Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1997). "Local polynomial variance function estimation", *Technometrics*, **39**, 262–273.

Sacks, F. and Ylvisaker, S. (1970). "Designs for regression problems with correlated errors III", *Annals of Mathematical Statistics*, **41**, 2057–2074.

Schucany, W.R. and Wang, S. (1991). "One-step bootstrapping for smooth iterative procedures", *Journal of the Royal Statistical Society, Series B*, **53**, 587–596.

Schuster, E. and Yakowitz, S. (1979). "Contribution to the theory of nonparametric regression, with application to system identification", *The Annals of Statistics*, **7**, 139–149.

Schwartz, J. (1994). "Nonparametric smoothing in the analysis of air pollution and respiratory illness", *Canadian Journal of Statistics*, **22**, 471–487.

Schwarz, G. (1978). "Estimating the dimension of a model", *The Annals of Statistics*, **6**, 461–464.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* New York: Wiley.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Severini, T.A. and Staniswalis, J.G. (1991). "Diagnostics for assessing regression models", *Journal of the American Statistical Association*, **86**, 684–692.

Severini, T.A. and Staniswalis, J.G. (1994). "Quasi-likelihood estimation in semiparametric models", *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T.A. and Wong, W.H. (1992). "Profile likelihood and conditionally parametric models", *The Annals of Statistics*, **20**, 1768–1802.

Shao, J. and Tu, D. (1995). *The Jacknife and the Bootstrap.* New York: Springer Verlag.

Shepard, T.H., Mackler, B. and Finch, C.A. (1980). "Reproductive studies in the iron-deficient rat", *Teratology*, **22**, 329–334.

Sherman, M. and le Cessie, S. (1997). "A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models", *Communications in Statistics - Simulations*, **26**, 901–925.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* New York: Springer Verlag.

Simonoff, J.S. and Tsai, C.-L. (1988). "Jackknifing and bootstrapping quasi-likelihood estimators", *Journal of Statistical Computation and Simulation*, **30**, 213–232.

Simonoff, J.S. and Tsai, C.-L. (1999). "Semiparametric and additive model selection using an improved AIC criterion", *Journal of Computational and Graphical Statistics*, **8**, to appear.

Skellam, J.G. (1948). "A probability distribution derived from the binomial distribution by the probability of success as variable between the sets of trials", *Journal of the Royal Statistical Society, Series B,* **10**, 257–261.

Smith, D.M. (1997). *Oswald: Object-oriented Software for the Analysis of Longitudinal Data in S.* World Wide Web page `http://www.maths.lancs.ac.uk/Software/Oswald/`.

Smith, M. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection", *Journal of Econometrics*, **75**, 317–344.

Sommerfeld, V. (1997). "Wild bootstrap versus moment-oriented bootstrap", Unpublished manuscript, Sonderforschungsbereich 373, Humboldt University, Berlin.

Staniswalis, J.G. (1989). "The Kernel estimate of a regression function in likelihood-based models", *Journal of the American Statistical Association*, **84**, 276–283.

Staniswalis, J.G. and Cooper, V. (1988). "Kernel estimates of dose response", *Biometrics*, **44**, 1103–1119.

Staniswalis, J.G. and Lee, J.J. (1998). "Nonparametric regression analysis of longitudinal data", *Journal of the American Statistical Association*, **93**, 1403–1418.

Staniswalis, J.G. and Severini, T.A. (1991). "Diagnostics for assessing regression models", *Journal of the American Statistical Association*, **86**, 684–692.

Staniswalis, J.G., Severini, T.A. and Moschopoulos, P.G. (1993). "On a data based power transformation for reducing skewness", *Journal of Statistical Computation and Simulation*, **46**, 91–100.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997). "Polynomial splines and their tensor products in extended linear modeling", *The Annals of Statistics*, **25**, 1371–1470.

Sutradhar, B.C. and Bartlett, R.F. (1993). "Monte Carlo comparison of Wald's, likelihood ratio and Rao's tests", *Journal of Statistical Computation and Simulation*, **46**, 23–33.

Tibshirani, R. and Hastie, T. (1987). "Local likelihood estimation", *Journal of the American Statistical Association*, **82**, 559–568.

Tutz, G. and Kauermann, G. (1997). "Local estimators in multivariate generalized linear models with varying coefficients", *Computational Statistics*, **12**, 193–208.

Verloove, S.P. and Verwey, R.Y. (1988). *Project on preterm and small-for-gestational age infants in the Netherlands, 1983 (Thesis, University of Leiden),* University Microfilms International, Ann Arbor, Michigan, USA, no. 8807276.

Viraswami, K. and Reid, N. (1996). "Higher-order asymptotics under model misspecification", *The Canadian Journal of Statistics*, **24**, 263–278.

Wand, M.P. (1997). "A comparison of regression spline smoothing procedures", Unpublished Manuscript, Harvard School of Public Health, Boston.

Wand, M.P. (1998). "A central limit theorem for local polynomial backfitting estimators", Unpublished Manuscript, Harvard School of Public Health, Boston.

Wand, M.P. (1999). "On the optimal amount of smoothing in penalized spline regression", *Biometrika*, to appear.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Wedderburn, R.W.M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method", *Biometrika*, **61**, 439–447.

White, H. (1982). "Maximum likelihood estimation of misspecified models", *Econometrica*, **50**, 1–26.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge: Cambridge University Press.

Wild, C.J. and Yee, T.W. (1996). "Additive extensions to generalized estimating equation models", *Journal of the Royal Statistical Society, Series B*, **58**, 711–725.

Woodroofe, M. (1982). "On model selection and the arc-sine laws", *The Annals of Statistics*, **10**, 1182–1194.

Wu, C.F.J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis", (with discussion), *The Annals of Statistics*, **14**, 1261–1295.

Xia, Y. (1998). "Bias-corrected confidence bands in nonparametric regression", *Journal of the Royal Statistical Society, Series B*, **60**, 797–811.

Yanagimoto, T. and 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.

Yee, T.W. and Wild, C.J. (1996). "Vector generalized additive models", *Journal of the Royal Statistical Society, Series B*, **58**, 481–493.

Young, S.G. and Bowman, A.W. (1995). "Non-parametric analysis of covariance", *Biometrics*, **51**, 920-931.

Yuan, K.-H. and Jennrich R.I. (1998) "Asymptotics of estimating equations under natural conditions", *Journal of Multivariate Analysis*, **65**, 245–260.

Zeger, S. and Liang, K.-Y. (1986). "Longitudinal data analysis for discrete and continuous outcomes", *Biometrics*, **42**, 121–130.

Zhao, P.-L. (1994). "Asymptotics of kernel estimators based on local maximum likelihood", *Journal of Nonparametric Statistics*, **4**, 79–90.

Zhao, L.P. and Prentice, R.L. (1990). "Correlated binary regression using a quadratic exponential model", *Biometrika*, **77**, 642–648.

Samenvatting

In de statistische literatuur wordt nog steeds veel aandacht besteed aan het "klassieke" regressie model $Y_i = \theta(x_i) + \varepsilon_i$, waarbij het gemiddelde van de (meestal continue en onafhankelijke) stochastische variabelen ε_i, de storingstermen, gelijk is aan nul. Hoofdzakelijk omwille van de directe interpretatie en de mathematische eenvoud van dit model zijn de statistische eigenschappen hiervan ten gronde bestudeerd. In vele praktische situaties is dit statistisch model echter te eenvoudig en is een meer complexe aanpak aangewezen; dit onder meer wanneer de geobserveerde subjecten tot "natuurlijke groepen" behoren. Bijvoorbeeld, in oogheelkundige studies vormen beide ogen van een persoon een natuurlijke groep. Ook leden van hetzelfde gezin of inwoners van een zelfde geografisch gebied dienen als groepen beschouwd te worden. Men kan verwachten dat groepsleden een gelijkaardig gedrag vertonen. We kunnen onafhankelijkheid tussen de groepen onderstellen, maar binnen een groep zijn de gegevens mogelijk gecorreleerd. Voor dit soort data zullen we een statistisch model gebruiken dat minstens twee parameters bezit. Eén parameter kan de gemiddelde uitkomst zijn en de anderen beschrijven associaties tussen subjecten van dezelfde groep. We zijn voornamelijk geïnteresseerd in gegroepeerde gegevens met een binaire responsvariabele. Verscheidene voorbeelden worden gegeven in Sectie 1.1.

Onze eerste uitdaging is om bestaande *niet-parametrische schattingsmethoden* ("smoothing methods") voor het klassieke regressiemodel uit te breiden naar *likelihood modellen met meerdere parameters* (zie Hoofdstuk 2). De populariteit van niet-parametrische methoden is deels te danken aan diens flexibiliteit; ze laten immers toe om curven te schatten zonder eerst een specifieke parametrische vergelijking voor de vorm van deze curven op te stellen.

In Hoofdstuk 2 worden lokale veeltermschatters in multiparameter "likelihood" (aannemelijkheids-) modellen gedefinieerd. In de lokale schattingsmethode is het

niet nodig dat voor iedere ongekende curve een veeltermbenadering met dezelfde graad wordt gebruikt. We bewijzen de zwakke consistentie en de gezamenlijke asymptotische normaliteit van de schatters voor de ongekende curven en voor diens afgeleiden tot en met de graad van de gekozen veeltermbenadering. De belangrijkste conclusie op basis van de asymptotische uitdrukking voor de vertekening van de schatters is dat de orde van deze vertekening hoofdzakelijk bepaald wordt door de veeltermbenadering met de kleinste graad. Een andere belangrijke vaststelling is dat indien veeltermen met een oneven graad worden gebruikt, de orde van de asymptotische vertekening in de randpunten gelijk is aan deze in inwendige punten (zie Sectie 2.2 voor de exacte definities). Een cross-validatie methode welke gebaseerd is op de likelihood biedt een gepaste data afhankelijke bandbreedte.

In Hoofdstuk 3 wordt verder gebouwd op dit idee. Onder de gepaste regulariteitsvoorwaarden bewijzen we dat de lokale veeltermschatters de eigenschap van sterke consistentie bezitten, en dit in een algemener kader: we onderstellen niet langer dat de likelihood functie correct is of volledig gespecificeerd wordt.

Eén van de voordelen van het gebruik van lokale veeltermschatters is dat de methode, naast schatters voor de ongekende curven, automatisch schatters voor alle afgeleiden van deze curven, tot en met de graad van de gekozen veelterm benadering, oplevert. In de uitdrukking voor de marginale vertekening vinden we echter hogere afgeleiden terug. De vraag die we beantwoorden in Sectie 3.3.3 is hoe deze afgeleiden te schatten. De schattingstechniek die we zullen gebruiken heeft nood aan een tweede bandbreedte, welke met een kleinere snelheid naar nul convergeert dan de bandbreedte gebruikt voor de lokale veeltermschatters zelf (zie bijvoorbeeld Corollarium 3.1). Gebruik makend van de schatters voor de hogere afgeleiden kan de eerder vermelde vertekening geschat worden. Ook een variantieschatter wordt voorgesteld.

In Hoofdstuk 3 zijn we echter vooral geïnteresseerd om met behulp van *bootstrap technieken* een schatter voor de volledige, ongekende verdeling van de lokale veeltermschatter te verkrijgen. Daar de residu-gebaseerde bootstrap niet direct eenduidig te veralgemenen is naar multiparameter likelihood modellen, volgen we een alternatieve weg: we definiëren een één-stap bootstrap methode. Deze heeft als bijkomend voordeel dat er geen extra iteratieve berekeningen vereist zijn. De bootstrap methode kan ook gebruikt worden bij de constructie van simultane betrouwbaarheidsintervallen (zie Sectie 3.5.1).

In Hoofdstuk 4 wordt de schattingstechniek van lokale veeltermen uitgebreid naar multiparameter additieve modellen. De schattingsvergelijkingen gedefinieerd in dit hoofdstuk geven de mogelijkheid om voor iedere curve een andere bandbreedte te gebruiken. Uit de theoretische studie blijkt dat als we verschillende bandbreedtes willen gebruiken voor het schatten van de verschillende curven dit niet altijd op de meest optimale wijze kan. Zowel in een niet-parametrisch additief model als in een multiparameter likelihood model met slechts één verklarende variabele wordt de asymptotische vertekening van ieder van deze schatters niet enkel bepaald door de specifieke curve die men schat, maar ook door alle andere curven die in het statistisch model aanwezig zijn. Voor multiparameter modellen is er geen probleem wanneer de curven orthogonaal zijn in de betekenis dat de niet-diagonale elementen van de Fisher informatie matrix gelijk zijn aan nul. In het algemeen kan voor deze statistische modellen een oplossing gevonden worden door middel van backfitting; een methode die echter niet optimaal werkt in de additieve modellen waar ze oorspronkelijk voor gedefinieerd werden.

Een uitbreiding naar semi-parametrische modellen geeft de mogelijkheid om bepaalde verklarende variabelen op een parametrische manier te modelleren (bijvoorbeeld de categorische variabelen) en anderen op een niet-parametrische manier, hetgeen resulteert in zeer flexibele statistische modellen.

Niet-parametrische schatters in deze modellen zijn uiteraard niet beperkt tot de lokale veeltermschatters. In Hoofdstuk 5 bestuderen we gepenalizeerde regressie *splines*. In deze techniek wordt een stel knooppunten ("knots") gespecificeerd. Tussen ieder van deze punten vervangt men de te schatten functie door een veelterm. Deze veeltermen worden op een "gladde" manier met elkaar verbonden in de verbindingspunten. Door een gepaste keuze van deze punten worden schatters voor de ongekende curve verkregen. Deze methode is niet-parametrisch omdat het volledige functioneel verband tussen de verklarende variabelen en de responsvariabele niet globaal gemodelleerd wordt, het lokale karakter van de methode zit bevat in de specificatie van de knooppunten. De resulterende schatter zal echter afhankelijk zijn van de plaats en van het aantal van deze punten. Om de invloed van deze afhankelijkheid te verminderen, wordt er een penalizatie constante toegevoegd. In Sectie 5.3.1 onderzoeken we onder andere de asymptotische gemiddelde kwadratische fout van de resulterende schatters en stellen we een methode voor om de penalizatie constante te selecteren.

In de hierop volgende hoofdstukken ligt de klemtoon op het toetsen van de hypothese dat een functie een welbepaalde parametrische vorm heeft. Deze functie kan één van de parameters in een regressie model zijn, bijvoorbeeld het gemiddelde van de responsvariabele, of de correlatie tussen uitkomsten van subjecten in dezelfde groep, ... (*lack-of-fit toetsen*), of het kan een volledige dichtheidsfunctie zijn waarvoor we een aanpassingstoets willen uitvoeren (*goodness-of-fit toetsen*). Eerst construeren we toetsen gebaseerd op niet-parametrische methoden welke consistent zijn voor essentieel alle afwijkingen van het parametrische model gespecificeerd onder de nulhypothese. De basis van deze toetsingsgrootheden is het weergeven van het verschil tussen de functie onder de nulhypothese en de "echte" curve door middel van een orthogonale reeks. Dit verschil wordt geschat door een eindig aantal termen van deze reeks te nemen, waarbij het nemen van nul termen overeenkomt met het specificeren van de functie onder de nulhypothese. Het principe van orde-selectie toetsen is om de selectie van dit aantal termen in de reeks te laten leiden door de data zelf. De nulhypothese kan verworpen worden als het door de data bepaalde aantal termen in de reeks groter of gelijk is aan één.

In Sectie 6.2 stellen we twee criteria voor om dit aantal termen te bepalen; beide zijn geldig in volledig gespecificeerde likelihood modellen. Een eerste methode is een aangepaste versie van het Akaike AIC criterium en maakt rechtstreeks gebruik van de likelihood functie. Een tweede criterium is gebaseerd op score toetsingsgrootheden. De asymptotische verdelingstheorie van de resulterende toetsen wordt gegeven in Secties 6.2.4 en 6.2.5. In geval de likelihood functie niet volledig of niet correct gespecificeerd wordt, stellen we een ander criterium voor, gebaseerd op de zogenaamde robuuste score toetsingsgrootheden. Dit laat toe om de lack-of-fit toetsen ook te gebruiken in de context van veralgemeende schattingsvergelijkingen (generalized estimating equations, GEE). Verscheidene toepassingen en resultaten van een simulatiestudie zijn gegeven in de tekst.

De penalizatie constante in de selectiecriteria is niet eenduidig bepaald. Verschillende mogelijkheden worden voorgesteld en bestudeerd in Sectie 7.2. Een vergelijking van het onderscheidingsvermogen van de resulterende toetsen levert de volgende conclusies. De orde-selectie test, met constante 4.18, is een zeer goede keuze als het model onder de alternatieve hypothese een lage orde heeft (zie Sectie 7.3 voor de exacte beschrijving van de simulatiestudie). Het onderscheidingsvermogen daalt echter snel als de orde van het alternatieve model toeneemt. De toets gebaseerd op

BIC, met constante $\log n$ (waarbij n de steekproefgrootte), toont ongeveer hetzelfde gedrag. Vooral bij hoge frequenties is de penalizatie erg groot. We onderzochten ook het gedrag van drie AIC gebaseerde toetsen (met constante 2). De conclusie hier is dat de gestandaardiseerde versie (vergelijking (7.5)) en de toetsingsgrootheid welke gedefinieerd is als de maximale waarde van het AIC criterium (vergelijking (7.7)) beide beter zijn dan de niet gestandaardiseerde versie, maar wanneer de orde van het model onder de alternatieve hypothese minstens gelijk is aan vier, dan is de niet gestandaardiseerde AIC toets de beste keuze.

Waar de voorgaande toetsen nog steeds gebaseerd zijn op regressie modellen met één verklarende variabele (of op univariate dichtheden), zullen we in Sectie 7.4 de techniek uitbreiden naar meerdimensionale situaties. Het probleem nu is het definiëren van een stijgende rij alternatieve modellen ("nested model sequence"). De keuze van zo een rij zal de aard van de toets bepalen en laat toe om specifieke toetsen te construeren, zoals bijvoorbeeld een toets voor additiviteit.

In Hoofdstuk 8 bestuderen we het toetsen van hypothesen waar de nulhypothese gecontrasteerd wordt met een heel specifieke alternatieve hypothese. We zijn geïnteresseerd in het gebruik van bootstrap methoden voor de benadering van P-waarden van likelihood ratio, Wald en score toetsingsgrootheden. De belangrijkste moeilijkheid is het genereren van bootstrap gegevens welke de nulhypothese weerspiegelen. Indien men onderstelt dat het datagenererende kansmodel gekend is, kan dit gebeuren door middel van een parametrische bootstrap methode. Hierbij worden ongekende parameters in het model onder de nulhypothese geschat; de geschatte verdeling wordt vervolgens gebruikt voor het genereren van nieuwe gegevens. In het bijzonder vestigen we de aandacht op de pseudo-likelihood schattingsvergelijkingen. We tonen theoretisch aan dat de parametrische bootstrap kan gebruikt worden als een alternatief voor de klassieke asymptotische chi-kwadraat verdeling van de hoger vermelde toetsingsgrootheden. Simulaties tonen dat het gebruik van de bootstrap vooral voordelen biedt voor de Wald toets welke, in de simulatiesetting zoals beschreven in Sectie 8.5.1, in verscheidene gevallen een veel te grote gesimuleerde type I fout vertoont.

Een belangrijk aspect bij het toepassen van een parametrische bootstrap methode is de noodzaak om de likelihood volledig en correct te specificeren. Bij een niet correcte specificatie zullen er data gegenereerd worden uit een verkeerde verdeling, hetgeen in de meeste gevallen foute resultaten zal opleveren. Een ander nadeel is

dat deze techniek erg rekenintensief is, de berekening van de schatters in algemene likelihood modellen dient immers op een iteratieve manier te gebeuren. Dit is de motivatie om in Hoofdstuk 9 een één-stap semi-parametrische bootstrap procedure te bestuderen. Om een weerspiegeling van het statistisch model onder de nulhypothese mogelijk te maken, worden de bootstrap schatters geconstrueerd met behulp van de schatters onder de nulhypothese (dit is het parametrische aspect van de methode), en door middel van trekken met terug leggen uit de collectie van score functies en afgeleiden van score functies, gebaseerd op de oorspronkelijke data (dit is het niet-parametrische aspect). Zowel een één-stap lineaire als een één-stap kwadratische bootstrap methode worden bestudeerd, met als toepassing de constructie van gerobustificeerde bootstrap Wald en score toetsingsgrootheden. Eén van de conclusies van de simulatiestudie in Sectie 9.4.2 is dat de veel te grote gesimuleerde type I fout van de robuuste Wald toets goed 'gecorrigeerd' wordt door de één-stap kwadratische bootstrap, maar niet altijd door de één-stap lineaire bootstrap.

In Hoofdstuk 10 bestuderen we het gebruik van de één-stap kwadratische bootstrap voor de constructie van schatters met een correctie voor de eindige steekproef vertekening, alsook voor de constructie van betrouwbaarheidsintervallen. Het blijkt immers dat de één-stap lineaire bootstrap niet geschikt is om te corrigeren voor de vertekening en ook problemen heeft bij de constructie van betrouwbaarheidsintervallen. Deze laatsten hebben wel een kleinere lengte (in vergelijking met de intervallen gebaseerd op asymptotische verdelingstheorie), maar de gesimuleerde betrouwbaarheid is eveneens kleiner. De één-stap kwadratische bootstrap daarentegen, levert ook kortere betrouwbaarheidsintervallen, maar met grotere betrouwbaarheid dan de overeenkomstige intervallen gebaseerd op de normale verdeling.

This is not the end
This is just a new beginning