Lack-of-fit tests and plots based on regional residuals

Ellen Deschepper

Supervisors: Prof. dr. ir. Olivier Thas

Prof. dr. Jean-Pierre Ottoy

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (Ph.D.) in Sciences, Mathematics

Academic year 2006–2007

Dankwoord

Met het schrijven van deze woorden van dank, kan ik eindelijk een lange periode van predoctoraal werk afsluiten. Met plezier schrijf ik dan ook dit meest gelezen stukje uit mijn doctoraat!

Eerst en vooral wil ik mij graag richten tot mijn promotoren, Prof. Thas en Prof. Ottoy, die mij in de eerste plaats de mogelijkheid boden om te doctoreren. Bedankt voor de zinvolle wetenschappelijke discussies en de mogelijkheid om internationale contacten te leggen op diverse wetenschappelijke congressen! Olivier, bedankt voor de nodige peptalk en de aangename momenten van ontspanning! Ik hoop dat we in de toekomst nog vele terrasjes mogen uittesten.

I would like to thank the members of the jury for their comments and efforts. In particular, Prof. Hart and Prof. Le Cessie thank you for being here. It is a great honour to have you as a member of my examination committee. Prof. Hart, thank you for your careful reading and useful suggestions!

Roos, bedankt voor alle logistieke hulp, je geduld en je bereidwilligheid om steeds in te springen wanneer de nood het hoogst was! Jelle en Koen, bedankt voor het gebruik van alle LaTeXtemplates! Lieven en Heidi, jullie hielpen zo een beetje op diverse vlakken! Ook aan alle andere collega's van de vakgroep die op één of andere manier, tot de kleinste bijdrage toe, hebben meegeholpen om dit werk tot een goed einde te brengen: bedankt! De grootste bijdrage van jullie allemaal is zeker en vast het creëren van een aangename werksfeer. Jullie staan immers ook garant voor de nodige ontspanning op en naast de werkvloer!

Maar de belangrijkste bron van morele steun komt natuurlijk van vrienden en familie! Een welgemeende dankjewel voor al jullie begrip, steun en geduld. In het bijzonder denk ik aan al wie bereid was om op onze kleine schat te passen wanneer ik alweer maar eens een weekend of een avond moest doorwerken. Mams en paps, jullie steun gaat natuurlijk veel verder dan alleen maar de afgelopen jaren. Bedankt om over de jaren heen, steeds de gepaste omgeving te creëren waardoor ik me door al mijn studies hebben kunnen slaan.

Met grote vreugde richt ik mij tenslotte tot diegene die me het meest dierbaar

zijn.
Bob, je bent een fantastische papa, bob de bouwer, en wetenschapper, maar voor mij ook zeker en vast een ondersteunende, liefde- en begripvolle wederhelft!
Rune, mijn kleine spruit, bedankt om me alles te doen relativeren en me te laten genieten van de kleine dingen in het leven! Je lach is goud waard!
En jij, klein stampertje in mijn buik, jij bent meer dan welkom in het leven na het doctoraat!
Aan jullie drie draag ik dan ook dit werk op!

Gent, Juni 2007
Ellen Deschepper

List of symbols and abbreviations

i.i.d.	identically and independently distributed
GLM	Generalized Linear Model
GOF	Goodness-of-Fit
LOF	Lack-of-Fit
σ^2	population variance
$\hat{\sigma}^2$	some consistent estimator of σ^2
S_n^2	sample variance estimator based on a sample of size n
$\hat{\sigma}_u^2$	unbiased sample variance estimator
$\hat{\sigma}_b^2$	biased sample variance estimator
$\hat{\sigma}_D^2$	first difference variance estimator
$\hat{\sigma}_M^2$	mean squares error variance estimator
$\hat{\sigma}_P^2$	pseudoresidual based error variance estimator
RR	regional residual test
RRS	regional residual test based on intervals, using the sample variance S_n^2
RRD	regional residual test based on intervals, using the first difference variance estimator $\hat{\sigma}_D^2$
RRP	regional residual test based on intervals, using the pseudoresidual based error variance estimator S_P^2
RRK	regional residual test based on intervals, with known variance σ^2
RRUn	unstandardized regional residual test with factor $1/\sqrt{n}$
RRUnij	unstandardized regional residual test with factor $1/\sqrt{n_{ij}}$
RRC	regional residual test based on arcs, using the sample variance S_n^2
RRGL	regional residual test based on marginal test statistics, using the sample variance S_n^2
SRRS	regional residual test based on spherical regional residuals, using the sample variance S_n^2
RRLR	regional residual test based on raw residuals in the logistic regression context
RRLD	regional residual test based on standardized deviance residuals in the logistic regression context
RRLP	regional residual test based on standardized Pearson residuals in the logistic regression context
RRL.	the collection of RRLR, RRLD and RRLP tests
sd(.)	standard deviation of (.)

Contents

CHAPTER 1

Introduction

Regression analysis is a very widely used statistical method, and its applications can be found in many different fields, like in biological or psychological experiments, medical science, business and finance, and many others. In regression analysis the effect of one or more predictor variables or covariates $\mathbf{x}$, like e.g. the geographical position of a city, on a response variable y, like e.g. the temperature, is investigated by means of a statistical model,

$$y = m(\mathbf{x}) + \epsilon. \tag{1.1}$$

The model consists of a systematic component $m(\mathbf{x})$ and an error component ϵ. The mean of the response variable is modeled conditional on the observed value of the predictor variables, i.e. $E(y\,|\mathbf{x}) = m(\mathbf{x})$. This function thus characterizes the average value of y among subjects with the same covariate values or covariate pattern $\mathbf{x}$. The response values y are assumed to be conditionally independent. The error component on the other hand, represents how much the value of the response variable y differs from the average value among subjects with the same covariate pattern $\mathbf{x}$. In practice, the true relationship (1.1) is rarely known. If the functional form of the systematic component and the distributional form of the error component are known, parameters can be estimated based on a sample of size n, e.g. by means of the least squares method. When the relationship between the mean response and a single covariate $x \in \mathbb{R}$ is known to be linear, the simplest form for $m(x)$ is $m(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x$. It belongs to a parametric family of regression functions, $\mathcal{M} = \left\{m(x, \boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_0, \theta_1) \in \Theta \subset \mathbb{R}^2\right\}$. If, in addition, the distribution of the random error terms is specified, e.g. the error terms are independently and identically normally distributed with zero mean and variance σ^2, then the model $y = \theta_0 + \theta_1 x + \epsilon$ is referred to as a fully parametric model. On the other hand, a regression model may be specified by infinitely many regression parameters, and then the model is called nonparametric. To illustrate the idea of a nonparametric regression model, assume equally spaced design points $x_i = (i - 0.5)/n,\ i = 1, \ldots, n$ of a single covariate $x \in [0, 1]$, and, let $m(x)$

be a square integrable function, that has the Fourier series representation,

$$m(x, \boldsymbol{\phi}) = \phi_0 + 2\sum_{j=1}^{\infty} \phi_j \cos(\pi j x), \quad \text{for almost all } x \text{ in } [0,1], \tag{1.2}$$

with Fourier coefficients

$$\phi_j = \int_0^1 m(x)\cos(\pi j x)dx, \quad j = 1, 2, \ldots. \tag{1.3}$$

Note that $m(x)$ contains an infinite dimensional vector of parameters $\boldsymbol{\phi}$ and is now thus said to be a nonparametric regression model. The previous representation is only one possible illustration of a nonparametric regression model, out of the wide class of nonparametric regression models. Of course in practice, to fit a nonparametric regression model to the data, the number of parameters has to be finite, and the model has to be approximated. The term 'nonparametric'rather refers to the enormous variety of functions that can be approximated by a nonparametric model without specifying a particular form in advance. Despite their flexibility, data analysts still often prefer parametric over nonparametric regression models. The popularity of the parametric models is perhaps due to the fact that the regression coefficients are more easily interpretable than those involved in nonparametric regression. Parametric models are also more familiar to practitioners, and are very easily fitted by statistical software packages. Estimation in nonparametric models, on the other hand, always requires a subjective or data-driven choice of a smoothing parameter.

Once the systematic component of the regression model is specified, the data-analyst can fit the model to the data, and is further interested in inferences on regression parameters and prediction. Before doing so, it is wise to perform a model check to verify whether the specified regression model is appropriate for the data at hand. Inferences and predictions can be wrong when the specified parametric model is not appropriate. The discussion in this dissertation will deal with the assessment of the fit of a parametric model, with which the data-analyst is preliminarily satisfied. This means that to the best of his knowledge, all relevant predictor variables are present in the model and have been entered in the correct functional form (like $\log(x)$, or inclusion of interaction terms). To validate the quality of the specified regression model, distance measures between observed, y, and fitted values, $\hat{y}$, should be examined both individually and collectively. If the model fits well, we expect global summary measures to be small and individual contributions of each pair $(y_i, \hat{y}_i)$, $i = 1, \ldots, n$ to be unsystematic and relatively small compared to

the error component. The particular field of statistics that is concerned with assessing the fit of parametric regression models is known as Lack-of-Fit (LOF). This will be the main focus of this dissertation. In the literature, there is not always a clear distinction between Goodness-of-Fit (GOF) and LOF tests and the two terms are sometimes mixed up. We would like GOF tests to correspond to the null hypothesis stating that a given sample has arisen from a specified distribution, while LOF tests are used to check whether a certain family of parametric regression models appropriately describes the relationship between the mean of a response variable y and one or more predictor variables $\mathbf{x}$.

The simplest distance measure between observed, and fitted values, is their difference, $y - \hat{y}$, called residual, which provides an estimate of the error component. Residuals or transformations of them are highly informative to assess the fit of the specified model. Large values of properly standardized residuals may indicate individually poorly fitting observations. When plotted with respect to the fitted values, or with respect to included and/or omitted covariates, they may

- visually show poorly fitting individual observations,
- allow the assessment of model assumptions like equal error variances or homoscedasticity,
- even suggest possible ameliorations to the specified model, as they may reveal effects of potential new covariates or suitable transformations of predictor variables already included in the model.

Residuals or transformations of them may be combined into a single overall LOF test statistic as a global measure of model quality. The fact that a LOF test is a single value to summarize a considerable amount of information is both an advantage as well as a disadvantage. Therefore, in any analysis, the use of a LOF test should be complemented with a careful examination of some individual measures or regression diagnostics. Diagnostics focus on individual observations and their influence on regression parameters and predictions. When, for example, in linear regression the values of a certain subject are *far* from the average predictor value, the design point is said to have a high leverage. Small perturbations of the response value of a high leverage point, may have considerable influence on the regression parameter estimates, predictions and inferences. Therefore, both LOF tests and regression diagnostics should be considered before any conclusion concerning the model fit is drawn. Without denying the importance of individual regression diagnostics, we only investigate the use of LOF tests in this dissertation.

A huge number of statistical tests in the literature are already available for this purpose. Chapter 3 provides a selective overview of lack-of-fit tests. We have no intention to give a complete overview, but we rather focus on those tests that are of historical importance, that are widely used in practice or that are related to our newly introduced tests in later chapters. First of all, there is the well known classical F test (Fisher, 1922) which is based on the pure sum of squared errors. This widely described and important test is unfortunately only applicable when multiple observations with the same covariate patterns are present in the data. This is however a severe limitation in practice. A related test, called the reduction method, does not require replicates of observations with the same covariate pattern, but requires the specification of an alternative model in advance. This model is often not available, as we believe that the model under study is appropriate to describe the relationship between the mean response and the covariates. In addition, we would like to test the appropriateness of the model under study to prevent us from misleading or incorrect inferences, without having a particular alternative model in mind. Nonparametric LOF tests answer this need. Two historically important nonparametric tests, the von Neumann (1941) and Buckley (1991) tests are discussed. Further, smoothing based LOF tests, nicely presented and summarized in the monograph of Hart (1997), form an important part of this section. In general, these tests are very powerful to a wide class of alternative models, but, unfortunately, their performance depends on the subjective choice of a smoothing parameter or type of smoother. However, data-driven selection criteria for smoothing parameters are available nowadays. Further, in case of multiple covariates, the performance highly depends on the order relation that has to be chosen for the residuals before they can be smoothed, unless tests are based on multivariate smoothers. Finally, LOF tests based on marked empirical processes (e.g. Stute (1997), Diebolt and Zuber (1999), Lin et al. (2002)) are introduced, as they are closely related to our tests described in Chapter 4 and later chapters. LOF tests for logistic regression models are introduced in a separate section, as the non-unique definition of the residuals have important consequences on the construction of tests (Hosmer and Lemeshow, 2000). As already briefly highlighted, graphical diagnostic displays are also very useful for detecting and examining anomalous features in the fit of a model to data. Among the graphical diagnostic tools, the classical residual plot is probably the best known. Only a few authors also suggest plots directly related to LOF tests. A brief overview is also provided. To conclude Chapter 3, we provide an overview of bootstrap schemes that can be used to approximate the null distribution of the test statistics described in this review chapter. Many asymptotic null distributions of test statistics are not suitable for use in small

samples. Further, some null distributions might be too complex, and therefore the bootstrap is an alternative that is preferred by many authors, e.g. Hart (1997), Stute et al. (1998), Fan and Huang (2001), among others. Also for our new tests, we prefer to apply one of the bootstrap schemes in this section.

One could wonder why to propose more LOF tests when there are already so many tests available. In the literature, the majority of tests focus on detecting *global* deviations from the null model. In other words, when the model is not appropriate, often the deviations occur over the entire range of the covariates. Tests are constructed to detect these global deviations by combining discrepancy measures of individual observations into an overall LOF test. Failure to assess *local* deviations in a particular region of the predictor space is a major problem of most tests. By *local* LOF, we thus mean the presence of small areas in the predictor space, where the regression model does not fit well locally.
The goal of this dissertation is to construct LOF tests that are able to detect both global *and* local deviations from a parametric regression model. In addition, the tests should not depend on the subjective choice of a smoothing parameter. They have to be applicable whether replicated observations are present or not. Further, we aim to construct a formal diagnostic plot that is directly associated with the LOF test. The plot should formally identify regions in the predictor space where the regression model does not fit well.
We therefore propose to consider discrepancy measures over both local and global regions in the predictor space, and to combine this information in one test statistic. The new test is able to detect both global *and* local deviations. It is independent of a subjective choice of a smoothing parameter and is applicable whether replicated observations are present or not.
In addition, we believe that we could do better than solely reducing information of an entire sample to a single test statistic. Therefore, we propose to use the information that is available in the individual discrepancy measures, calculated over both local and global regions in the predictor space, to construct plots. In this way, the plots are clearly associated to the new tests, and provide a better insight into the underlying deviations present in a certain model fit. Moreover, the graphs allow formal conclusions, which is a major advantage over other diagnostic plots directly associated to LOF tests. Possible local deviations in the latter may be detected by the human eye, but the data analyst does not know whether the observed discrepancies are statistically significant. When LOF is detected by the new statistical test, the corresponding plot allows the data-analyst to locate specific regions in the predictor space where the model does not fit well and suggest in which area remedial measures may be necessary. The new plots help the data-analyst to formally identify regions in

the predictor space that deserve special attention of the experimenter.
Chapter 4 describes these new tests and plots in case of a single predictor variable on the real line. Chapter 6 discusses extensions of the new tests to multiple covariates, while in the literature often only the univariate case is considered. The use of the new tests in generalized linear models has to be studied in more detail as the residuals are not uniquely defined. This is done in Chapter 7. To answer the need of a rarely discussed problem, the methodology is applicable in the context of circular-linear regression models. When one of the predictor variables is measured on a circular, rather than on a linear scale, many classical LOF tests are no longer applicable as a specific origin for the circular variate has to be chosen. As our methodology is origin-independent, it is straightforward to consider *LOF on a circle*. All chapters include specific data examples on which the tests and plots are illustrated. They all contain a simulation study for comparing the performance of the new methods with some classical tests, in the specific context of each chapter. As some general guidelines for all simulation studies, we mention that all tests in this dissertation are performed at the 5% significance level. Typically, a larger number of Monte Carlo and bootstrap samples than the ones used in this thesis are necessary to accurately estimate the empirical powers. We believe, however, that our results are indicative of the comparison between the different tests. In particular, we obtain good empirical powers for our tests in case of local deviations from the hypothesized model.
In the last chapter (Chapter 8), some large sample properties of the newly introduced tests are discussed. Although we argue that the bootstrap is more suitable for practical use, we provide some theoretical basis for our proposed tests.

Before starting the discussion of LOF tests, we first introduce and discuss a case study from the literature in Chapter 2 to fully appreciate the underlying problem and goal.

CHAPTER 2

The US temperature data set

The US Temperatures data (Peixoto, 1990) gives the normal average January minimum temperature, y, in degrees Fahrenheit, between 1931 and 1960, with the longitude, x_1, and latitude, x_2, of 56 United States (US) cities. The longitude of the US cities is measured in degrees west of the prime meridian and the latitude in degrees north of the equator. The geographical positions of the 56 US cities are illustrated in the map and scatter plot in Figure 2.1. Note that the longitude is plotted in negative values to obtain the conventional map of the US.

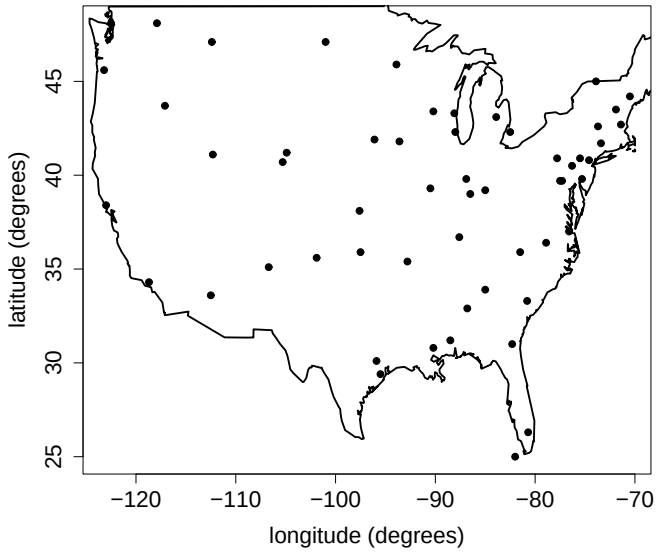

FIGURE 2.1: *US Temperature data (Peixoto, 1990), map and scatter plot that represents the geographical positions of the 56 US cities. The latitude is measured in degrees north of the equator and the longitude in degrees west of the prime meridian. The longitude is thus plotted with negative values to obtain the conventional map of the US.*

TABLE 2.1: *Estimated parameters and their standard errors when the linear model (2.1) is fit to the US temperature data, together with the calculated values of the statistical tests for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ and their corresponding p-values.*

Coefficients	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	26.51786	0.92667	28.616	< 0.0001
Longitude	0.13396	0.06314	2.122	0.0386
Latitude	-2.16355	0.17570	-12.314	< 0.0001

The data file is available online from the Data and Stories Library (DASL) at `http://lib.stat.cmu.edu/DASL/Stories/USTemperatures.html`.
The simplest model of the average minimum temperature as a function of the longitude and the latitude is the linear model,

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon. \tag{2.1}$$

Assuming normally distributed error terms with constant variances σ^2, we fit this model to the data using least squares. Note that the predictor variables were centered before the model was fit, so as to reduce multicollinearity. We summarize some results in Table 2.1, from which the fitted parametric model can be written as

$$\hat{y} = 26.518 + 0.134 x_1 - 2.164 x_2. \tag{2.2}$$

Figure 2.2 shows the observed data with the fitted linear regression plane. It is very hard to visually find out whether the fitted model is appropriate for the data, because the representation needs to be done in three dimensions. Before valid inferences could be drawn from the model, or before the model could be used for predictions, we need to be sure that no systematic deviations from the fitted model are present. In other words, we need to check whether no lack-of-fit is present. In what follows, we discuss the results of the regression analysis and some traditional graphics to evaluate the quality of the fitted model.
In Table 2.1, both predictor variables, longitude and latitude, show significant linear relationships at the 5 % significance level: $p = 0.0386$ for testing $H_0 : \theta_1 = 0$ versus $H_1 : \theta_1 \neq 0$, and $p \leq 0.0001$ for testing $H_0 : \theta_2 = 0$ versus $H_1 : \theta_2 \neq 0$. It actually tests whether a model with this specific predictor tells us more about the outcome variable, than a model that does not include that variable. The F test for a regression relationship (F-statistic = 75.88 and corresponding p-value ≤ 0.0001) indicates the existence of a regression relationship between

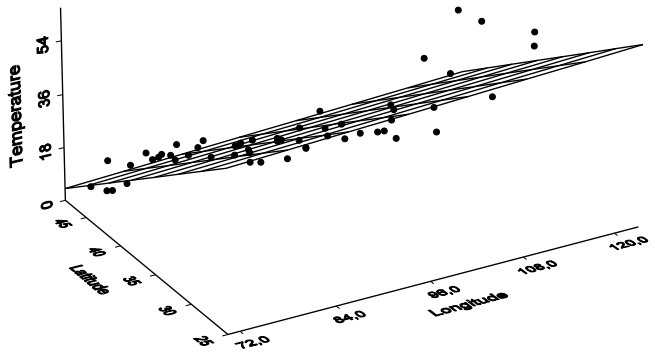

FIGURE 2.2: *Observed (data) and fitted linear regression model (plane) of the US temperature data (Peixoto, 1990).*

the average January minimum temperature and longitude and latitude, but of course it does not ensure that useful predictions can be made by using it. The adjusted R-squared, which represents the percentage of the total variability of the response variable that is explained by the regression model, is 0.73, which is neither very good nor bad. The predicted values are a more or less accurate representation of the observed values.

We plot several graphs to check the model assumptions and to validate the model quality. The QQ plot in the upper left panel of Figure 2.3 does not indicate severe deviations from normality for the residuals. The scatter plot of the observed temperature versus the fitted values is shown in the lower left panel of this figure. The dotted line represents the bisector, and the smoothed trend line (full line) is a loess smoother with span = 0.75. When good model predictions are available, we expect to see a narrow cloud around the bisector, which indicate a good correspondence between the observed and the fitted response variable. The smoothed line will then more or less coincide with the bisector. This graph shows a rather good correspondence when fitting the linear model. Although the cloud is not that narrow, most points are scattered nicely around the bisector (dotted line) and the smoothed line does not deviate considerably from the bisector.

Further, the left panels of Figure 2.4 show the residual plots against the fitted values (upper panel), against latitude (middle panel) and against longitude

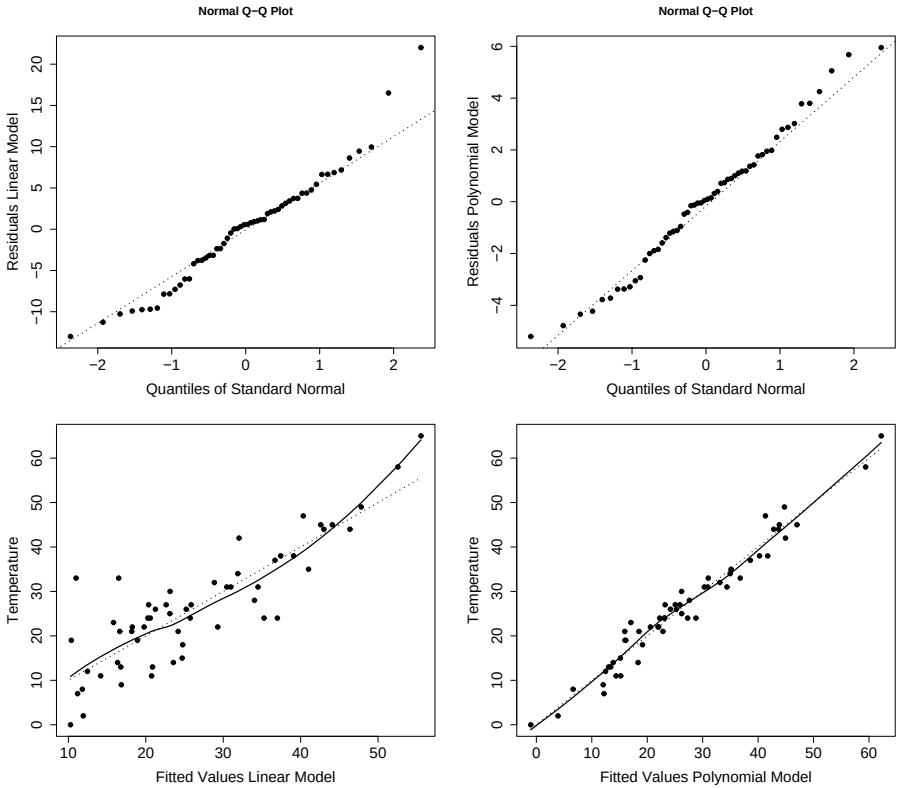

FIGURE 2.3: *(Upper panels) QQ-plot of the residuals; (Lower panels) Scatter plots of the observed temperature values versus the fitted values. The dotted line represents the bisector, and the smoothed trend line (full line) is a loess smoother with span = 0.75. The plots in the left panels are those for the linear model (2.2), while in the right panels are those for the third order polynomial model (2.4) are plotted.*

(lower panel) for the linear model fit. When no model deviations are present, we expect the residuals to have zero mean and constant variance. In these scatter plots, this can be translated into a random cloud of residuals around zero, equally wide for small and large values of the variable on the x-axis. We expect then the solid line, a loess smoother with span = 0.75, to coincide with the constant line at zero (dotted line). From these plots, we conclude that the assumption of homoscedasticity or constant variances is reasonable.

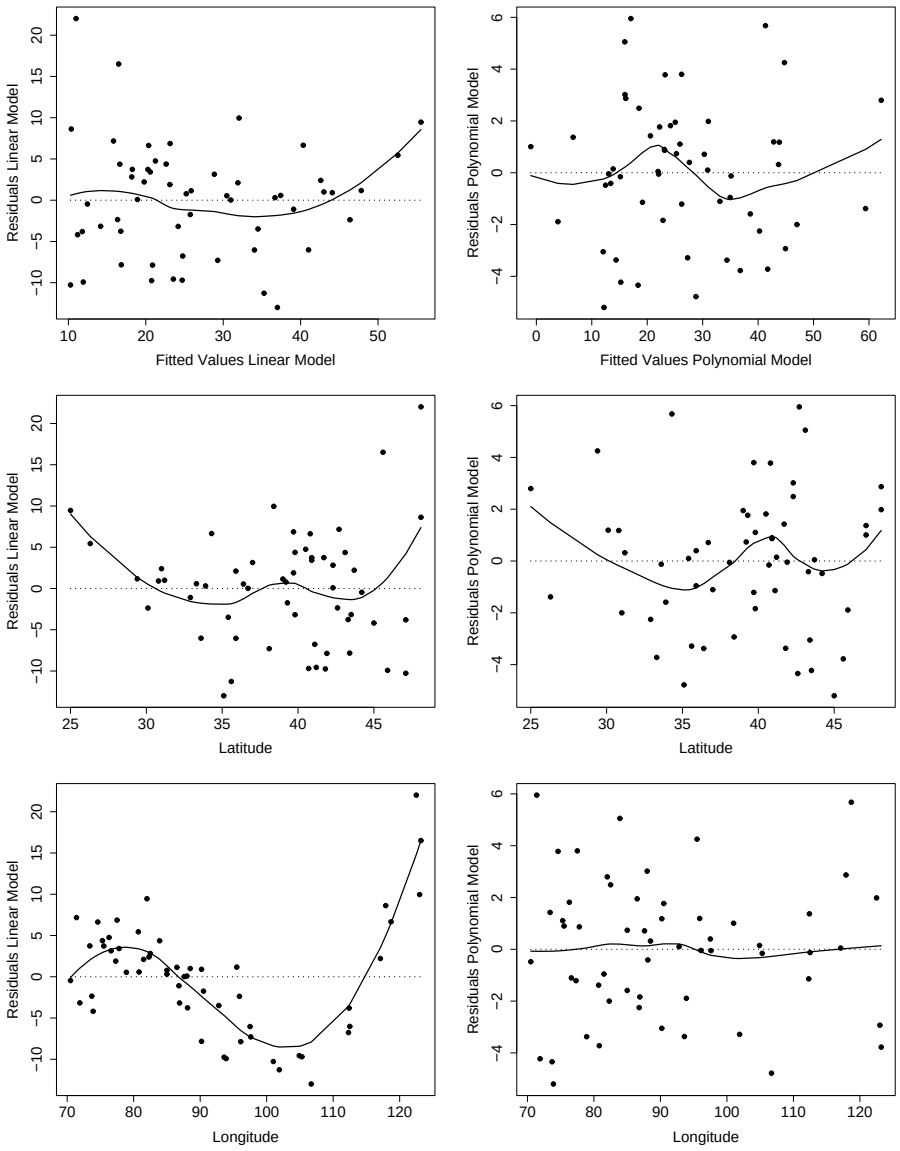

FIGURE 2.4: *Scatter plots of the residuals versus the fitted values (upper panels), versus latitude (middle panels) versus longitude (lower panels). The full line corresponds to a smoothed trend line (loess smoother, span = 0.75), the dotted line is the constant mean model at zero. The plots in the left panels are those for the linear model (2.2), while in the right panels are those for the third order polynomial model (2.4).*

The smoothed trend line for the residuals versus longitude, however, shows a clear trend. It suggests that some regression model with higher order polynomials might be more appropriate. For the fitted values and latitude, the scatter plots are less clear. We also have to be aware of possible boundary effects that drag the smoothed trend line unjustly towards one direction at the boundaries. Partial regression plots might help to reveal the relationship between average January minimum temperature and latitude and longitude. The left panel in Figure 2.5 reveals that the relationship between January temperature and latitude, after removing the effects of longitude, is linear and negative. However, after removing the effects of latitude, the relationship between January temperature and longitude is cubic polynomial (right panel). Of course, these plots do not involve a statistical test, so no formal conclusions can be drawn from them. Although we did not perform a statistical lack-of-fit test yet, we could already suspect that the linear model is inappropriate to describe the relationship between the normal average January minimum temperature, and latitude and longitude. In Chapters 3 and 6 we will confirm this suspicion with some statistical tests.

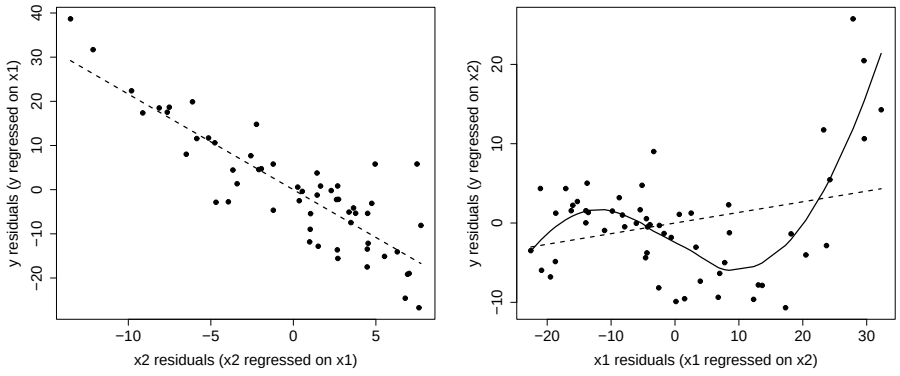

FIGURE 2.5: *Partial residual plots for latitude (left panel) and longitude (right panel). The dashed line is the linear least squares fit, the full line is a loess smooth fit (span = 0.75).*

According to Peixoto (1990), a cubic polynomial model in longitude and first order in latitude is a more appropriate model for the US temperatures data than the first order polynomial in both predictors.
For the parametric regression model,

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_1^2 x_2 + \theta_6 x_1^3 + \theta_7 x_1^3 x_2 + \epsilon, \qquad (2.3)$$

we obtain the least squares fit

$$\hat{y} = 22.83 - 0.518x_1 - 2.522x_2 - 0.0008x_1x_2 + 0.0042x_1^2 + 0.0003x_1^2x_2 + 0.001x_1^3 + 0.00003x_1^3x_2. \quad (2.4)$$

Previous measures and plots are reconsidered for the new model. We discuss some remarkable changes. Firstly, the adjusted R-squared seriously improved to 0.95 and this also results in a more narrow scatter around the diagonal in the scatter plot of the observed temperature values versus the fitted values (Figure 2.3, lower right panel). The residual scatter plots versus the fitted values, latitude and longitude (Figure 2.4, right panels) show a remarkable improvement in the trend for longitude and a huge reduction in the range of the residuals as compared to the plots in the left panels. Although it is very hard to judge whether a trend is still present in the residual scatter plots versus fitted values and latitude, we conclude that the model suggested by Peixoto is an enormous improvement as compared to the linear regression model in Equation (2.1). To conclude that the new model is really appropriate for the data at hand and no lack-of-fit is present, we do need a statistical test to assess the parametric model fit. Moreover, if there would be a lack-of-fit, it would be most welcome to have some graphical tools that formally locate lack-of-fit in the predictor space. The latter will be the main topic of this dissertation. So rather than dealing with model building or variable selection techniques, we mainly focus in the following chapters on the validation of the quality of a selected model that is assumed to be useful by the data analyst.

CHAPTER 3

Review of some lack-of-fit tests

In this chapter, a limited review of some lack-of-fit tests is provided. It is not intended to be a complete literature review of LOF tests. There is a huge amount of literature available, but not all of them are relevant for the research presented in this dissertation. The selected tests are included because of there historical importance, because they are frequently used in real case studies, because of their good power properties or because of their relation with the main results in this dissertation.

We consider the general setting where

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

describes the regression relationship between the mean of the response variable y and one single predictor $x \in \mathbb{R}$, where the x_i, $i = 1, \ldots, n$ are assumed to be known and fixed, and n denotes the sample size. Assume the error terms ϵ_i, $i = 1, \ldots, n$, where n denotes the sample size, to be independently and identically distributed with mean zero and variance σ^2. The central null hypothesis states that m belongs to a given parametric family of functions,

$$H_0 : m \in \mathcal{M} = \{m(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}, \tag{3.2}$$

where Θ is a p-dimensional proper parameter set in $\mathbb{R}^p$.

We assume rather severe distributional conditions at first sight, but these are mainly to obtain nice limit distributions for the different statistics described in this chapter. Throughout this chapter, possible extensions, such as random designs, multiple covariates, or heteroscedasticity, will be discussed wherever relevant. However, for most test statistics, one of the bootstrap schemes described in Section 3.5 will be appropriate, so that most of them are easily applicable in real case studies.

3.1 Classical lack-of-fit F test and similar approaches

3.1.1 Pure error lack-of-fit F test

Probably the best known and most frequently described lack-of-fit test in regression textbooks (e.g. Neter et al. (1996), Draper and Smith (1998)) is the classical LOF F test, or the pure error F test (Fisher, 1922). This test is able to detect any kind of deviations from a parametric linear regression model, but assumes ideal circumstances for lack-of-fit testing. More precisely, more than one replicate should be available for at least one of the different design points. To highlight the need for repeated observations, we rewrite Model (3.1) as

$$y_{ij} = m(x_i, \boldsymbol{\theta}) + \epsilon_{ij}, \quad i = 1, \ldots, c, \;\; j = 1, \ldots, n_i, \tag{3.3}$$

where the index j stresses the presence of n_i replicates at the i^{th} design point. Note that $\mathbf{y}$ is not a matrix but represents the $n \times 1$ vector of observable response values with $n = \sum_{i=1}^{c} n_i$. As we assume $m(x_i, \boldsymbol{\theta})$ to be linear, $m(x_i, \boldsymbol{\theta}) = \sum_{j=0}^{p-1} m_j(x_i)\theta_j$, where $\boldsymbol{\theta}$ is a p vector of unknown parameters, $m_0(x_i) = 1$ for all i, and $m_j(x_i), j = 1, \ldots, p-1$, are known functional forms of the predictor value x for the i^{th} covariate pattern, e.g. a power function or a logarithmic transform of x, etc. Note that the value of the latter are the same for all replicates in design point i. The errors are assumed to be i.i.d. $N(0, \sigma^2)$. The pure error LOF test statistic contrasts the error sum of squares of a so-called *full* model with respect to a *reduced* model. Let F denote the full model that imposes no restrictions on the means of different design points,

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, c, \tag{3.4}$$

where the μ_i are parameters which may be different for each design point. This model does not assume any predefined relationship between the mean of the response variable and the values of the regressors, which is actually the one-way analysis of variance model. Model (3.3) is the model under the null hypothesis and is referred to as the reduced model, denoted by R. The F test allows a formal decision about whether the more complex, full model should be preferred over the reduced model under the null hypothesis. The corresponding hypotheses have the form

$$H_0 : m(x_i) = \sum_{j=0}^{p-1} m_j(x_i)\theta_j \text{ versus } H_a : m(x_i) = \mu_i, \quad i = 1, \ldots, c.$$

Note that we assume that $p \leq c$ so that the reduced model involves less parameters than the full model. In addition, assume $n - p > c - 1$. This means that

a sufficient amount of replicates, compared to the number of groups, have to be available in the data set. By finding the error sums of squares (SSE) in both models, the test statistic,

$$F = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}, \tag{3.5}$$

can be computed, with df_R and df_F the degrees of freedom of $SSE(R)$ and $SSE(F)$ respectively. The denominator corresponds to the pure error mean squares, $\frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}{n-c}$, where $\overline{y}_i$ denotes the mean of the response variable at design point i. Whether the null model is appropriate or not, this estimator always provides an unbiased estimator of the error variance σ^2. It is a model-free variance estimator, since no parametric model for m is involved. The numerator equals the LOF mean sum of squares, $\frac{\sum_{i=1}^{c} n_i(\overline{y}_i-\hat{y}_i)^2}{c-p}$, with $\hat{y}_i$ the fitted value of the response variable for the i^{th} group of replicates, obtained by least squares regression of the reduced model. Under the null, this estimator is also an unbiased estimator of σ^2. This one requires a predefined parametric model for m and is therefore called a model-based variance estimator. However, when the true relationship between the response and the predictor variable considerably deviates from the specified null model, the estimator will tend to overestimate the error variance. These two measures of deviations are illustrated in Figure 3.1. The F-ratio is a good measure of lack-of-fit, since the estimator in the denominator is an unbiased estimator of the error term variance, no matter what the true regression function is, i.e. under both the null and the alternative hypothesis. The estimator in the numerator is constructed to be unbiased under the null hypothesis, but is biased upwards if the hypothesized regression model is not appropriate. Thus, when say a higher order polynomial model would be more appropriate compared to a simple linear regression model, large values of the test statistic will probably show up. The null hypothesis is only rejected when the ratio is sufficiently larger than one. Under the null hypothesis of no lack-of-fit, this statistic is F distributed with $df_R - df_F$ and df_F degrees of freedom.
The idea of considering a test statistic which is a ratio of a model-based variance estimator and a model-free variance estimator as a measure for deviations from the null model, will be frequently used in this and further sections.

Example 1 Ice crystal data. *The ice crystal data set is discussed in Draper and Smith (1981), example R on p. 66, but originally the data comes from Ryan et al. (1976). Ice crystals are introduced into a chamber, the interior of which is maintained at a fixed temperature ($-5°C$) and a fixed level of saturation of air with water. The growth of*

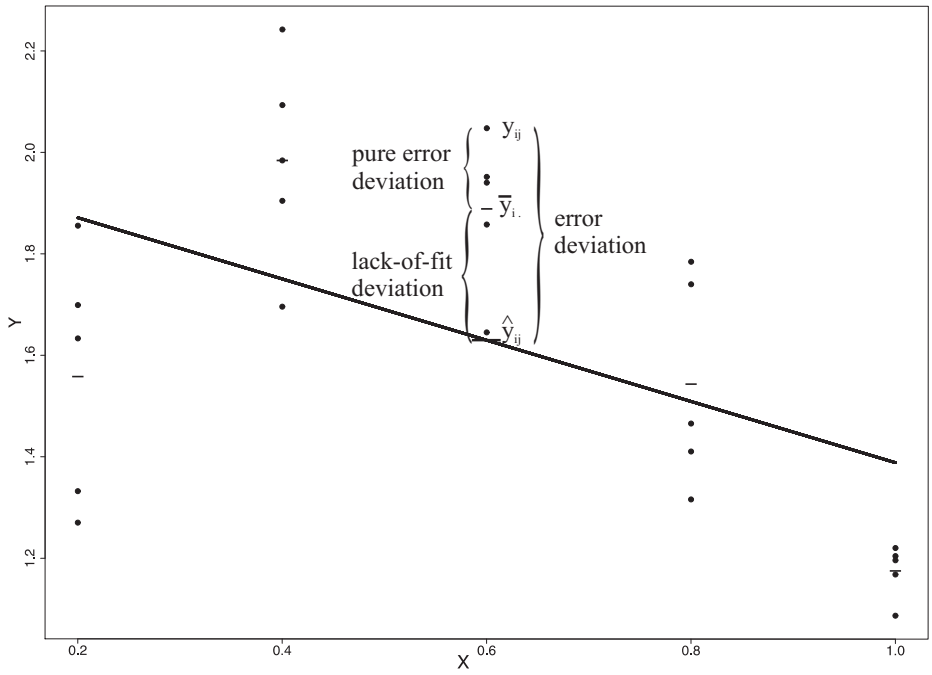

FIGURE 3.1: *Illustration of the decomposition of the error deviation $y_{ij} - \hat{y}_i$, with j indicating the j^{th} replicate of x_i, $i = 1, \ldots, c$, $j = 1, \ldots, n_i$ and $\overline{y}_i$ the mean of the response variable in the i^{th} group of replicates.*

crystals with time is observed. The 43 sets of measurements are of axial length of the crystals (A) in micrometers for times (T) of 50 seconds to 180 seconds from the introduction of the crystals. Each measurement represents a single complete experiment. The experiments were conducted over a number of days, and were randomized with respect to observations time. It was desired to learn whether a straight line model, $A = \theta_0 + \theta_1 T + \epsilon$ provided an adequate representation of the growth with time of the mean axial length of the ice crystal.
Exact replicates are available for this example, thus the classical F-test is applicable. The pure error lack-of-fit test statistic equals 0.79 (p=0.70). Although the fit of a loess smoother in Figure 3.2 indicates the presence of a small bump in the mid range of the predictor variable time, no sufficient evidence of lack-of-fit is found for this example.

Example 2 Motorcycle data. *Figure 3.3 shows the motorcycle data (Härdle, 1990). The x-values denote time (in milliseconds) after a simulated impact with motorcycles. The response variable y is the head acceleration (in g) of a post mortem human test*

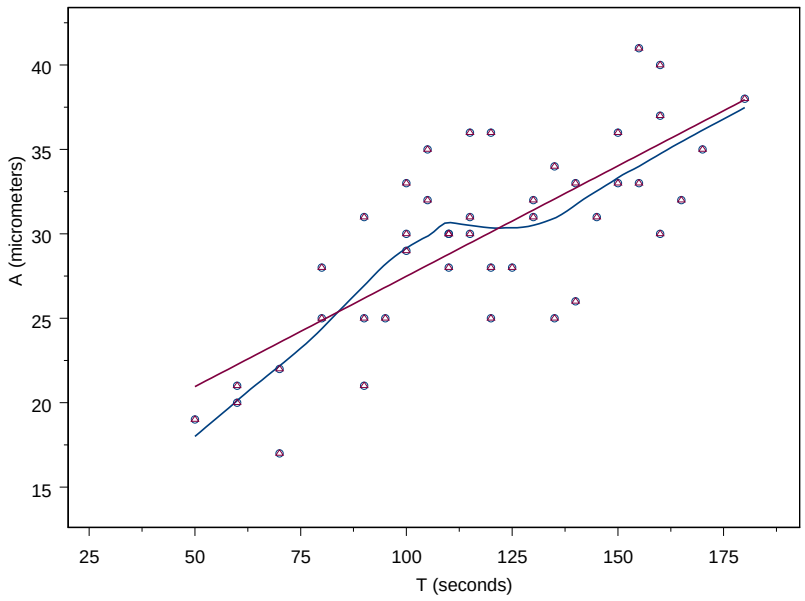

FIGURE 3.2: *Ice Crystal Data (Ryan et al., 1976); A = axial length of the ice crystal in micrometers; T = times in seconds from the introduction of the crystals. The straight line represents the least squares fit of a linear model, the smoothed line is the fit of a loess smoother to the data (span=0.75).*

subject. The smoothed line is the fit of a loess smoother to the data (span=0.30). When the no-effect hypothesis is tested against a full model for the motorcycle data, a clear lack-of-fit is found in both the graphical representation and by means of the pure error LOF test. The value for the test statistic equals 4.56, which corresponds to a p-value smaller than 0.0001.

Although well known and easily applicable, this test is subject to a number of constraints which makes it only applicable in a limited number of datasets. Firstly, repeated observations at one or more x levels are required. Secondly, the null distribution of the test statistic is only exact when the error terms are Gaussian with constant variance σ^2, the model is linear, and the parameters are estimated by least squares.
To overcome the limitations caused by the requirement of replicates, an extensive research has been done during the nineteen seventies and eighties to pro-

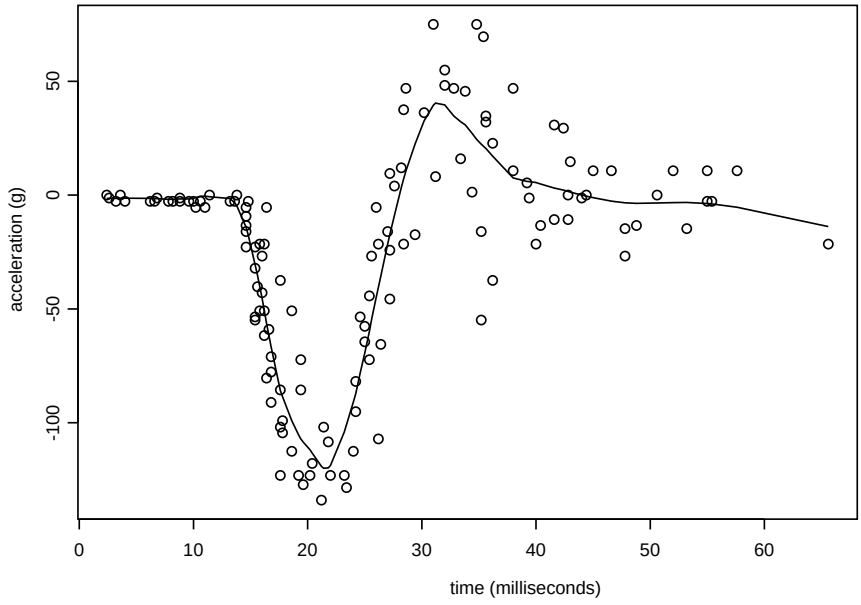

FIGURE 3.3: *Motorcycle Data (Härdle, 1990); y = the head acceleration (in g) of a post mortem human test object; x = time (in milliseconds) after a simulated impact with motorcycles. The smoothed line is the fit of a loess smoother to the data (span=0.30).*

vide tests that work when replicates are not available, e.g. in non-designed or observational experiments. Many of these procedures use the concept of *pseudo-pure error* estimates of the error variance. These estimates can be constructed based on clusters of near-replicates, near-neighbour pairs, piecewise regression or low leverage points. For a detailed review of these procedures, the reader is referred to e.g. Neill and Johnson (1984), Joglekar et al. (1989), Christensen (1991), Su and Yang (2006), and the references therein. Another option in case of a sufficiently smooth regression function is to replace the pure error mean sum of squares in test statistic (3.5) by the consistent variance estimator suggested by Gasser et al. (1986) based on *pseudo-residuals*. It uses the same idea of treating neighbouring design points as near-replicates, but does not involve the subjective choice of defining clusters of near-replicates. Pseudo-residuals, say $\tilde{e}_i$, $i = 2, \ldots, n-1$, are obtained by taking triples of subsequent design points, i.e. x_{i-1}, x_i, x_{i+1}, and fitting a straight line between the outer two. The pseudo-

residual $\tilde{e}_i$ is the difference between the observed value and the one predicted by the straight line,

$$\begin{aligned}\tilde{e}_i &= \frac{x_{i+1}-x_i}{x_{i+1}-x_{i-1}}y_{i-1} + \frac{x_i - x_{i-1}}{x_{i+1}-x_{i-1}}y_{i+1} - y_i \\ &= a_i y_{i-1} + b_i y_{i+1} - y_i.\end{aligned}$$

The variance estimator $\hat{\sigma}_P^2$ based on these pseudo-residuals is defined as

$$\hat{\sigma}_P^2 = \frac{1}{n-2}\sum_{i=2}^{n-1} c_i^2 \tilde{e}_i^2, \tag{3.6}$$

where $c_i^2 = (a_i^2 + b_i^2 + 1)^{-1}$. If some multiple measurements are present in practice, a modification of this estimator is suggested by Gasser et al. (1986). The probability distribution for this alternative test statistic needs to be derived before a formal test can be conducted, or it needs to be approximated by one of the procedures described later in this chapter. Other model-free variance estimators may be used as well (Hart, 1997).
Concerning the second remark about the normality and the linearity of the null model, one can in practice often bootstrap the null distribution instead of using large sample distribution theory. This issue will be discussed in Section 3.5.

The extension to multiple predictor variables $\mathbf{x} \in \mathbb{R}^d$ is straightforward.

3.1.2 Reduction method

Another closely related approach to assess the fit of a parametric model that does not require exact replicates is the reduction method. The idea is to overfit the data by introducing supplementary regression terms to the null model. When both the null and supplementary terms are linear in the parameters, the alternative model can be written as

$$y_i = \sum_{j=0}^{p-1} m_j(x_i)\theta_j + \sum_{k=1}^{q} g_k(x_i)\gamma_k + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.7}$$

where m_j and g_k are known functions and θ_j and γ_k are unknown parameters. The null hypothesis reduces to H_0:$\gamma_1 = \ldots = \gamma_k = 0$. It is now straightforward to fit both models by least squares and to compare the corresponding variance estimators as in (3.5). Consider thus the alternative model as the full model in the classical F-test and obtain the test statistic

$$F = \frac{\frac{SSE_0 - SSE_a}{q}}{\frac{SSE_a}{n-p-q}}, \tag{3.8}$$

in accordance with Equation (3.5), where SSE_0 and SSE_a represent the error sums of squares under the null and the alternative model, respectively. This ratio actually measures the reduction of the error sum of squares by fitting the alternative model, which explains the name "reduction method". No exact replicates are necessary in this setting. In the special case of Gaussian error terms with constant variance σ^2, and under H_0, the test statistic is F distributed with q and $n - p - q$ degrees of freedom. Among a class of invariant tests, the reduction test is uniformly most powerful for testing nested linear models (Lehmann, 1959).

Example 3 *The reduction method is extremely useful if one has a particular alternative in mind, like for example in case of the US temperatures data. To illustrate this method, the null model, a first order polynomial model in longitude, x_1, and latitude, x_2,*

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon,$$

is tested against an alternative model suggested by Peixoto (1990), a third order polynomial model in longitude and first order in latitude,

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \gamma_1 x_1 x_2 + \gamma_2 x_1^2 + \gamma_3 x_1^2 x_2 + \gamma_4 x_1^3 + \gamma_5 x_1^3 x_2 + \epsilon.$$

Using least squares, the reduction test statistic equals to 53.23 ($p < 0.0001$). Their is no doubt that the alternative model is superior to the first order polynomial model in longitude and latitude.

3.2 Nonparametric and smoothing based lack-of-fit tests

Most of this section is taken from the monograph of Hart (1997).

The reduction method turns out to be very useful, as it does not require any replicates and has optimal power for testing nested linear models. However, it might have no power at all against certain other types of alternatives. In the literature, this is called a directional test. Omnibus tests, on the other hand, have some power against all kinds of alternatives, and thus do not need the specification of any kind of alternative model in advance. The ideal setting would thus be to find an omnibus test, that has rather high power against a wide range of important or interesting alternatives. Tests based on nonparametric regression models are designed with this purpose in mind. Many nonparametric test statistics are constructed as the ratio of two variance estimators of which one no longer depends on the fit of a particular parametric

alternative that has been suggested beforehand by the data-analyst.

The true regression model (3.1) and the central null hypothesis (3.2) remain the focus of this section, though several tests are only designed to test the no-effect hypothesis,

$$H_0 : m(x, \theta) = \theta_0. \tag{3.9}$$

3.2.1 Some historical nonparametric tests

von Neumann test

The von Neumann statistic dates to 1941, and is used in a LOF test for testing the no-effects null hypothesis in case of a single predictor variable x. von Neumann et al. (1941) pointed out that for observations with a constant variance, but with a smooth trend in mean, it is not appropriate to calculate the variance without a correction. This results in an overestimation of the true population variance. Instead they suggested to use the mean sum of squares of successive differences, i.e. $\hat{\delta}^2 = \frac{\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2}{n-1}$, where y_i corresponds to the concomitant of the i^{th} order statistic of the covariate x, i.e. the response values y are ordered with respect to x. This variance estimator is less sensitive to the effect of the trend than the conventional sample variance estimator. Originally, von Neumann (1941) suggested to use the ratio of the mean sum of squares of successive differences and the biased sample variance, $\hat{\sigma}_b^2$,

$$\frac{\hat{\delta}^2}{\hat{\sigma}_b^2} = \frac{\frac{\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2}{(n-1)}}{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n}},$$

to detect possible trends as opposed to the no effect hypothesis. Small values of the test statistic indicate possible deviations from the null model.
Over the years, this statistic was adjusted several times. Since both estimators are biased, $\mathrm{E}(\hat{\delta}^2) = 2\sigma^2$ and $\mathrm{E}(\hat{\sigma}_b^2) = \frac{n-1}{n}\sigma^2$, Harper (1967), among others, suggested the ratio of unbiased estimators

$$\frac{\frac{\hat{\delta}^2}{2}}{\hat{\sigma}_u^2} = \frac{\frac{\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2}{2(n-1)}}{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}}.$$

In order to obtain the more familiar ratio of two unbiased estimators, where the numerator is more sensitive to model deviations than the denominator, the more common version of the test statistic becomes

$$T_N = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_D^2} = \frac{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}}{\frac{\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2}{2(n-1)}}, \tag{3.10}$$

but it is still referred to as the von Neumann test statistic (see e.g. Hart (1997) and others) due to the minor adjustments.
Note that half the mean sum of squares of successive differences,

$$\hat{\sigma}_D^2 = \frac{\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2}{2(n-1)}, \tag{3.11}$$

is an unbiased estimator of σ^2 which is actually the variance estimator that is nowadays well known as the Rice variance estimator (Rice, 1984).
The T_N statistic can be rewritten in matrix notation,

$$T_N = \frac{\frac{\mathbf{Y}^t(\mathbf{I}_n - n^{-1}\mathbf{J}_n)\mathbf{Y}}{\mathrm{tr}(\mathbf{I}_n - n^{-1}\mathbf{J}_n)}}{\frac{\mathbf{Y}^t\mathbf{D}\mathbf{Y}}{\mathrm{tr}(\mathbf{D})}}, \tag{3.12}$$

where $\mathbf{Y}$ is the $n \times 1$ response matrix, $\mathbf{I}_n$ the $n \times n$ identity matrix, $\mathbf{J}_n$ an $n \times n$ matrix of all 1's, and $\mathbf{D}$ the $n \times n$ tridiagonal matrix

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

The use of $\mathbf{D}$ in (3.12) is to obtain the first differences. The degrees of freedom corresponding with the variance estimators are expressed in terms of the trace of the matrices, $\mathrm{tr}(\mathbf{D}) = 2\mathrm{tr}(\mathbf{I}_n - n^{-1}\mathbf{J}_n) = 2(n-1)$.

von Neumann's test is equivalent to the Durbin Watson test (Durbin and Watson, 1950) for testing for positive serial correlation in a sequence of constant-mean variables.

Generalization of the von Neumann test

Consider the $(p-1)^{th}$ order polynomial regression model $y_i = \sum_{j=0}^{p-1} x_i^j \theta_j + \epsilon_i$ in a single predictor x, which is to be tested against an unspecified alternative model. Let $\mathbf{X}$ denote the $n \times p$ design matrix for the polynomial regression model, and $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ the $n \times n$ hat matrix. In this setting, the mean

squared error,

$$\hat{\sigma}^2_M = \frac{\mathbf{Y}^t(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\mathrm{tr}(\mathbf{I}_n - \mathbf{H})}, \tag{3.13}$$

could be used as the model based variance estimator in the generalized statistic, while a generalization of the half mean sum of squares of successive differences variance estimator, $\hat{\sigma}^2_D = \frac{\mathbf{Y}^t(\mathbf{I}_n-\mathbf{H})^t\mathbf{D}(\mathbf{I}_n-\mathbf{H})\mathbf{Y}}{\mathrm{tr}(\mathbf{D}(\mathbf{I}_n-\mathbf{H}))}$, is utilized as a model free estimator. The generalized test statistic is thus the ratio

$$T_N = \frac{\hat{\sigma}^2_M}{\hat{\sigma}^2_D} = \frac{\frac{\mathbf{Y}^t(\mathbf{I}_n-\mathbf{H})\mathbf{Y}}{\mathrm{tr}(\mathbf{I}_n-\mathbf{H})}}{\frac{\mathbf{Y}^t(\mathbf{I}_n-\mathbf{H})^t\mathbf{D}(\mathbf{I}_n-\mathbf{H})\mathbf{Y}}{\mathrm{tr}(\mathbf{D}(\mathbf{I}_n-\mathbf{H}))}}. \tag{3.14}$$

Distribution of a ratio of two quadratic forms

Hart (1997) describes a procedure to obtain the probability distribution of any ratio of quadratic forms. In particular, let

$$V_n = \frac{\mathbf{Y}^t\mathbf{A}\mathbf{Y}}{\mathbf{Y}^t\mathbf{B}\mathbf{Y}}, \tag{3.15}$$

where $\mathbf{A}$ and $\mathbf{B}$ are known matrices, and assume the error terms to be Gaussian random variables. Under the null hypothesis, for an observed value v of the statistic V_n,

$$P(V_n \leq v) = P\left(\sum_{j=1}^{r} \lambda_{jn}(v) Z_j^2 \leq 0\right),$$

where $r = \mathrm{rank}(\mathbf{A} - v\mathbf{B})$, $\lambda_{jn}(j = 1, \ldots, r)$ are the non-zero eigenvalues of the matrix $\mathbf{A} - v\mathbf{B}$, and $Z_1, \ldots, Z_r$ are i.i.d. $\mathrm{N}(0, \sigma^2)$. In this way, the p-value that corresponds to an observed value v of any test statistic V_n can be numerically approximated.

When the error terms are non-Gaussian, many of these tests must rely on the bootstrap methods described in Section 3.5.

This simple and powerful von Neumann test to assess the fit of linear models against unspecified alternatives, will be included further in this dissertation in the analysis of data examples and simulation studies with a single predictor variable.

Remark

Basically, for generalizations of the von Neumann test, the original idea is applied to the residuals from the linear model. The same idea could be used in testing the fit of a nonlinear model (Hart, 1997). The test statistic is based on the residuals obtained from the nonlinear model fit, say $e_i = y_i - m(x_i, \hat{\boldsymbol{\theta}}_n)$, and has the form $\frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=2}^{n}(e_i - e_{i-1})^2}$. The nonlinearity of the model results in the additional problem that the null distribution becomes dependent on the values of the unknown regression parameters. A double bootstrap procedure as described in Section 3.5 provides a suitable solution.

As a second remark, note that the von Neumann test implicitly assumes that no replicates are present in the data, as the performance of the test would depend on the order of the response values among tied observations.

Cusum test of Buckley

The numerator of the von Neumann statistic measures departures from the hypothesized model in a rather non smooth way. In particular, it squares individual residuals and then sums them. Buckley (1991) proposed a first attempt to measure deviations in a smoother way by using cumulative sums of residuals to obtain an estimate of the residual variance. For equally spaced data and to test the no-effect hypothesis, this estimate is proportional to $n^{-2}\sum_{j=1}^{n}(\sum_{i=1}^{j} y_i - \overline{y})^2$. Here, the residuals are first summed, and then squared. For smooth departures from the null model, this estimator is more sensitive than for example the mean squared error. From the previous section we know that the difference based estimator is rather insensitive to smooth model departures. As Buckley's statistic is another example of a variance ratio of two unbiased estimators under the null hypothesis, it can be written as a ratio of two quadratic forms, so that the null distribution can be easily approximated. To test for the $(p-1)^{th}$ order polynomial regression model, $y_i = \sum_{j=0}^{p-1} x_i^j \theta_j + \epsilon_i$, against an unspecified alternative model, this test statistic has the form

$$T_B = \frac{\frac{\mathbf{Y}^t(\mathbf{I}_n - \mathbf{H})^t \mathbf{S}^t \mathbf{S}(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\mathrm{tr}(\mathbf{S}(\mathbf{I}_n - \mathbf{H})\mathbf{Y})}}{\frac{\mathbf{Y}^t\mathbf{D}\mathbf{Y}}{\mathrm{tr}(\mathbf{D})}}, \tag{3.16}$$

where $\mathbf{S}$ is the p^{th} order cusum operator as defined in Buckley (1991).

Example 4 Windmill data. *The windmill data (Montgomery and Peck, 1982) contain information on the Direct Current (DC) Output, and the wind velocity (miles per hour). The data are shown in the left panel of Figure 3.4, together with the least squares fit of a simple linear regression model of the mean of the DC output as a function of the wind velocity. This scatterplot already suggests that the linear model is not appropriate to describe this relationship. Both the von Neumann test and Buckley's cusum test are applied to test for a linear relationship between the DC Output, y, and the wind velocity, x. Their p-values are found by means of the procedure to determine the probability distribution of a ratio of two quadratic forms. The value of the test statistics are 3.888 ($p < 0.0001$) and 54.478 ($p < 0.0001$) for the the von Neumann test and Buckley's cusum test, respectively. Both tests allow us to formally conclude an inadequate model fit.*

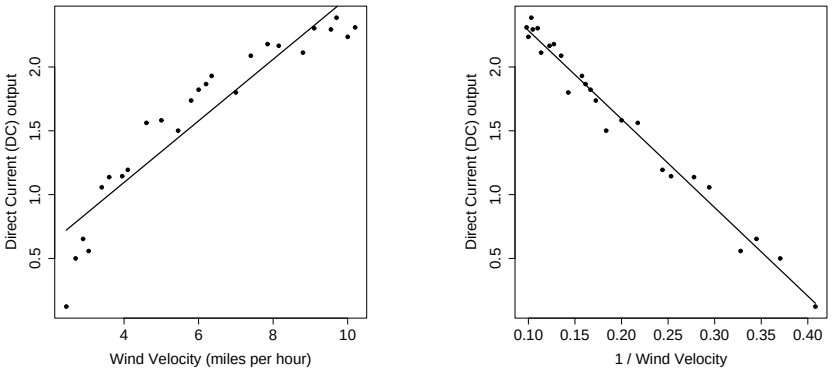

FIGURE 3.4: *(Left panel) Windmill Data (Montgomery and Peck, 1992); y = Direct Current (DC) Output; x = Wind Velocity (miles per hour). (Right panel) Windmill Data; y = Direct Current (DC) Output; x = Reciprocal Transformation on Wind Velocity. The line represents the fitted linear regression line.*

In the literature, it has already been suggested that for this particular dataset, a reciprocal transformation on x is appropriate to fit a linear relationship with the mean of y. The result is shown in the right panel of Figure 3.4. If we apply both tests on the transformed data, the corresponding values of the test statistics become, T_N = 0.919 (p=0.843) and T_B = 0.654 (p=0.505). The p-values and the graph no longer suggest any evidence of LOF.

Both tests are simple and can be used to assess the fit of linear models against unspecified alternatives. Small sample power properties are studied later in this and following chapters.

Weighted sums of squared sample Fourier coefficients

Eubank and Hart (1993) provided a canonical decomposition of both the von Neumann and the Buckley statistics in terms of sample Fourier coefficients for testing the no effect null hypothesis (3.9). Assume equally spaced design points $x_i = \frac{i-0.5}{n}$, $(i = 1, \ldots, n)$ of a single covariate $x \in [0, 1]$. Both test statistics may be represented as

$$T = \frac{2\sum_{j=1}^{n-1} w_{j,n}\hat{\phi}_{j,n}^2}{\hat{\sigma}^2}, \tag{3.17}$$

where

$$\hat{\phi}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} y_i \cos(\pi j x_i), \quad j = 1, \ldots, n-1, \tag{3.18}$$

are the sample Fourier coefficients obtained by least squares and $\hat{\sigma}^2$ is a consistent estimator of σ^2. The von Neumann and Buckley statistics are obtained by assigning different weights $w_{j,n}$ to the Fourier coefficients in Equation (3.17). The von Neumann statistic weights all $n-1$ Fourier coefficients equally, more precisely, $w_{j,n} = 1$. The weights for the Buckley statistic are $w_{j,n} = \frac{n}{(2n\sin(j\pi/(2n)))^2}$, and thus put more weights on the first coefficients, which correspond to low frequency alternatives or smooth deviations. Buckley's test is therefore more powerful to detect smooth deviations from the null model and is in this case superior to the von Neumann statistic. The latter, on the other hand, has equally good power for both low and high frequency alternatives and is thus favourable in case of high frequency alternatives.

The performance of both tests can also be discussed by studying their large sample powers. The von Neumann and Buckley tests are both consistent against any non constant, sufficiently smooth function. The von Neumann test has non-trivial power against alternatives converging to the null model at the rate $n^{1/4}$, while the Buckley test is superior in the sense that it achieves the parametric rate of $n^{1/2}$. Note that these results are only valid for large samples in the limit. More details on the local alternatives considered for these tests can be found in Eubank and Hart (1993) or in Hart (1997).

Simulation Study 1 *To get an idea about the agreement between the large and small sample powers of the tests, we will add from time to time in this chapter a small simulation study, originally performed by Eubank and Hart (1993), but now applied to a selected number of LOF tests discussed in this chapter. In the literature, an important distinction is made between low and high frequency alternatives. As an example of a*

low frequency alternative, the function m_1 is chosen as

$$m_1(x) = \beta \left(e^{4x} - \frac{e^4 - 1}{4} \right) \left(\frac{e^8 - 1}{8} - \left(\frac{e^4 - 1}{4} \right)^2 \right)^{-\frac{1}{2}},$$

and as a representative of high frequency alternatives,

$$m_2(x) = 2\beta \left(20 \left(x - \frac{1}{2} \right)^3 - 3 \left(x - \frac{1}{2} \right) \right),$$

where $\beta = 0, 0.1, \ldots, 1.0$ is the degree of LOF. In particular $\beta = 0$ corresponds to the no effect null hypothesis (3.9) and $\beta = 1$ results in the functions shown in the left panel of Figure 3.5. The simulation results are obtained for an evenly spaced, fixed design $x_i = \frac{i-0.5}{n}$, $i = 1, \ldots, n$, with $n = 40$ and standard normally distributed error terms. The plots are based on 5000 Monte Carlo simulation runs. All tests are performed at the 5% significance level. In the middle panel, one can clearly see the power advantage of Buckley's cusum test against low frequency alternatives, where as the von Neumann test is remarkably superior in case of the high frequency alternative (right panel).

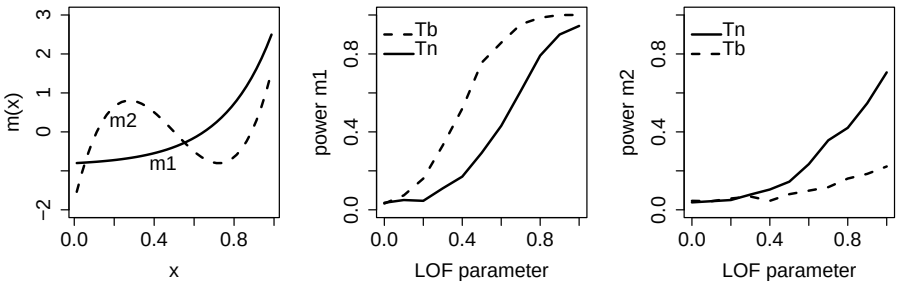

FIGURE 3.5: *(Left panel) Illustration of the low (m_1) and high frequency (m_2) alternative model with parameter $\beta = 1.0$. (Middle panel) Empirical power curves for the different values of the parameter β for m_1. (Right Panel) Empirical power curves for the high frequency alternative m_2.*

Neyman smooth test

The Neyman smooth test is well known in the goodness-of-fit context (Rayner and Best, 1989), but also has an equivalent in the regression context. Consider

the k^{th} order smooth alternative,

$$m(x, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{k} \theta_j \psi_{j,n}(x), \tag{3.19}$$

where the functions $\psi_{j,n}$ $(j = 1, \ldots, k)$ form a set of orthonormal functions, over equally spaced fixed design points $x_r = \frac{r-0.5}{n}$, $(r = 1, \ldots, n)$ such that for any $0 \leq i, j \leq k$,

$$\frac{1}{n} \sum_{r=1}^{n} \psi_{i,n}(x_r)\psi_{j,n}(x_r) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j, \end{cases}$$

and $\psi_{0,n} \equiv 1$. The Neyman smooth test now tests the no effect hypothesis (3.9) against the k^{th} order smooth alternative model (3.19) by means of the test statistic

$$T_{N,k} = \frac{n \sum_{j=1}^{k} \hat{\theta}_j^2}{\hat{\sigma}^2}, \tag{3.20}$$

where $\hat{\sigma}^2$ is some consistent estimator of σ^2 and $\hat{\theta}_j$ is the least squares estimator $\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} y_i \psi_{j,n}(x_i)$. Under the null hypothesis, $T_{N,k}$ is asymptotically distributed as a χ^2 random variable with k degrees of freedom. The k^{th} order Neyman smooth test has power against k^{th} order local alternatives converging to H_0 at the parametric rate $n^{\frac{1}{2}}$.

If the set of orthonormal functions are cosine functions, then the Neyman test statistic is a weighted sum of Fourier coefficients as in (3.17) with weighting scheme

$$w_{j,n} = \begin{cases} 1, & 1 \leq j \leq k \\ 0, & k < j < n. \end{cases} \tag{3.21}$$

Recall that Buckley's test was powerful in detecting low frequency alternatives, i.e. when most energy is situated in the first two Fourier coefficients, but failed to detect high frequency alternatives. The von Neumann test, on the other hand, achieves lower power than Buckley's test for low frequency alternatives, but has equally good power for low and high frequency alternatives, because it puts equal weight on all Fourier coefficients. Taking weighting scheme (3.21) into account, it becomes clear that the Neyman test can be seen as a compromise between these two tests by putting equal weights on the first k Fourier coefficients, but ignoring the higher order terms. We refer to Hart (1997) for more details on the Neyman smooth test.

This test will however not be used in practice since it requires a predefined alternative, and a poor choice of the order k can severely diminish the power. It was mentioned here because of its historical importance. Finally, note that the parameter k is basically a smoothing parameter.

3.2.2 Smooth lack-of-fit tests with fixed smoothing parameters

The Neyman smooth test presented in the previous subsection can be seen as a smoothing based test with fixed smoothing parameter k. An alternative way is to compare a parametric fit $m_{\hat{\boldsymbol{\theta}}_n}$ versus a non parametric fit, say m_n. This nonparametric fit can be based on an orthonormal series expansion as in (3.19), or it can be based on, e.g., kernel or spline estimators. A proper distance measure $d(m_{\hat{\boldsymbol{\theta}}_n}, m_n)$ is taken as a test statistic. If the observed value of the distance measure exceeds a critical value, the null hypothesis of no lack-of-fit is rejected (e.g. Azzalini et al. (1989), Hart (1997) and the references therein). Unfortunately, bias related problems may occur due to the bias in the nonparametric estimate m_n. le Cessie and van Houwelingen (1991) avoided the bias problem by using smoothed function of the residuals, instead of a smoothed version of the regression function. When a test statistic is based on a smooth estimate of the residuals, the bias disappears as these residuals have expectation zero. The tests of le Cessie and van Houwelingen (1991, 1995) are discussed in more detail in Section 7.1 in the special case of logistic regression models.

One major disadvantage of this kind of test is its dependence on the smoothing parameter, which needs to be specified in advance. The performance of these tests is highly sensitive to a wrong choice of this parameter. We do not provide more details on these tests, but a wide literature on data-driven smoothing-based test is available. Many of these data driven tests do no longer suffer from this shortcoming. They are presented in the next subsection.

3.2.3 Tests based on data-driven smoothing parameters

All tests in the previous subsection involve choosing a fixed smoothing parameter. Different choices of this parameter result in different p-values, which is an undesirable property. Therefore, we next describe a class of tests that utilize data-driven smoothing parameters. This means that the smoothing parameter is selected from the data by means of a selection criterion, e.g. cross validation, Akaike's information criterion (AIC), the Bayes information criterion (BIC), estimators of risk, etc. The latter have been widely studied in the regression context by, among others, Yanagimoto and Yanagimoto (1987), Barry and Hartigan (1990), Eubank and Hart (1992), Barry (1993), Eubank et al. (1995), Fan (1996), Kuchibhatla and Hart (1996), Hart (1997), Lee and Hart (1998), Aerts et al. (1999), Aerts et al. (2000), Fan and Huang (2001). In this subsection, we only describe a few of the proposed data-driven smooth tests.
Any orthonormal series estimator could be used in these smoothing based tests,

but to keep the discussion as lucid as possible, we only consider trigonometric series estimators in what follows. Nevertheless, the set of basis functions has considerable impact on the power properties of these tests under different types of alternatives (Hart, 1997).
The descriptions of the test statistics are as in Eubank and Hart (1992), Kuchibhatla and Hart (1996) and Hart (1997), and therefore focus on testing the no effect hypothesis (3.9) against a smooth alternative, over an equally spaced fixed design $x_i = \frac{i-0.5}{n},\ i = 1, \ldots, n$.

The Order Selection Test

To address the bothersome aspect of smooth tests based on a predefined smoothing parameter, Eubank and Hart (1992) proposed to use the data-driven selected smoothing parameter itself as a test statistic. Let $\hat{k}$ denote the selected smoothing parameter. It is defined as the maximizer of the risk criterion

$$r(k, c_\alpha) = \begin{cases} 0 & k = 0 \\ \sum_{j=1}^{k} \frac{2n\hat{\phi}_{j,n}^2}{\hat{\sigma}^2} - kc_\alpha & k = 1, \ldots, n-1, \end{cases} \tag{3.22}$$

where $\hat{\phi}_{j,n}^2$ is defined in Equation 3.18, $\hat{\sigma}^2$ is a consistent estimator, and $c_\alpha > 1$ is a constant so that the desired level of the test can be asymptotically obtained. For an asymptotic level of $\alpha = 0.05$, c_α equals 4.18. We sometimes write k_α and $\hat{k}_\alpha$ to stress the dependence on the level α. This risk criterion is referred to as the Mallows-like criterion since maximizing this risk function for $c_\alpha = 2$, corresponds to Mallows' criterion for selecting the order of terms added in the regression smoother

$$\hat{m}(x, \hat{\boldsymbol{\phi}}, \hat{k}_\alpha) = \hat{\phi}_{0,n} + \sum_{j=1}^{\hat{k}_\alpha} \hat{\phi}_{j,n} \cos(\pi j x). \tag{3.23}$$

Under the null hypothesis, $r(k_\alpha, c_\alpha)$ is very likely to be maximized at k_α equal to zero, as $\hat{m}(x, \hat{\boldsymbol{\phi}})$ then equals $\hat{\phi}_{0,n} = \bar{y}$. This means that the null hypothesis is rejected at level α only if $\hat{k}_\alpha > 0$. In this case, the absolute value of at least one of the sample Fourier coefficients $|\hat{\phi}_{j,n}|$ is nonzero, which entails a nonconstant mean function. This test will be referred to as the Order Selection (OS) test. When the null hypothesis is rejected, the graph of the smooth estimate of the regression function, $\hat{m}(x, \hat{\boldsymbol{\phi}}, \hat{k}_\alpha)$, provides an impression of the true nature of the relationship between x and y. If the graph is non constant, there is evidence against the null hypothesis.

Kuchibhatla and Hart (1996) provided an equivalent form of the test statistic $\hat{k}_\alpha$ that allows a straightforward computation of the p-value. Note that $\hat{k}_\alpha = 0$ if and only if

$$\frac{1}{m}\sum_{j=1}^{m}\frac{2n\hat{\phi}_{j,n}^2}{\hat{\sigma}^2} \leq c_\alpha \quad \text{for all } m = 1, \ldots, n-1.$$

Let T_{OS} denote the equivalent test statistic,

$$T_{OS} = \max_{1\leq m\leq n-1}\frac{1}{m}\sum_{j=1}^{m}\frac{2n\hat{\phi}_{j,n}^2}{\hat{\sigma}^2}. \tag{3.24}$$

The null hypothesis is rejected if T_{OS} is larger than c_α. When t_{obs} denotes the observed value of T_{OS} for a particular data set, the p-value, $p = 1 - P(T_{OS} \leq t_{obs})$, can be approximated by the limiting distribution provided in Eubank and Hart (1992). In particular,

$$p \approx 1 - \exp\left\{-\sum_{j=1}^{M}\frac{P(\chi_j^2 \leq jt_{obs})}{j}\right\},$$

where χ_j^2 denotes a χ^2 distributed random variable with j degrees of freedom and M has to be taken sufficiently large to obtain the desired accuracy. In small sample sizes, one might prefer the bootstrap procedure described in Section 3.5 to obtain a better approximation (Chen et al., 2001).

Eubank and Hart (1992) showed that their OS test is consistent against smooth departures from the null hypothesis and is able to detect local alternatives that converge to the null at the parametric rate of $n^{1/2}$. Finally, the performance of the OS test depends on the choice of the consistent estimator $\hat{\sigma}^2$ of σ^2. Examples include the unbiased sample variance estimator, $\hat{\sigma}_u^2$, $\hat{\sigma}_P^2$ (Equation 3.6), $\hat{\sigma}_M^2$ (Equation 3.13) and $\hat{\sigma}_D^2$ (Equation 3.11).

Extensions to (non)linear models and random designs

If the null model is a polynomial or nonlinear model in the single predictor variable, the previous data-driven smooth tests can be applied to the residuals rather than to the response variable y. More specifically, the sample Fourier coefficients in the test statistics are now defined as

$$\hat{\phi}_{j,n} = \frac{1}{n}\sum_{i=1}^{n} e_i \cos(\pi j x_i), \quad j = 1, \ldots, n-1, \tag{3.25}$$

where e_i are the residuals $y_i - m(x_i, \hat{\boldsymbol{\theta}}_n)$.
Note that the intuitive motivation of all these tests does not depend on the assumption of a fixed design. The only concern is the effect that weakening the assumptions has on the null distribution of the test statistic. For more details on this issue, we refer to the cited references. However, no practical implications are involved, since most authors suggest to bootstrap the sampling distribution anyway.

Extensions to a more general context

The order selection test is extended by Aerts et al. (1999) and Aerts et al. (2000) to more a general context. Their lack-of-fit tests are also based on orthogonal series estimators and use data-driven selection criteria. Next to penalized likelihood criteria, they use penalized score statistics, which only require computation of null parameter estimates. Their methodology is more widely applicable, e.g. in generalized linear models, spectral analysis, the goodness-of-fit problem, and longitudinal data analysis. Alternatives to the null hypothesis are modeled by a sequence of nested orthogonal series or some other appropriate function approximators. For multiple predictor variables, this means that a path in the alternative model space has to be chosen, as many model sequences are possible. They also suggested robust versions of their test statistics against likelihood misspecification.

More data-driven Neyman smooth tests

The Neyman Smooth test as it was presented in (3.20) can be seen as a smoothing based test with fixed smoothing parameter k. The same general idea of maximizing a risk criterion to obtain a data-driven choice of this parameter applies to this test statistic. Kuchibhatla and Hart (1996) suggested use of the Mallows-like criterion (3.22) with $c_\alpha = 2$ defined for the OS test, to obtain a data-driven choice of the smoothing parameter k. They used

$$T_{KH} = \begin{cases} 0 & \hat{k} = 0 \\ \sum_{j=1}^{\hat{k}} \frac{2n\hat{\phi}_{j,n}^2}{\hat{\sigma}^2} & \hat{k} > 0 \end{cases}$$

as a test statistic. The asymptotic probability distribution under H_0 is a mixture of a continuous distribution and one that is degenerate at 0 and is discussed in Kuchibhatla and Hart (1996). However, the authors suggested use of the bootstrap procedure in Section 3.5. Based on their simulation results, they also reported that their test performs in general best when the natural unbiased

variance estimator $\hat{\sigma}_u^2$ is used in the risk function and the half mean difference estimator $\hat{\sigma}_D^2$ in the denominator of the test statistic. In particular, the test turns out to have very good power against high frequency alternatives. This test is referred to as the KH test.
Instead of the Mallows-like criterion (3.22), one could also use the Schwarz criterion (Ledwina, 1994).

Simulation Study 2 *We extend the small simulation study 1 with the OS test and the adaptive Neyman test described by Kuchibhatla and Hart (1996). Empirical powers are approximated by taking 499 Monte Carlo loops and 159 bootstrap loops. Typically, a larger number of Monte Carlo and bootstrap samples are necessary to accurately estimate the power, but we believe that our results are indicative of the comparison between the different tests. In Figure 3.6 one can clearly see the good power of the OS test for the studied alternatives. The OS test performs very well for both the low and high frequency alternative. However, Aerts et al. (2000) showed that for higher frequency alternatives its power decreases rapidly. For the low frequency alternative m_1, Buckley's cusum test seems to remain the highest power in general, while for the high frequency alternative m_2, the adaptive Neyman test seems to perform best overall.*

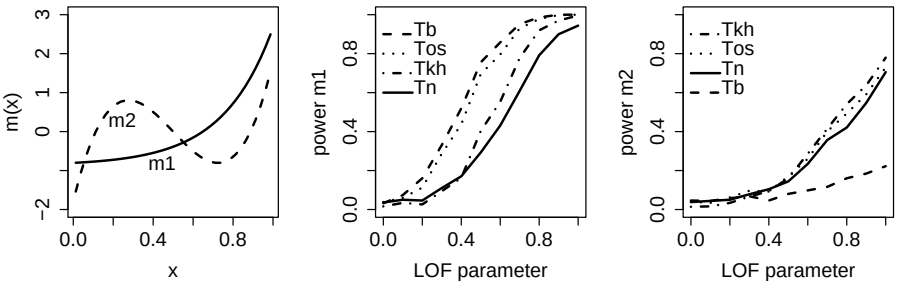

FIGURE 3.6: *(Left panel) Illustration of the low (m_1) and high frequency (m_2) alternative model with parameter $\beta = 1.0$. (Middle panel) Empirical power curves for the different values of the parameter β for m_1. (Right Panel) Empirical power curves for the high frequency alternative m_2.*

Many more variations on this theme are available. Fan and Huang (2001) also formalized the traditional residual plot where a covariate x_j is plotted against the residual, e_j, by testing whether the bias of the vector of residuals is negligible. Instead of using a trigonometric series estimator that is only based on cosine functions, let $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_n)$ be the discrete Fourier transform of the

residual vector **e**, based on both a cosine and sine bases,

$$\hat{\gamma}_{2j-1} = (2/n)^{1/2} \sum_{i=1}^{n} \cos(2\pi ij/n) e_i,$$
$$\hat{\gamma}_{2j} = (2/n)^{1/2} \sum_{i=1}^{n} \sin(2\pi ij/n) e_i, \qquad j = 1, \ldots, [n/2].$$

When n is odd, an additional term $\hat{\gamma}_n = (1/\sqrt{n/2}) \sum_{i=1}^{n} e_i$ is needed, but for linear regression with an intercept, this term is simply zero. The test statistic of this adaptive Neyman test is defined as

$$T_{FH} = \max_{i \le m \le n} \frac{1}{\sqrt{2m\hat{\sigma}^4}} \sum_{i=1}^{m} (\hat{\gamma}_i^2 - \hat{\sigma}^2),$$

where $\hat{\sigma}^2$ is a $n^{1/2}$ consistent estimator of σ^2 under both the null and the alternative hypotheses. For example,

$$\hat{\sigma}^2 = \frac{1}{n - i_n} \sum_{i=i_n+1}^{n} \hat{\gamma}_i^2 - \left(\frac{1}{n - i_n} \sum_{i=i_n+1}^{n} \hat{\gamma}_i \right)^2,$$

for some given i_n ($= [n/4]$, say). The asymptotic null distribution of T_{FH} is given by Fan and Huang (2001), but the approximation is not so good in small samples. Therefore, the bootstrap approximation (Section 3.5) is recommended in practice.
The FH test looks similar to the test proposed by Kuchibhatla and Hart (1996), but tends to select a smaller smoothing parameter. As a consequence this adaptive Neyman test is more powerful than the KH test in detecting very smooth alternatives.

Fan and Huang (2001) further introduced the wavelet-thresholding test. This test combines truncation and thresholding. More specifically, the order of the series is not important, but instead the absolute values of the estimators of the series coefficients are. The term with the largest coefficient estimate enters first in the model, the second largest next, and so forth until the coefficients become lower than a certain threshold. Instead of using the Fourier transform, LOF tests are constructed based on the discrete wavelet transforms. For more details, the reader is referred to Fan and Huang (2001).

Extensions to multiple covariates

If multiple covariates are involved in the null model, the extensions are not straightforward. The performances of the data-driven smooth tests depend

highly on the order of the residuals according to which they are arranged before computing the test statistics. The performance will be optimal if the sequence of the residuals is as smooth as possible, so that large Fourier coefficients are concentrated on low frequencies. One approach suggested in Kuchibhatla and Hart (1996), Hart (1997) and Fan and Huang (2001) is to use these tests by regressing residuals on a scalar function of $\mathbf{x}$. For $\mathbf{x} \in \mathbb{R}^d$, the order relation may be defined as $\mathbf{x}_i \leq \mathbf{x}_j$ if

- all components of $\mathbf{x}_i$ are smaller than or equal to those of $\mathbf{x}_j$, this means $x_{ik} \leq x_{jk}$, where $k = 1, \ldots, d$,
- the k^{th} component of $\mathbf{x}_i$ is smaller than or equal to that of $\mathbf{x}_j$, thus $x_{ik} \leq x_{jk}$ for a specified $k \leq d$,
- $s_i \leq s_j$, where s_i is the score of a specified function of $\mathbf{x}_i$, e.g. the first principal component.
- $\hat{y}_i \leq \hat{y}_j$, where $\hat{y}$ denotes the predicted values of the fitted regression model.

When the tests are calculated in several directions, for example, with respect to each covariate direction separately, the Bonferroni adjustment should be applied to the combined test to obtain a global conclusion with a family-wise error rate.
Instead of fitting a smooth trigonometric series to the sequence of residuals, one could apply a multidimensional smoother to the residuals over the predictor space, see le Cessie and van Houwelingen (1991).
Further, Aerts et al. (2000) constructed lack-of-fit tests based on orthogonal series estimators which involve choosing a nested model sequence in the multiple regression setting. They described different orders of model sequences in the case of two covariates.
In Chapter 6 of this dissertation, we will introduce another solution based on a distance measure in the predictor space, which avoids an ordering of the residuals in advance, or the choice of a smoothing parameter and has nevertheless nice smoothing properties. The price that has to be paid is a rather heavy computational burden.

3.2.4 Tests based on residual cusum processes

In the nineteen nineties, a series of methods were proposed that avoid smoothing. Motivated by the fact that the least squares residuals in case of no lack-of-fit should fluctuate randomly around zero, a number of authors suggested using test statistics based on cumulative sums of residuals to validate the quality of

a model fit. To be more specific, such tests are based on the residual cusum process,

$$\hat{\mathbb{B}}_n(x) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(x_i; \hat{\boldsymbol{\theta}}_n)) I(x_i \leq x), \quad x \in \mathbb{R}, \tag{3.26}$$

which constitutes a *marked empirical process*, where the marks are given by the residuals. If the model is not appropriate, large sequences of positive or negative residuals will occur, provided that the true model and the model under the null hypothesis do not intersect too much. This will result in large values of a predefined norm of the cumulative sums of residuals, which turns out to be a useful test statistic to detect LOF.
Zuber (1996) studied such tests for testing the no-effect hypothesis with constant variance σ^2 and fixed design points. He compared the performance of Kolmogorov-Smirnov and Cramér-von Mises type tests, and concluded that they perform rather similarly for the alternatives under study, with a slight advantage of the Cramér-von Mises type test. Stute (1997) investigated the Kolmogorov-Smirnov type to assess the fit of linear regression models in case of homoscedasticity and random design points, while Su and Wei (1991) described this procedure specifically to assess the fit of generalized linear models (McCullagh and Nelder, 1989). They obtained a sensitive test to detect both missing predictor variables in the hypothesized model and a misspecified link function. For linear regression models, they expect their test to be very powerful against quadratic deviations and less powerful against higher order polynomials. A number of mistakes in their distributional theory were pointed out by Stute (1997). Finally, Diebolt and Zuber (1999), and Zuber (1999) extended the results for possibly nonlinear, heteroscedastic regression models on fixed designs.

In case of multiple predictor variables, Su and Wei (1991) suggested considering the supremum of the process

$$\hat{\mathbb{B}}_n(\mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) I(\mathbf{x}_i \leq \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d,$$

where the order relation is defined as $\mathbf{x}_i \leq \mathbf{x}$ if all components of $\mathbf{x}_i$ are smaller than or equal to those of $\mathbf{x}$, i.e. $x_{ij} \leq x_j$, $(j = 1, \ldots, d)$. In particular, this includes the special case suggested in Lin et al. (2002), who advise checking the functional form of the j^{th} component of the covariate vector $\mathbf{x}$, by the process

$$\hat{\mathbb{B}}_{n,j}(x_j) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) I(x_{ij} \leq x_j).$$

Since processes based on cumulative sums of residuals tend to be dominated by observations with small covariate values, they further discuss the use of moving sums of residuals with respect to one component of the covariate vector $\mathbf{x}$. For blocks with fixed size b, they suggest using the modified process

$$\hat{\mathbb{B}}_{n,j}(x_j, b) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) I(x_j - b \leq x_{ij} \leq x_j). \tag{3.27}$$

However, the moving sums are based on blocks of the same size b, so the number of observations in the blocks can be quite different when the covariate values are not evenly distributed. Therefore, moving averages were also studied,

$$\overline{\hat{\mathbb{B}}}_{n,j}(x_j, b) = \frac{n^{1/2} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)) I(x_j - b \leq x_{ij} \leq x_j)}{\sum_{i=1}^{n} I(x_j - b \leq x_{ij} \leq x_j)}. \tag{3.28}$$

The performance of these different processes will be further discussed in Chapter 4. We expect LOF tests based on the latter two processes to be powerful in detecting deviations in the functional form of the j^{th} component of the same size as the fixed block size b. Large block sizes are favourable to detect global LOF, over more or less the entire range of a component of the covariate vector $\mathbf{x}$, while small block sizes will be sensitive to local deviations. Of course, the block size has to be chosen in advance and different block sizes may lead to conflicting conclusions.

The authors try to get an indication of the nature of the deviations by studying prototype mean functions for the moving sums of residuals for several block sizes. Resemblance of an observed pattern of the residual processes with one of the prototype functions may suggest the nature of the misspecification.

One of the newly proposed approaches in this dissertation, outlined in Chapter 4, solves the dependencies on the fixed block size of the LOF tests based on processes (3.27) and (3.28) by considering all possible intervals obtained with respect to each covariate x_j, which results in powerful tests for both global and local lack-of-fit. Since our new tests are closely related to these processes, we provide some more distributional details on the marked empirical process based on residuals (3.26). The large sample results in Chapter 4 and Chapter 8 are mainly based on the next theorems, which are taken from Diebolt and Zuber (1999) and Zuber (1999). The distributional theory is subject to a number of assumptions.

Assumption 1 *The moment $E(\epsilon_1^2)$ is finite.*

Assumption 2 *The distribution of the design, say $F(x), x \in \mathbb{R}$, is continuous and strictly increasing.*

Assumption 3 *The function $\sigma^*(u) = \sigma(F^{-1}(u)), u \in [0,1]$ is positive and continuous on $[0,1]$.*

Assumption 4 *The regression function $m(x;\boldsymbol{\theta})$ and its first two partial derivatives with respect to $\boldsymbol{\theta}$ are continuous in $x \in \mathbb{R}$ for each $\boldsymbol{\theta}$, its first partial derivative is bounded for each $\boldsymbol{\theta}$ and the integrals $\int_{-\infty}^{\infty} \left| \frac{\partial m(y;\boldsymbol{\theta})}{\partial \theta_k |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}} \right| dF(y)$ are finite for $k = 0, \ldots, p-1$. There exist*

- *a real number $r_0 > 0$ such that the closed ball $\overline{B}(\boldsymbol{\theta}_0, r_0) \subset \Theta$, and*
- *a known function $M_2 \geq 0$ such that $\int_{-\infty}^{\infty} M_2(y) dF(y) < \infty$,*

that satisfy the condition

$$\sup_{\boldsymbol{\theta} \in \overline{B}(\boldsymbol{\theta}_0, r_0)} \left| \frac{\partial^2 m(x;\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right| \leq M_2(x)$$

for all $x \in \mathbb{R}$ and for all $j, k = 0, \ldots, p-1$.

Assumption 5 *The sequence of estimators $\{\hat{\boldsymbol{\theta}}_n\}$ of $\boldsymbol{\theta}_0$ converges almost surely to $\boldsymbol{\theta}_0$, and satisfies the condition*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\varphi}_0(x_i)\epsilon_i + o_{\mathbb{P}}(1),$$

with $\boldsymbol{\varphi}_0$ a function such that $\int_{-\infty}^{\infty} \|\boldsymbol{\varphi}_0\|^2 dF$ is finite for a certain norm on $\mathbb{R}^p$.

For the discussion of the distributional properties, we will from now on assume that $x \in [0,1]$. This includes no restrictions, since for $x \in \mathbb{R}$, and by putting $\hat{\mathbb{B}}_n(-\infty) = 0$ and $\hat{\mathbb{B}}_n(\infty) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(x_i; \hat{\boldsymbol{\theta}}_n))$, $\hat{\mathbb{B}}_n$ takes its values in the Skorokhod Space $D[-\infty, \infty]$. A classical quantile transformation, $u_i = F(x_i)$, allows us to work in the more familiar space $D[0,1]$ by considering the marked empirical process based on residuals from a uniform design on the unit interval $[0,1]$, $\hat{\mathbb{B}}'_n(u) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(F^{-1}(u_i); \hat{\boldsymbol{\theta}}_n)) I(u_i \leq u)$. Assumption 2, allows us to work with the inverse function F^{-1} of F, without ambiguity.

Under Assumption 5, as $n \to \infty$, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges to a p-dimensional normal random variable with zero mean and variance matrix $\boldsymbol{\Gamma}_0 = \int_{-\infty}^{\infty} \boldsymbol{\varphi}_0 \boldsymbol{\varphi}_0^T dF$.

Theorem 1 establishes the limiting centered Gaussian process $\hat{\mathbb{B}}$ of $\hat{\mathbb{B}}_n$.

Theorem 1 *Under H_0 and the Assumptions 1 - 5, $\hat{\mathbb{B}}_n \xrightarrow{w} \hat{\mathbb{B}}$, as $n \to \infty$, in the Skorokhod space $D[-\infty, \infty]$, where $\hat{\mathbb{B}}$ is a centered Gaussian process with covariance function*

$$r(x,y) = G(x \wedge y) - \mathbf{g}_0^T(x)\mathbf{h}_0(y) - \mathbf{g}_0^T(y)\mathbf{h}_0(x) + \mathbf{g}_0^T(x)\boldsymbol{\Gamma}_0\mathbf{g}_0(y),$$

where

$$\begin{aligned} G(x) &= \int_{-\infty}^{x} \sigma^2(u)dF(u), \\ \mathbf{g}_0(x) &= \int_{-\infty}^{x} \boldsymbol{\nabla}\mathbf{m}_0(u)dF(u), \\ \mathbf{h}_0(x) &= \int_{-\infty}^{x} \sigma(u)\boldsymbol{\varphi}_0(u)dF(u), \end{aligned}$$

with $\boldsymbol{\nabla}\mathbf{m}_0 = \boldsymbol{\nabla}\mathbf{m}_{\boldsymbol{\theta}|\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ the gradient with respect to $\boldsymbol{\theta}$ of $m(x,\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$.

However, Theorem 1 shows that the limit process depends on the null model and can take rather complicated structures. Obtaining critical values by means of the bootstrap seems to be more straightforward. Stute et al. (1998) showed that the wild bootstrap (Section 3.5.3) yields a consistent approximation of the distribution of the limit process. The simpler residual based bootstrap (Section 3.5.2) is only valid in case of homoscedasticity.

Simulation Study 3 *Finally, we add the cusum based test of Zuber (1996) to simulation study 2. Let $T_{Z,KS}$ denote the Kolmogorov-Smirnov type and $T_{Z,CM}$ denote the Cramér von Mises type of test statistic based on process $\hat{\mathbb{B}}_n(.)$. More specifically,*

$$T_{Z,KS} = \sup_{x \in \mathbb{R}} \left| \hat{\mathbb{B}}_n(x) \right| \tag{3.29}$$

and

$$T_{Z,CM} = \left(\int_{-\infty}^{+\infty} \hat{\mathbb{B}}_n(x)^2 dx \right)^{1/2} \approx \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{B}}_n(x_i)^2 \right)^{1/2}. \tag{3.30}$$

Empirical powers are approximated by taking 499 Monte Carlo loops and 159 bootstrap loops. In Figure 3.7, both tests show very good powers for the low frequency alternative under study, with a slight power advantage for the Cramér von Mises type. On the

other hand, they have hardly any power at all for the high frequency alternative. This can be expected, since systematic patterns of negative and positive values will cancel out. However, to show that much depends on the sample size n, we redo the simulation study for $n = 200$. The OS test performs very well for both the low and high frequency alternative. For the low frequency alternative m_1, Buckley's cusum test seems to remain the highest power in general, while for the high frequency alternative m_2, the adaptive Neyman test (KH) seems to perform best overall.

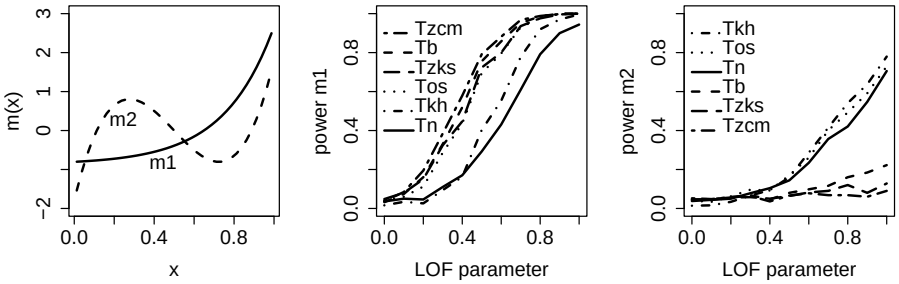

FIGURE 3.7: *(Left panel) Illustration of the low (m_1) and high frequency (m_2) alternative model with parameter $\beta = 1.0$. (Middle panel) Empirical power curves for the different values of the parameter β for m_1 and $n = 40$. (Right Panel) Empirical power curves for the high frequency alternative m_2.*

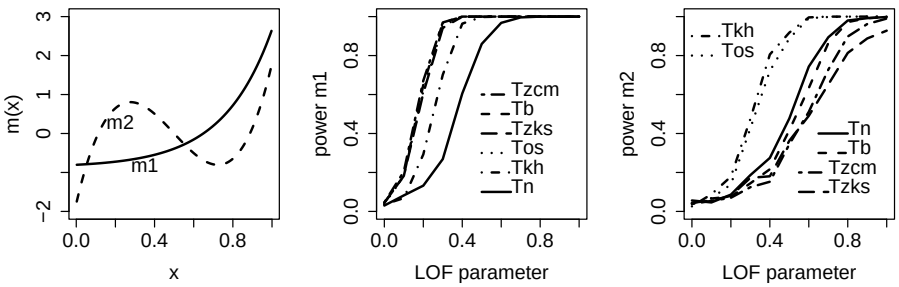

FIGURE 3.8: *(Left panel) Illustration of the low (m_1) and high frequency (m_2) alternative model with parameter $\beta = 1.0$. (Middle panel) Empirical power curves for the different values of the parameter β for m_1 and $n = 200$.(Right Panel) Empirical power curves for the high frequency alternative m_2.*

Instead of taking the norm of the process $\hat{\mathbb{B}}_n$ as a test statistic, Diebolt and Zuber

(1999) and Zuber (1999) proposed to use the Karhunen-Loève expansion of the Gaussian limit process to obtain a χ^2 test statistic. Their test has equal power against low and high frequency alternatives and is able to detect alternatives that approach the null hypothesis at rate $n^{1/2}$. The Karhunen-Loève expansion can be seen as the principal components analysis of the Gaussian process $\hat{\mathbb{B}}(.)$ and allows large sample power investigations, as well as the derivation of smooth and directional tests (Stute, 1997). For more details, we refer the reader to Diebolt and Zuber (1999) and Zuber (1999).
In practice, the principal components are often difficult to obtain. Therefore, Stute et al. (1998) proposed to replace the cusum process by its innovation martingale. For the new processes, principal components are readily available and the resulting tests turn out to be asymptotically distribution free under composite null models. The authors showed how to derive optimal directional tests based on their innovation process approach.

3.3 LOF tests in the context of logistic regression models

Logistic regression models belong to the family of generalized linear models (McCullagh and Nelder, 1989). In logistic regression analysis, the error terms are no longer continuously distributed. The response variable, y_i is binary and thus only takes the values 0 or 1, often called failure and success, respectively. In particular, the conditional distribution of this response is Bernoulli with parameter $\pi(\mathbf{x}_i) = P\{y_i = 1 \,|\mathbf{x} = \mathbf{x}_i\}$. When no replicates are available, the residuals only take the values

$$e_i = y_i - \hat{\pi}(\mathbf{x}_i) = \begin{cases} 1 - \hat{\pi}(\mathbf{x}_i) & \text{if } y_i = 1, \\ -\hat{\pi}(\mathbf{x}_i) & \text{if } y_i = 0, \end{cases}$$

where $\hat{\pi}$ denotes an estimator of π. In the logistic model, the logit of this probability is modeled as a linear function of the predictor variables,

$$\text{logit}(\pi(\mathbf{x}_i)) = \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \sum_{j=0}^{p-1} m_j(\mathbf{x}_i)\theta_j,$$

where $\mathbf{m}(\mathbf{x}_i)^t$ is the $p-$dimensional vector of the functional forms of d fixed covariates. The estimator $\hat{\pi}(\mathbf{x})$ is obtained by replacing the θ_j's in this linear function by their maximum likelihood or weighted least squares estimators.
More generally, we will denote by n_i the number of replicated observations available at the i^{th} design point, called covariate pattern, $n_T = \sum_{i=1}^{n} n_i$ the total number of observations for the n different covariate patterns, and y_i the number of successes for the specified design point. In logistic regression there are

several possible ways to measure the difference between the observed and the fitted values. Three different types of residuals are frequently used in the literature,

- the *raw* or *working* residuals, $e_{r,i} = y_i - n_i\hat{\pi}(\mathbf{x}_i)$,
- the *Pearson* residuals, $e_{P,i} = \frac{y_i - n_i\hat{\pi}(\mathbf{x}_i)}{\sqrt{n_i\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))}}$,
- the *deviance* residuals,
  $e_{d,i} = \text{sign}(y_i - n_i\hat{\pi}(\mathbf{x}_i))\sqrt{2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}(\mathbf{x}_i)}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}(\mathbf{x}_i)}\right)\right]}$.

The Pearson residuals are the scaled measures of the differences of observed to fitted values. The deviance residuals are the contributions to the deviance due to the difference in the observed and fitted values. Note that both Pearson and deviance residuals, are the signed square roots of the individual contributions of the different design points to the Pearson test statistic and the deviance function respectively, i.e.

$$\chi^2_P = \sum_{i=1}^{n} e^2_{P,i} \quad \text{and} \quad D = \sum_{i=1}^{n} e^2_{d,i}.$$

Both statistics could be used as measures of lack-of-fit. Under a number of assumptions these two statistics are assumed to be asymptotically or approximately distributed as χ^2_{n-p}, where p denotes the number of parameter estimates under the null hypothesis. However, as pointed out in McCullagh and Nelder (1989), these assumptions are certainly not met when most of the n_i are small. Therefore, Hosmer et al. (1991) suggested comparing the value of both test statistics with their degrees of freedom. If the value of the test statistic is much larger than the corresponding degrees of freedom, a strong indication of LOF is present. There is no doubt that more appropriate LOF tests should be used in the assessment of a logistic regression model.

3.3.1 Early alternatives to the Pearson χ^2 test statistic

To solve the distributional problem for the Pearson χ^2 test statistic that occurs when most n_i are small, Tsiatis (1980) suggested a partitioning of the predictor space into k distinct regions. However, the choice of the k distinct regions is arbitrary and the performance of this test is greatly affected by this subjective choice (Su and Wei (1991)).

Unweighted residual sum-of-squares test

Copas (1989) introduced a very simple LOF test, the Unweighted Residual Sum-of-Squares (URSS) test, which is based on $S = \sum_{i=1}^{n}(y_i - n_i\hat{\pi}(x_i))^2$. Since the χ^2 test is only asymptotically valid when the expected frequencies $n_i\pi(x_i)$ and $n_i(1-\pi(x_i))$ are sufficiently large, Copas suggested to give less weight to those covariate patterns with small values of n_i. In simulation studies later on, we calculate the asymptotic moments of S as suggested in Hosmer et al. (1997) and use the standardized test statistic

$$\frac{S - \mathrm{tr}(\hat{\mathbf{V}})}{\hat{\mathbf{d}}^t\hat{\mathbf{V}}^{1/2}(\mathbf{I}_n - \mathbf{H})\hat{\mathbf{V}}^{1/2}\hat{\mathbf{d}}},$$

where $\hat{\mathbf{d}}$ is a vector with i^{th} element $\hat{d}_i = (1 - 2\hat{\pi}(x_i))$ and $\hat{\mathbf{V}}$ is the $n \times n$ diagonal variance covariance matrix of y with i^{th}-element $n_i\hat{\pi}(x_i)(1-\hat{\pi}(x_i))$. The standard normal distribution can now be used as an approximate null distribution.

The Hosmer - Lemeshow tests

Hosmer and Lemeshow (1980), Lemeshow and Hosmer (1982) and Hosmer et al. (1988) proposed and discussed the use of χ^2-like lack-of-fit tests based on grouping the values of the estimated probabilities. In summary, they advise using $g = 10$ groups based on the percentiles of the estimated probabilities, especially when many of the estimated probabilities are small. This means that the first group contains the $n_1 = n_T/10$ subjects having the smallest estimated probabilities and so on. The Hosmer-Lemeshow test statistic, is given by

$$C = \sum_{k=1}^{g}\frac{(o_k - n_k\overline{\pi}_k)^2}{n_k\overline{\pi}_k(1-\overline{\pi}_k)},$$

where for the k^{th} decile, n_k denotes the total number of subjects, c_k is the number of different covariate patterns, o_k denotes the observed number of successes, $o_k = \sum_{j=1}^{c_k} y_j$, and $\overline{\pi}_k = \sum_{j=1}^{c_k}\frac{n_j\hat{\pi}_j}{n_k}$, the average estimated probability. Under the null hypothesis of no LOF, the distribution of the test statistic C is approximated by the χ^2 distribution with $g-2$ degrees of freedom. If the null hypothesis is rejected, inspection of the g individual components may indicate regions where the model does not fit satisfactorily.

Although very valuable and well known among practitioners, important local deviations in regions of the covariate space that yield the same estimated probabilities, may be missed by this test statistic. To overcome this disadvantage, the authors suggested using the individual contributions to the test statistic as a first check for possible local deviations within one of the deciles. Further, groups constructed by means of a grouping strategy like the one suggested above, may contain subjects with widely different values of covariates.
Many extensions are suggested to overcome this problem, e.g. Pulkstenis and Robinson suggested to perform this grouping within the cross-classification of all categorical covariates in the model. However, this extension is only useful when the logistic regression model contains both categorical and continuous covariates. Moons et al. (2004) suggested to construct groups based on the recursive partitioning algorithm underlying classification trees. This approach has a beneficial effect on the power characteristics of the test, and can easily handle large datasets with a high dimensional covariate space. However, many have to be made for the practical implementation. For example, the choice of partitioning scheme and pruning process, including the number of final nodes and the number of observations in final nodes.

3.3.2 Smoothing based LOF in logistic regression

To address the issue of a subjective choice of partitioning the predictor space, a smoothing based approach for generalized linear models was introduced by le Cessie and van Houwelingen (1991 and 1995). They used smoothed residuals, i.e. weighted averages of residuals in the neighbourhood of a design point $\mathbf{x}_i$,

$$\hat{r}_{s,i} = \sum_{j=1}^{n} w_{ij} \frac{(y_j - \hat{\pi}_j)}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}},$$

where the weights w_{ij} depend on the type and size of smoothing. Their test statistic is a weighted sum of squares of the smoothed standardized residuals,

$$T = \sum_{i=1}^{n} \frac{\hat{r}_{s,i}^2}{\widehat{\text{var}}(\hat{r}_{s,i}^2)}.$$

In this way, they obtain a procedure that adequately handles continuous covariates rather than subjectively partitioning the range of the covariate. However, the problems of partitioning the predictor space as in Tsiatis (1980) or Hosmer and Lemeshow (1980) more or less remain, because the issue is now choosing the type of kernel and the bandwidth. The authors report that the performance of T depends heavily on the bandwidth. If it is chosen too small,

the test has no power and if it is too large, local deviations are smoothed away.

We include the test based on smoothed residuals in our simulation studies in Chapter 7. The residuals are smoothed in the predictor space, using a uniform kernel as suggested in Hosmer et al. (1997). The weights w_{ij}, defined by the distance between the i^{th} and j^{th} covariate pattern, are

$$w_{ij} = \prod_{k=1}^{p} u(x_{ik}, x_{jk}), \text{ where } u(x_{ik}, x_{jk}) = \begin{cases} 1 & \text{if } \left| x_{ik} - x_{jk} \right| / s_k \leq c_u, \\ 0 & \text{otherwise,} \end{cases}$$

where s_k denotes the sample standard deviation of the k^{th} covariate and c_u is taken to be $c_u = \frac{1}{2}(4/n^{1/(2p)})$, so that about $\sqrt{n}$ subjects have non-zero weights. In addition, we consider a similar test that uses smoothed residuals in the response space, as described in Hosmer et al. (1997). They used cubic weights w_{ij},

$$w_{ij} = \begin{cases} 1 - (\left| \hat{\pi}_i - \hat{\pi}_j \right| / c_{ci})^3 & \text{if } \left| \hat{\pi}_i - \hat{\pi}_j \right| \leq c_{ci}, \\ 0 & \text{otherwise,} \end{cases},$$

where c_{ci} depends on i and is chosen such that again about $\sqrt{n}$ subjects have non-zero weights for each subject.
Information about why the model does not fit can be found by plotting the smooth residuals.
The asymptotic null distribution of these test statistics is a scaled χ^2 distribution. For computational details, we refer to le Cessie and van Houwelingen (1991) and Hosmer et al. (1997).

3.3.3 Tests based on residual cusum processes

The tests of Su and Wei (1991), described in Section 3.2.4, were actually introduced in the context of generalized linear models and are thus in particular applicable in the logistic regression context. We refer to Section 3.2.4 for the general description of the test, but we would like to add that the authors claim that their methodology is sensitive to detect a misspecified link function or to detect the omission of relevant independent variables.
Hosmer and Hjort (2002) extended the ideas of Su and Wei (1991) and proposed the partial sums to be computed over partitions of the estimated logit. In addition, the individual residuals are weighted by functions that are derived to be optimal for detecting particular alternatives to the fitted model. Both authors suggested a bootstrap simulation approach (Section 3.5.1) to approximate the limiting distribution.
The test of Su and Wei (1991) is extended to generalized linear mixed models by Pan and Lin (2005).

3.4 Graphical diagnostic tools

Residuals are not only the basic building block for LOF tests, they are also widely used in graphical diagnostic tools. Most graphics to assess the adequacy of regression models are illustrative and indicative, and the results depend on the data analyst. In general, graphical methods allow visualization of possible discrepancies between the fitted model and the data. Nevertheless, judging whether the observed discrepancies are really present or not is often a major problem and systematic departures smaller than the noise level can often not be observed. Among the graphical diagnostic tools, the classical residual plot, where the residuals are plotted against a covariate or the fitted values, is probably the best known. It is often used as a descriptive method to assess lack-of-fit in a regression analysis. The OS, KH and FH tests in Section 3.2.3 formalize the classical residual plot. When the null hypothesis is rejected, the graph of the smooth estimate of the regression function provides an impression of the true nature of the relationship between x and y. If the graph is non constant, there is evidence against the null hypothesis. This approach is an attempt to obtain a graphical diagnostic tool, that is directly associated with a lack-of-fit test, but a major disadvantage of this procedure is its dependence on the choice of the smoother and the bandwidth. Sometimes the procedure can be made adaptive by using a data-driven choice of the bandwidth.

Landwehr et al. (1984) introduced a variety of residual and partial residual plots appropriate for logistic regression. Further, the smooth residuals of le Cessie and van Houwelingen (1991) are plotted to collect information about why the model does not fit or to get an idea of where the LOF is located in the predictor space. However, no formal interpretation can be given to deviations that are observed from this plot.

Another approach is to contrast the fit of a consistent nonparametric estimate with the fit of a parametric model to assess the parametric fit. The discrepancies between the two fits can be visualized by comparing two curves graphically over the range of data (e.g. by creating reference bands as described by Bowman and Young (1996)). The graphs can however not be used as an inferential tool on there own. A test statistic (e.g. a pseudolikelihood ratio test by Azzalini and Bowman (1993)) is needed to judge the observed discrepancies in a formal way. Pardoe (2001) introduced a Bayesian sampling approach to regression model checking that uses Bayes Marginal Model Plots (BMMP's), based on earlier work by Cook and Weisberg (1997). Unfortunately, no formal interpretation is given to the BMMP's.

Finally, Lin et al. (2002) introduced prototype plots of their cumulative sums of residuals for different types of LOF. Resemblance of an observed pattern of the

residual processes with one of the prototype functions may suggest the nature of the misspecification.
As it is most welcome to have a diagnostic plot that formally locates LOF in the predictor space, we introduce a LOF test and its associated formal graphical diagnostic tool in Chapter 4.

3.5 Bootstrap methods in regression

Approximating the sampling distribution in practice by a limit distribution, as is done in the previous section, might not work well in small samples. Much depends on the sample size, the error distribution, the design and the choice of the error variance estimators. Often severe assumptions, like fixed and equally spaced designs or restrictive distributional assumptions on the error distribution, are necessary for obtaining nice limit distributions and other favourable asymptotic properties. Nevertheless, most of these assumptions cannot be justified in a random design setup in real case studies. In addition, the limit distribution of a test statistic is often available, but the convergence might be very slow. Therefore, many authors (Hart (1997), Stute et al. (1998), Fan and Huang (2001), Chen et al. (2001), Hosmer and Hjort (2002), among others) suggested to approximate the sampling distribution of LOF test statistics by means of a bootstrap procedure. Often, better approximations are obtained and bootstrap p-values are then used in hypothesis testing. This is often an elegant solution, which is easy to implement, though sometimes computationally heavy. Although this issue may become of minor importance in the near future, due to the rapidly growing gain in computer power nowadays.
Several bootstrap schemes will be discussed in this section. Which one to choose in practice depends on the distributional assumptions of the error terms and on the regression model under the null hypothesis. In fact, to approximate the limit distribution of the test statistic under the null hypothesis, we need to simulate data from a model specified under the null hypothesis. We follow the advice in Davison and Hinkley (1997) and take the design points in the resampling model the same as in the original data. This means that even when they are randomly sampled, they are not bootstrapped themselves, but treated as fixed. This basically means that the conditional distribution $F_{\mathbf{x}}$ of y given $\mathbf{x}$ is studied.

3.5.1 Parametric versus nonparametric bootstrap schemes

Parametric bootstrap schemes involve some distributional assumptions. For example, assume that it is known that the true regression model is linear, $m(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x$ and that the error terms are normally distributed random

variables with zero means and common, but unknown variances σ^2. A bootstrap sample based on parametric bootstrapping is obtained by sampling from a normal distribution $N(m(x_i, \hat{\boldsymbol{\theta}}_n), \hat{\sigma}^2)$ for all i, where $\hat{\boldsymbol{\theta}}_n$ is the least squares estimator of $\boldsymbol{\theta}$ and $\hat{\sigma}^2$ is one of the consistent variance estimators discussed in this chapter.
For logistic regression models, a parametric bootstrap sample for the binomial response y could be constructed by sampling from $\text{Bin}(n_i, \hat{\pi}(\mathbf{x}_i))$, for all i. More specifically, to obtain the parametric bootstrap distribution of a specific test statistic T in logistic regression models, we proceed as described in Hosmer and Hjort (2002).
For $b = 1, \ldots B$,

1. Obtain a random bootstrap sample of new outcomes, say y_i^*, $i = 1, \ldots, n$ using the fitted values $\hat{\pi}_i$. Take

$$y_i^* = \begin{cases} 1 & \text{if } u_i \leq \hat{\pi}_i \\ 0 & \text{otherwise} \end{cases}, \text{ where } u_i \sim U(0,1). \tag{3.31}$$

2. Fit the logistic regression model using the data $(\mathbf{x}_i, y_i^*)$, resulting in $\hat{\pi}_i^*$ and $\hat{\boldsymbol{\theta}}_*$.

3. Calculate the bootstrap residuals, denoted by $e_1^*, \ldots, e_n^*$.

4. Calculate the statistic $T^*(e_1^*, \ldots, e_n^*)$, which is further denoted by t_b^*.

The bootstrap p-value is the probability $1 - P^*(|T^*| \leq t)$, where P^* denotes the probability under the bootstrap distribution.

A disadvantage of the parametric bootstrap algorithm is that, in general, data sets generated by a poorly fitting regression model do not contain the same statistical properties as the original data set (Davison and Hinkley, 1997). In practice, we often prefer nonparametric bootstrap schemes, since there are less assumptions on the error distribution, say G. All bootstrap schemes discussed below are nonparametric procedures. For more details, we refer the reader to Davison and Hinkley (1997).

3.5.2 Residual based bootstrap in linear regression

This first nonparametric bootstrap scheme is only valid when homoscedasticity holds and when a consistent estimator $\hat{\boldsymbol{\theta}}_n$, e.g. the least squares estimator, is used to estimate the regression parameters $\boldsymbol{\theta}$. Consider the linear regression model

$$y_i = \mathbf{m}(\mathbf{x}_i)^t \boldsymbol{\theta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.32}$$

where $\mathbf{m}(\mathbf{x}_i)^t$ denotes the $p-$dimensional vector of the functional forms of d covariates. The ϵ_i's are i.i.d. with zero means and equal variances σ^2. Let G denote the common error distribution. The empirical distribution function $\hat{G}$ of the raw residuals, obtained after fitting the hypothesized null model, provides a consistent estimator of G (Davison and Hinkley, 1997). To obtain an approximation of the sampling distribution of a certain test statistic T, find the observed value of the test statistic in the original sample, $T(y_1, \ldots, y_n) = t$. Proceed as follows:

For $b = 1, \ldots B$,

1. Obtain the residuals $e_i = y_i - \mathbf{m}(\mathbf{x}_i)^t \hat{\boldsymbol{\theta}}_n$, $i = 1, \ldots, n$.
2. Construct a bootstrap sample $(e_1^*, \ldots, e_n^*)$ by n times drawing with replacement from the set $\{e_1, \ldots, e_n\}$.
3. Set $y_i^* = \mathbf{m}(\mathbf{x}_i)^t \hat{\boldsymbol{\theta}}_n + e_i^*$, $i = 1, \ldots, n$.
4. Calculate the statistic $T^*(y_1^*, \ldots, y_n^*) = t_b^*$. Note that if the hat matrix, $\mathbf{H}$, is used in the statistic T, it remains unchanged since $\mathbf{x}_i^* \equiv \mathbf{x}_i$. Thus, if the test statistic is based on residuals, rather than on the original values, the statistic becomes $T^*((\mathbf{I}_n - \mathbf{H})y_1^*, \ldots, (\mathbf{I}_n - \mathbf{H})y_n^*) = t_b^*$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The bootstrap p-value is the probability $1 - P^*(|T^*| \leq t)$, where P^* denotes the probability under the bootstrap distribution.
By resampling the least squares residuals, the data generating distribution assumes the null model to be true, and obeys the assumptions on the error distribution, since $\mathrm{E}^*(e_i^*) = 0$ and $\mathrm{E}^*(e_i^*)^2 = e_i^2$.

3.5.3 Wild bootstrap

The second bootstrap procedure is more robust against failure of model assumptions like homoscedasticity. The wild bootstrap procedure was originally proposed by Wu (1986), but received its name in Härdle and Mammen (1993), since n different distributions are estimated from n residuals. The procedure only differs from the residual based bootstrap procedure in the construction of the residual bootstrap sample. Define the bootstrap data $y_1^*, \ldots, y_n^*$ by $y_i^* = \mathbf{m}(\mathbf{x}_i)^t \hat{\boldsymbol{\theta}}_n + e_i \nu_i^*$, where ν_i^* is a random variable with expectation $E(\nu_i^*) = 0$, variance $E(\nu_i^{*2}) = 1$ and third moment $E(\nu_i^{*3}) = 1$. Mammen (1993) suggested the most popular choice for the distribution of ν_i^*,

$$F_1 : \nu_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } p = (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } 1-p. \end{cases} \tag{3.33}$$

Recently, Davidson and Flachaire (2001) showed that the Rademacher distribution

$$F_2 : v_i^* = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases} \tag{3.34}$$

may lead to better results than the popular version F_1. The null distribution of the test statistic $T(y_1, \ldots, y_n)$ can be found by constructing B wild bootstrap samples and the corresponding values of the test statistic $t_b^* = T^*(y_1^*, \ldots, y_n^*)$. Intuitively, one can feel why the wild bootstrap works, since at least in many cases and for large sample sizes, it ensures that the first three moments of the bootstrap distribution match the corresponding null distribution.

3.5.4 Double bootstrap

When the parametric null model is a nonlinear model, often the null distribution of a lack-of-fit statistic depends on $\boldsymbol{\theta}$. As a consequence, approximating the null distribution of the statistic by one of the above bootstrap procedures, may lead to a test whose actual level differs from the nominal level. An iterated, or double, bootstrap procedure will often solve this problem. However, the double bootstrap procedure is a rather heavy computational procedure. If B_1 bootstrap samples are taken from the original set, the procedure requires taking another B_2 samples from each set of the B_1 bootstrap samples, leading to a total of $B_1 B_2$ bootstrap samples.
For more details on the double bootstrap procedure we refer to Section 8.3 in Hart (1997), or to Davison and Hinkley (1997).

3.6 Global versus local lack-of-fit

In the previous sections the main focus was on globally assessing the fit of the parametric model in terms of an inappropriate family of functions for the true regression relationship, a misspecified link function, an omitted predictor variable or the presence of the wrong transformation of a predictor variable. It might also be possible that the specified model only shows local deviations from the data in a small subset of the predictor space. Global statistical tests are designed to accept the null hypothesis if the deviation could be reasonably explained by noise. As it is hard to distinguish between small systematic deviations and pure error in such small areas of the predictor space, we expect them to have low power properties in case of local departures from the null model. As is shown in the simulation study in Chapter 4, most global tests from previous sections have typically low power in case of local departures from the null model, and therefore miss an important group of deviations.

Moreover, most LOF tests have both the advantage as well as the disadvantage to summarize a considerable amount of observation into one single value. It might happen that these global tests have missed some important local deviations, and thus before one concludes that a model fits well, it should be examined whether or not the fit is supported over the entire range of covariate patterns. One way to deal with it, is of course the use of individual diagnostics, but another way is to consider a test that is also able to detect local deviations from the null model and locates them in the predictor space. Opsomer and Francisco-Fernández (2006) addresses the same issue. However, their solution is a local LOF test that applies to a subset of the data that is suspected for the presence of local deviations. The test statistic is a local version of the Cramér-von Mises test statistic presented in Alcalá et al. (1999). It is based on the difference of a global parametric and nonparametric fit, evaluated only over the suspicious subset in the predictor space. They point out themselves that their significance levels might not be correct, as one deals with a situation of so called *data snooping*, but argue that low p-values at least provide an indication of a suspicious local pattern in the data. Moreover, the predictor space could be partitioned into several intervals and their local test could be applied to each of the intervals and the corresponding p-values corrected by means of the Bonferroni correction. We believe, however, that this new method suffers from the same disadvantages as previous LOF tests, as a partition of the predictor space should be provided and no guidelines to do so are available. In addition, the power advantage of the local procedure gets lost due to the conservative Bonferroni correction that has to be applied to the p-values. Finally, the performance highly depends on the number and the choice of the partitions, subjectively chosen by the data analyst.

In what follows, we will propose a new type of test statistic that is able to detect both global and local deviations. In addition, we introduce new types of plots to visualize where subsets of deviating observations occur in the predictor space. We start in the next section by introducing these tests and diagnostic plots in case of simple linear regression.

CHAPTER 4

Interval based regional residual plots and tests

The selected review of existing LOF tests (Chapter 3) makes clear that residuals are highly informative for validating the quality of a parametric model. Model deviations are often reflected in a systematic pattern of the residuals and many tests focus on this property. Buckley (1991), Su and Wei (1991), Stute (1997), Diebolt and Zuber (1999), among others, based their tests on cumulative sums of residuals, since large patterns of positive or negative residuals indicate evidence of model deviations. However, these sums accumulate all the residuals associated with covariate values less than x and therefore, the test statistics are dominated by observations with low covariate values. Lin et al. (2002) solves this problem by considering moving sums, where the sums of residuals associated with covariate values in a certain window are taken. In addition, they introduced a test based on moving averages, since for unequally spaced designs the number of observations within moving windows of fixed block size in the predictor space can be quite unstable. However, the major problem with moving sums and moving averages of fixed block size, is that the performances of the tests highly depend on the block size, which has to be defined prior to the analysis. From this point of view, we propose in this chapter LOF tests based on so called *regional residuals*, which are averages of residuals over partitions in the predictor space. These partitions include all possible block sizes. Firstly, we introduce the building blocks of our test statistics, the regional residuals, and develop the corresponding tests. Secondly, to answer the need of formal graphical tools to visualize lack-of-fit in the predictor space, we construct regional residual plots. In a later chapter, we will provide a sketch of the large sample null distribution of the test statistics, although in practice, we advise using one of the bootstrap schemes of section 3.5.

This chapter [1] deals with assessing the quality of a parametric model fit in a

[1] Most of this chapter is published in Deschepper E., Thas O., Ottoy J.P. (2006) *Regional Residual Plots for Assessing the Fit of Linear Regression Models.* Computational Statistics and Data Analysis, 50, 1995-2013.

single predictor variable x. The extension to more predictor variables is the topic of Chapter 6.

4.1 Construction of a LOF test and a graphical diagnostic tool

4.1.1 Regional residuals

Given the independent observations $(x_i, y_i), i = 1, \ldots, n$, let $m(x)$ denote the parametric regression model for the mean of the response variable y in the single predictor variable $x \in \mathbb{R}$,

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_i's are i.i.d. random variables with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. For simplicity, assume that the observations are ordered with respect to the predictor variable x. More specifically, assume that x_i is the i^{th} order statistic $x_{[i]}$ of x, and $y_i = y_{[i]}$ is the y-value associated with $x_{[i]}$, the concomitant.
Consider testing the central null hypothesis in this dissertation, namely, m belongs to a given parametric family of functions,

$$H_0 : m \in \mathcal{M} = \{m(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}, \tag{4.1}$$

where Θ is a p-dimensional proper parameter set in $\mathbb{R}^p$. Note that testing the null hypothesis may include testing for a simple linear regression model $m(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x$, as well as testing for a nonlinear relationship, e.g. $m(x, \boldsymbol{\theta}) = \theta_0 \exp(-\theta_1 x)$, as well as testing for a polynomial regression model in x, like the family of cubic polynomials $m(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3) \in \Theta \subset \mathbb{R}^4$, or simply testing for the no-effect hypothesis $m(x, \theta) = \theta_0$.

Ordinary residuals are defined as $e_i = y_i - m(x_i, \hat{\boldsymbol{\theta}}_n)$, $(i = 1, \ldots, n)$, where $\hat{\boldsymbol{\theta}}_n$ is assumed to be a consistent estimator of $\boldsymbol{\theta}$, e.g. the least squares estimator. A *regional residual* is defined as the average of ordinary residuals in the subset $A_{ij} = [x_i, x_j],\ i, j = 1, \ldots, n,\ i \leq j$, i.e.

$$R(A_{ij}) = \frac{\sum_{k=1}^{n} e_k I(x_i \leq x_k \leq x_j)}{\sum_{k=1}^{n} I(x_i \leq x_k \leq x_j)} = \frac{1}{n_{ij}} \sum_{k=1}^{n} e_k I(x_i \leq x_k \leq x_j), \tag{4.2}$$

where $n_{ij} = \sum_{k=1}^{n} I(x_i \leq x_k \leq x_j)$ is the number of observations in the subset A_{ij}. When no replicated design points are present, a regional residual calculated over an interval A_{ii} is simply the ordinary residual at design point i. However, for design points with multiple measurements, this regional residual is equal to the average of all the multiple classical residuals at that design

point. This means that the availability or the presence of replicated design points is not an issue for future tests based on regional residuals.

Sometimes we use the matrix notation of (4.2),

$$R(A_{ij}) = (\mathbf{I}_{A_{ij}}^t \mathbf{I}_{A_{ij}})^{-1} \mathbf{I}_{A_{ij}}^t (\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

where $\mathbf{I}_{A_{ij}}$ is the $n \times 1$ column matrix for which the k^{th} element is 1 if $x_k \in A_{ij}$ and 0 otherwise. Let $\mathbf{I}_n$ be the $n \times n$ identity matrix, $\mathbf{Y}$ the $n \times 1$ response matrix, and $\mathbf{H}$ the hat matrix. The form of the hat matrix depends on the model. For a linear model, let $\mathbf{X}$ denote the $n \times p$ design matrix, containing the values for all functional forms of the covariate included in the null model. The hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t, \tag{4.3}$$

is the matrix that provides the fitted values as the projection of the outcome variable onto the covariate space. In case of a nonlinear regression model, the hat matrix is given by

$$\mathbf{H} = \mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t, \tag{4.4}$$

where $\mathbf{V}$ denotes the $n \times p$ matrix with elements $m_i^r = \frac{\partial m_i}{\partial \theta_r}$, $i = 1, \ldots, n$, $r = 1, \ldots, p$. In practice, all derivatives have to be evaluated at the estimated parameter $\hat{\boldsymbol{\theta}}_n$.

Under the null hypothesis of no lack-of-fit, these regional residuals have zero mean. The variance of $R(A_{ij})$ under the null hypothesis is given by $n_{ij}^{-1}\sigma^2 h_{ij}^2$, where $h_{ij}^2 = (\mathbf{I}_{A_{ij}}^t \mathbf{I}_{A_{ij}})^{-1} \mathbf{I}_{A_{ij}}^t (\mathbf{I}_n - \mathbf{H})\mathbf{I}_{A_{ij}}$. Sometimes we write $h_{ij}^2 = h_{ij}^2(\mathbf{X})$ to stress the dependence on $\mathbf{X}$.
Standardization of the regional residuals is an important issue in making the regional residuals comparable with each other. However, in practice, the residual variance σ^2 is unknown, but can be replaced by a consistent estimator. The most natural estimator, $S_n^2 = (n-p)^{-1} \sum_{i=1}^n (y_i - m(x_i, \hat{\boldsymbol{\theta}}_n))^2$, is considered first, resulting in the standardized regional residuals

$$R_{S_n^2}(A_{ij}) = \sqrt{n_{ij}} \frac{R(A_{ij})}{S_n h_{ij}}.$$

Lemma 1 *For a linear regression model m and normally distributed error terms, under the null hypothesis of no lack-of-fit,* $R_{S_n^2}(A_{ij}) \overset{H_0}{\sim} t_{n-p}$.
Proof. Under the null hypothesis of no lack-of-fit and in the particular case of normally distributed errors terms, straightforward calculations give

$$\sqrt{n_{ij}} R(A_{ij}) \overset{H_0}{\sim} N(0, \sigma^2 h_{ij}^2(\mathbf{X})).$$

Since $\frac{(n-p)S_n^2}{\sigma^2} \overset{H_0}{\sim} \chi^2_{n-p}$, we find for the standardized regional residual,

$$R_{S_n^2}(A_{ij}) = \frac{\frac{\sqrt{n_{ij}}R(A_{ij})}{\sigma h_{ij}}}{\frac{S_n}{\sigma}} \overset{H_0}{\sim} \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \sim t_{n-p}.$$

□

4.1.2 A lack-of-fit test

As argued in Chapter 3, ordinary residuals often play an important role in assessing the fit of regression models. If the model is correct, all residuals have expectation zero. Thus, averages of residuals over any subset of the predictor space also have expectation zero. We propose to use such averages, regional residuals, to detect possible deviations from the null model. Regional residuals are very suitable building blocks for constructing a lack-of-fit test. If deviations from the null model occur in a certain region of the predictor space, patterns of positive or negative residuals will show up in that neighbourhood, resulting in large absolute values of standardized regional residuals over these regions. Therefore, we suggest to calculate regional residuals over all possible intervals of the covariate x, $A_{ij} = [x_i, x_j]$, $i, j = 1, \ldots, n;\ i \leq j$, instead of a priori specifying a fixed interval length as was done in Lin et al. (2002). Large absolute values of standardized regional residuals suggest a possible lack-of-fit of the hypothesized model, located in the corresponding interval in the predictor space. To overcome the problem of multiplicity and to obtain a global measure of lack-of-fit, taking a norm of all the calculated standardized regional residuals is proposed as a test statistic,

$$T_{RRS} = \sup_{i \leq j; i,j=1,\ldots,n} \left| R_{S_n^2}(A_{ij}) \right|. \tag{4.5}$$

This statistic is sensitive to both *global* and *local* deviations from the hypothesized model (Section 4.2), where global and local refer to large and small intervals in the predictor space, respectively.

We use the subscript RRS to indicate that the test statistic is based on Regional Residuals that are standardized by using the most natural estimator, S_n^2. We refer to this test as the RRS test. In Section 4.1.5 alternative variance estimators to construct standardized regional residuals are discussed, together with their corresponding test statistics.

The following theorem states the asymptotic null distribution of T_{RRS} under the no-effects null hypothesis, $m(x;\theta) = \theta_0$ for a fixed, uniform design. The proof is given in Chapter 8.

Theorem 2 *Let $\mathbb{Z}$ denote a standard Brownian Bridge on [0,1], and let $0 < c < 1$ denote a small nonzero constant, and define $\mathcal{S} = \{(s,t) \in [0,1]^2 : c < t - s < 1 - c\}$. Then, under the no-effect null hypothesis and for a fixed, uniform design, the test statistic T_{RRS} converges in distribution to the supremum norm of $\frac{1}{\sqrt{(t-s)(1-(t-s))}} (\mathbb{Z}(t) - \mathbb{Z}(s))$ over $\mathcal{S}$.*

The condition $c < t - s < 1 - c$ is necessary to let T_{RRS} have a proper limiting distribution. The reason is that the weight function $[(t-s)(1-(t-s))]^{-1/2}$ gets too large for small $t - s$ or $1 - (t - s)$. Note that in fact, the definition of the T_{RRS} needs a slight modification. We additionally assume that $n_{ij} > cn$. However, in practice this assumption always holds, since the test statistic is defined over the design points and even when i equals j, there exists such a constant c. For more details on the derivation of the asymptotic distribution of T_{RRS} for a more general regression model, the reader is deferred to Chapter 8. A more formal argument is given in the proof of Theorem 17.2.1 of Shorack and Wellner (1986).

Since it is closely related to the marked empirical process based on residuals (3.26), the same limitations on using the asymptotic distribution for the test statistics described in Section 3.2.4, will apply here as well. As pointed out there, the limit process depends on the null model and can take rather complicated structures. Obtaining critical values by means of the bootstrap (Section 3.5) seems therefore to be recommended. In particular, we advise using an approximation of the null distribution of test statistic T_{RRS} by means of the residual based bootstrap scheme of Section 3.5.2, allowing both fixed and random designs. If the assumption of a constant error variance σ^2 is relaxed, the wild bootstrap scheme should be used to deal with heteroscedastic errors (Section 3.5.3).

4.1.3 Regional residual plots

Exploratory regional residual plot

We believe that important information is lost by summarizing all discrepancy measures into a single value. We therefore propose to complement the LOF test with a visualization of the individual regional residuals. We suggest two types of regional residual plots. In one plot, the standardized regional residuals, $R_{S_n^2}(A_{ij})$, are plotted in the (i,j) plane. This can, e.g., take the form of a heat map (Figure 4.1, right panel). The x-axis (y-axis) of the heat map shows the

starting point (end point) of the interval for which the standardized regional residual is calculated. Although the regional residuals are only defined for $i \leq j$, the regional residual plot is made symmetrical by filling up the half plane $i > j$ with $R_{S_n^2}(A_{ji})$. The colors of the heat map range from red for large negative values of the standardized regional residuals, to orange for values around zero, to light yellow and white for large positive values. Thus, red or white areas suggest possible regions of lack-of-fit. Alternatively, under normality assumptions, the t-distribution may be used to obtain pointwise p-values which may be plotted in a similar heat map (Figure not shown). However, the interpretation of these regional residual plots has only a pointwise nature. Therefore, these plots are referred to as *exploratory regional residual plots*. It should be clear that the resulting plots only explore a possible lack-of-fit and do not provide a formal way to assess lack-of-fit, for multiplicity is not taken into account. Thus even when there is actually no lack-of-fit, we may expect to see at least one individual p-value smaller than the nominal significance level α, with a probability larger than α. Although these plots are thus too sensitive to include a proper lack-of-fit test, red and white coloured areas will still indicate regions where the standardized average deviations between the observed response values and the mean fitted values are rather large.
The use of these plots is illustrated on an artificial example data set in Figure 4.1. In the left panel of this figure the simulated data are shown, together with a least squares fit (straight line). A sinusoidal deviation is locally added to the linear relationship $y = 5 - 2x$, where $x_i = \frac{i-0.5}{n}$ are equally spaced design points, $i = 1, \ldots, n = 50$, and hence

$$y = \begin{cases} 5 - 2x + 0.6\sin(19x) & \text{if } x \in [0.33, 0.65], \\ 5 - 2x & \text{otherwise.} \end{cases}$$

Gaussian noise with zero mean and a constant variance $\sigma^2 = 0.01$ is added as well. A clear lack-of-fit is thus situated in [0.33,0.65]. The heat map in the right panel shows red areas for small intervals around the interval [0.49,0.65] and for larger intervals that include the interval [0.49,0.65], indicating a local pattern of mainly negative residuals in the area around [0.49,0.65]. On the other hand, white areas occur for small intervals around the interval [0.33,0.49] and for larger intervals that include the interval [0.33,0.49], suggesting an underestimation of the data in this region.

We would like to stress once more that no statistical test is involved by representing the standardized regional residuals in a heat map and that the exploratory regional residuals plot does not have a formal interpretation. Next, we introduce a *formal regional residual plot*, which protects correctly for a

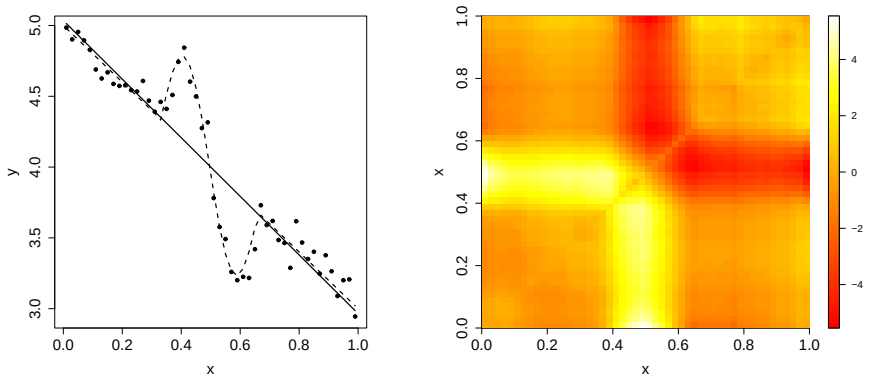

FIGURE 4.1: *(Left panel) True (dashed line) and fitted (solid line) regression model ($n = 50$); lack-of-fit situated in [0.33;0.65]; (Right panel) Exploratory regional residual plot. This plot shows the individual standardized regional residuals to visualize possible areas of LOF.*

family-wise error rate of α.

Formal regional residual plot

Apart from the exploratory regional residual plot, the lack-of-fit test in section 4.1.2 can be complemented with a two-dimensional formal graphical tool, which is called a *formal regional residual plot*. This plot does take the multiplicity into account, and is constructed by only indicating the intervals for which the absolute value of the standardized regional residual exceeds the bootstrap critical value of the test statistic T_{RRS}. An example of this diagnostic tool is shown in Figure 4.2 (right panel). White areas in the formal regional residual plot refer to large positive standardized regional residuals that exceed the bootstrap $\alpha-$level critical value of T_{RRS}. The null hypothesis would already be rejected based on this value alone. In those particular regions, a statistical significant underestimation of the data by the null model is detected at the specified alpha level. Similarly, red areas indicate regions with large negative standardized regional residuals, for which the absolute value exceeds the bootstrap critical value. Such areas indicate statistically significant overestimation. "Non-suspicious" regions are coloured orange. Hence, whenever one white or red spot appears in this formal regional residual plot, the null hypothesis of no lack-of-fit is rejected at the $\alpha-$level of significance, and, in addition, the plot

locates regions of lack-of-fit. These regions can be very small, a few neighbouring observations or even a single outlying observation, or very large in case of global deviations from the null model.

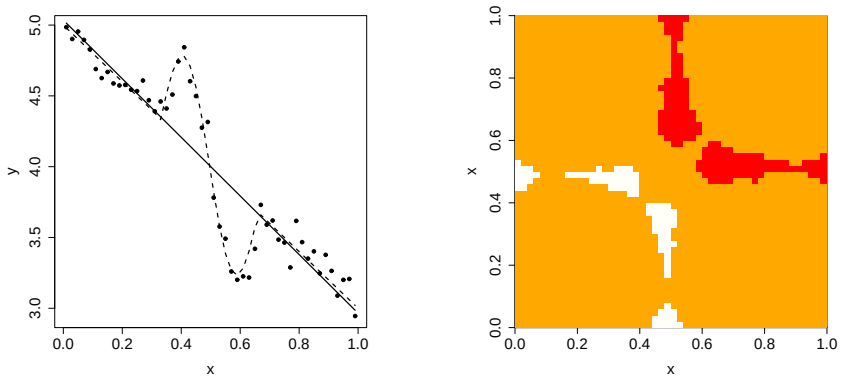

FIGURE 4.2: *(Left panel) True (dashed line) and fitted (solid line) regression model ($n = 50$); LOF situated in [0.33,0.65]; (Right panel) Formal regional residual plot. This plot locates the areas where statistically significant LOF occurs ($\alpha = 0.05$). White areas identify areas of significant underestimation, red of overestimation.*

We now return to our artificial data example for which the formal regional residual plot is shown in Figure 4.2. The heat map in the right panel shows red areas for small intervals around the interval [0.49,0.65] and for larger intervals that include the interval [0.49,0.65] and white areas for small intervals around the interval [0.33,0.49] and for larger intervals that include the interval [0.33,0.49]. This plot shows a statistically significant underestimation of the data around [0.33,0.49], followed by a statistically significant overestimation situated around [0.49,0.65]. This conclusion corresponds to a p-value $p < 0.001$ based on 1000 bootstrap samples. Note that in this particular case, other classical LOF tests are also able to reject the null hypothesis and thus detect a significant LOF, e.g. the von Neumann test ($p = 0.000$), the Buckley test ($p = 0.004$), the OS test ($p = 0.000$), the adaptive Neyman test by Kuchibhatla and Hart (1996) ($p = 0.000$), etc. However, they do not possess the ability to formally locate the deviations in the predictor space.

Added value of the regional residual plots

In the next few paragraphs we illustrate further the added value of both types of plots. Later, in Section 4.2 we give results of a particularly designed empirical study to assess the characteristics of the plots.

To appreciate more the added values of both types of plots, we consider the same deviation from the null model, but we increase the error variance in the artificial data example from $\sigma^2 = 0.01$ to $\sigma^2 = 0.1$. A sample drawn under the new conditions is shown in Figure 4.3. The left panel shows the simulated values of the response variable y and the least squares fit of a simple linear regression model. By inspecting this scatter plot no real indications of LOF are available as no obvious systematic trend is seen in this plot.

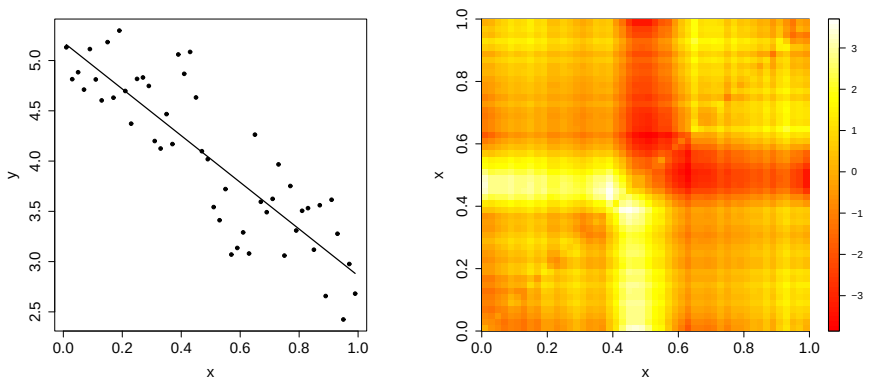

FIGURE 4.3: *(Left panel) True (dashed line) and fitted (solid line) regression model ($n = 50$); lack-of-fit situated in [0.33,0.65]; (Right panel) Exploratory regional residual plot. This plot shows the individual standardized regional residuals to visualize possible areas of LOF.*

For this particular dataset, some of the classical LOF tests show borderline statistical significance at the 5% level, or are even unable to reject the null hypothesis: e.g. the von Neumann test ($p = 0.051$), the Buckley test ($p = 0.529$), the OS test ($p = 0.032$), the adaptive Neyman test by Kuchibhatla and Hart (1996) ($p = 0.061$), etc. Even if the null hypothesis is rejected, the data-analyst has, by purely inspecting the scatter plot in Figure 4.3, no idea whether the model is inappropriate for the entire range of the x-variable or only for one or more small subsets. By inspecting the exploratory regional residual plot (Figure 4.3, right panel) the indication of possible under- and overestimation around the

intervals [0.33,0.49] and [0.49,0.65] immediately becomes clear and is formally confirmed in the formal regional residual plot ($p = 0.001$) in Figure 4.4. The red spots in the formal regional residual plot (Figure 4.4), correspond to four regional residuals calculated over four intervals, ranging from 0.47, 0.49, 0.51 and 0.53 up to 0.63. The light yellow spot corresponds to two regional residuals calculated over two intervals, ranging from 0.39 to 0.43 and 0.45 respectively. This illustrates that the plot fairly well indicates in which region the lack-of-fit is located.

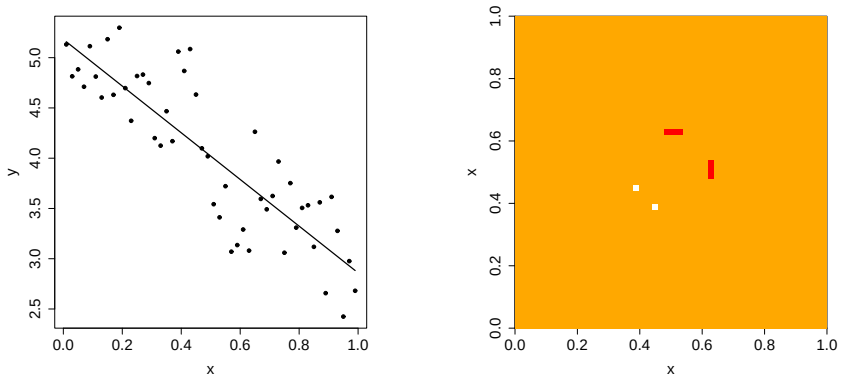

FIGURE 4.4: *(Left panel) The least squares fit (solid line) of the simple linear regression model is added to the scatter plot of y versus x; lack-of-fit situated in [0.33,0.65]; (Right panel) Formal regional residual plot. White areas identify areas of underestimation, red of overestimation.*

A popular exploratory method to assess this fit of a regression model, is to apply a smoother to the data or to the residuals, and add this fit to the scatter plot. This is shown in Figure 4.5 (left panel) with a loess smoother with span=0.75. However, the impression the data-analyst gets, depends on the choice of the smoother and the smoothing parameter. In addition, no formal conclusion can be drawn from these plots. They merely provide the same information as the exploratory regional residual plots. There are LOF tests based on smooths, e.g. the OS test or the adaptive Neyman test of Kuchibhatla and Hart (1996). The latter provides a graph of the smooth estimate of the residuals. An example is shown in the right panel of Figure 4.5. When the null hypothesis is rejected, the graph of the smooth estimate provides an impression of the true nature of the relationship between the residuals and their index. If the graph is non constant, the data-analyst is still not able to formally locate the LOF in the

predictor space.

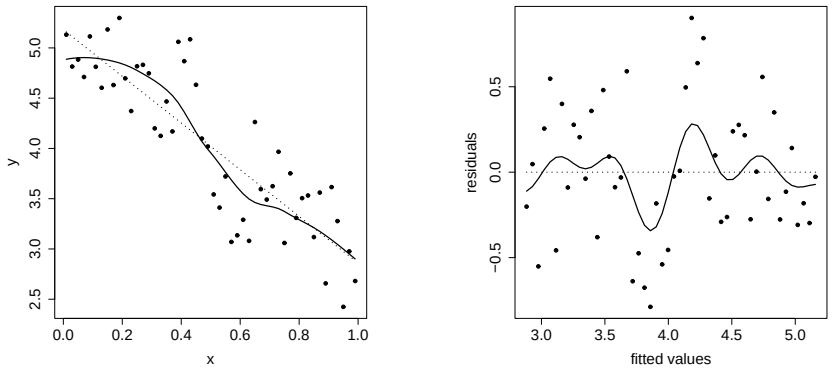

FIGURE 4.5: *(Left panel) Scatterplot of y versus x with the least squares fit (dotted line) and the loess fit with span = 0.75 (solid line) for the artificial data example with* $\sigma^2 = 0.1$. *(Right panel) Scatterplot of the residuals versus the fitted values with the constant line* $y = 0$ *(dotted line) and the trigonometric series smooth fit (3.23) with* $\hat{k} = 9$ *(solid line) for the artificial data example with* $\sigma^2 = 0.1$.

We would like to end the discussion on the regional residual plot with a remark. Regional residual plots are designed to help the data-analyst in identifying areas that deserve special attention. Note that the plots only provide an idea of over- or underestimation of the observations in that specific area. They are not a real tool that suggests how to ameliorate the model. However, one could consider the idea of constructing prototype regional residual plots for different types of lack-of-fit as was done in Lin et al. (2002) for the cusumplot. By comparing a plot obtained for a particular data set with the prototype plots, one could try to recognize a pattern of the prototype plots to get an idea of the particular type of LOF at hand.

4.1.4 Related test statistics

As already indicated in the introduction of this chapter, the proposed test statistic is closely related to those of Stute (1997), Diebolt and Zuber (1999), Lin et al. (2002), among others. Recall that Stute (1997) and Diebolt and Zuber (1999)

studied the process

$$\hat{\mathbb{B}}_n(x) = n^{-1/2} \sum_{i=1}^{n} I(X_i \leq x) e_i.$$

Their test statistic is constructed as the supremum norm of $\hat{\mathbb{B}}_n(x)$. Since the process $\hat{\mathbb{B}}_n$ accumulates all the residuals associated with covariate values less than x, it tends to be dominated by residuals with small covariate values. This problem can be overcome by considering moving sums or moving averages of residuals as proposed by Lin et al. (2002), who also use the supremum norm to obtain a global measure of lack-of-fit. The moving sums of residuals are calculated over blocks of fixed size b,

$$\hat{\mathbb{B}}_n(x; b) = n^{-1/2} \sum_{i=1}^{n} I(x - b < X_i \leq x) e_i.$$

Since the moving sums are based on blocks of the same size, the number of observations in the blocks can be quite different when the covariate values are not evenly distributed. Therefore, also moving averages were studied,

$$\overline{\hat{\mathbb{B}}}_n(x; b) = \frac{n^{-1/2} \sum_{i=1}^{n} I(x - b < X_i \leq x) e_i}{\sum_{i=1}^{n} I(x - b < X_i \leq x)}.$$

The powers of these tests depend on the choice of b. Larger values of b will lead to more powerful tests when a lack-of-fit is situated over the entire range of the predictor variable x (global LOF), while smaller values of b are needed to detect local deviations (Section 4.2) with good power. The method proposed in this thesis solves this problem by considering all possible intervals. Note that our new test thus considers more regions than the tests of Stute (1997), Diebolt and Zuber (1999), Lin et al. (2002) do, as our regional residuals include the cumulative sums of residuals of Stute (1997) and Diebolt and Zuber (1999) and also the moving sums of Lin et al. (2002) for all block sizes. We allow for a non-fixed block size, as in most cases no prior knowledge on the size of the area of LOF is available. We therefore expect our tests to be powerful in case of both global and local lack-of-fit. However, probably this is at a price. If Lin et al. (2002) choose the "right" block size b, their tests will be more powerful. On the other hand, for a "bad" block size, their tests will hardly detect the LOF present.

4.1.5 Difference based variance estimators

As mentioned before, we standardize the regional residuals to make them comparable to each other. Ideally, the variance is known and we obtain immediately the standardized regional residuals and the test statistic. However, in practice,

the residual variance is unknown, and needs to be replaced by a variance estimator that is consistent under both the null and the alternative hypotheses to obtain a powerful LOF test. Unfortunately, the estimator S_n^2, of the residual variance often overestimates under a lack-of-fit situation. The estimated standardized regional residuals appear then to be smaller than they really are, which might result in low power. The use of variance estimators which are more robust against deviations from the null model may therefore be more appropriate. A number of nonparametric variance estimators have been proposed in the literature. We refer to Dette et al. (1998) and Munk et al. (2005) and the references therein for an overview and discussion on variance estimation in nonparametric regression. In what follows, we focus on two popular choices of difference based variance estimators. The first one is half the mean sum of squares of successive differences estimator, $\hat{\sigma}_D^2$, which was introduced by von Neumann (1941) (Equation 3.11), but has also been used by Rice (1984), and is therefore often known as the Rice estimator. The second is the variance estimator $\hat{\sigma}_P^2$ based on pseudo-residuals of Gasser et al. (1986) (Equation 3.6). Both variance estimators are attractive from a practical point of view, as they are computationally simple and often have a small bias for small sample sizes (Dette et al., 1998). When multiple observations at some of the design points are present, appropriate modifications to these nonparametric variance estimators have to be made.
By replacing the residual variance estimator S_n^2 by these estimators, the standardized regional residuals are given by

$$R_D(A_{ij}) = \sqrt{n_{ij}}\frac{R(A_{ij})}{\hat{\sigma}_D h_{ij}} \quad \text{and} \quad R_P(A_{ij}) = \sqrt{n_{ij}}\frac{R(A_{ij})}{\hat{\sigma}_P h_{ij}}.$$

The corresponding test statistics T_{RRD} and T_{RRP} become

$$T_{RRD} = \sup_{i\leq j; i,j=1,\ldots,n} \left|R_D(A_{ij})\right| \quad \text{and} \quad T_{RRP} = \sup_{i\leq j; i,j=1,\ldots,n} \left|R_P(A_{ij})\right|. \tag{4.6}$$

These tests are abbreviated as the RRD and the RRP tests.

4.2 Simulation results

To learn about the small sample power characteristics of the proposed tests in simple linear regression, a simulation study is performed, comparing the empirical powers of the RRS, RRD and RRP tests with those of the closely related tests of Lin et al. (2002) and three classical lack-of-fit tests. The supremum test with cumulative sums of residuals (Stute (1997), Lin et al. (2002)) is abbreviated

as the S test. Since the powers of the tests based on moving sums depend on the choice of b, the fixed block size, three different block sizes are included in the study, corresponding to the range of the lowest 10%, 30% and 50% of the covariate values, which are referred to as the MB10, MB30 and the MB50 tests. The three classical lack-of-fit tests are the generalization of the von Neumann (1941) test, the Buckley's cusum test, and the smoothing-based lack-of-fit test proposed by Kuchibhatla and Hart (1996), denoted as the N, B and KH test, respectively.
In this study, the asymptotic null distributions of the test statistics T_N and T_B are used, while the residual based bootstrap procedure discussed in Section 3.5.2 was used for all other tests. Calculations were performed using R and C++. To reduce the computing time for the estimation of the power of the bootstrap tests, a Monte Carlo power study was set up based on the simple linear extrapolation method proposed in Boos and Zhang (2000). For each scenario, $O = 1000$ data sets are generated under the alternative, resulting in O estimated p-values, $\hat{p}_{I,1}, \ldots, \hat{p}_{I,O}$, each of which is obtained from resampling $I = 59$ times in the bootstrap loop. A linear extrapolation procedure further results in a bias-adjusted power estimate. A sufficiently accurate approximation of the nominal level is observed in all cases (Tables 4.2, 4.3, and 4.5). Typically a larger number of Monte Carlo and bootstrap samples would estimate the power more accurately, but we believe our results are indicative in their comparison of the different tests.
In what follows, three main questions are discussed in the next few subsections:

- How good is the small sample performance of the new tests as compared to other tests, assuming homoscedasticity and Gaussian error terms?
- How do the three new tests perform in case of heteroscedasticity?
- How do they behave for heavy-tailed error distributions?

A fixed, equidistant design with one covariate will be considered.

4.2.1 Homoscedasticity and Gaussian error terms

Two different parametric null models are considered to investigate the first question: the constant mean model and a simple linear regression model. The performance in case of global and local deviations from these null models is studied.

Global LOF

We first reconsider the simulation study of Eubank and Hart (1993), introduced in Chapter 3, Section 3.2.1. In Figure 4.6, our three tests RRS, RRD and RRP

seem to have reasonable power properties in case of the *global* low frequency alternative m_1, and good power properties in case of the *global* high frequency alternative m_2, especially when a nonparametric variance estimator is used. Note that only the two tests that had the highest and lowest power for the specific alternative in Figure 3.7 are plotted in Figure 4.6 so as to keep the plot as lucid as possible. Note that ZCM in the graphs refers to the Cramér von Mises type of test statistic (3.30) based on process $\hat{\mathbb{B}}_n(.)$. We stress once more the fact that the lack-of-fit occurs over the entire range of the predictor variable, and thus represents situations of *global* LOF, in contrast to the next simulation studies where we focus on *local* LOF. In the latter, the deviations from the null only occur in a subset of the predictor variable, while the fit outside that specific region is well-modeled by the parametric null model under study.

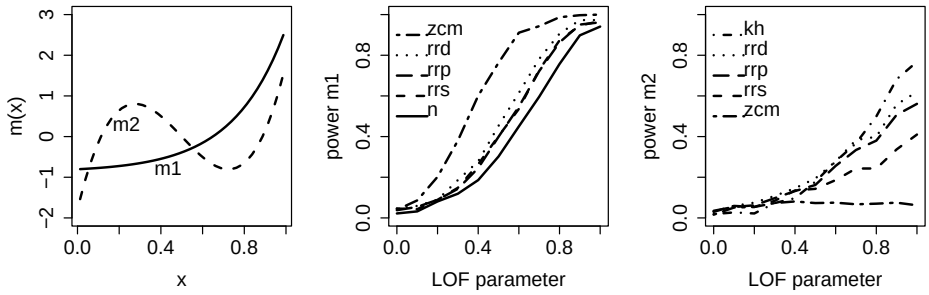

FIGURE 4.6: *(Left panel) Illustration of the low (m_1) and high frequency (m_2) alternative model with parameter $\beta = 1.0$. (Middle panel) Empirical power curves for the different values of the parameter β for m_1. (Right Panel) Empirical power curves for the high frequency alternative m_2.*

Global versus local LOF

As a general setting in the next paragraphs, we consider the null hypothesis of a linear model, $m(x_i, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$ where the vector x of the independent variate is fixed by design and $x_i \in [0,1]$. Continuous data are generated as $Y_i \sim N(m(x_i, \boldsymbol{\theta}), \sigma^2)$ with $\sigma^2 = 0.1$ and with an equidistant design $x_i = (i - 0.5)/n, i = 1, \ldots, n$, for different sample sizes, n = 20, 50 and 100. High frequency alternatives are studied with both larger and smaller regions of LOF, representing global and local lack-of-fit. The lack-of-fit is introduced as

TABLE 4.1: *Estimated powers in case of global and local lack-of-fit (γ = 12.5, 19 and 36, $\lambda = 0.5$, n=50) situated in the lower and mid range of the predictor variable.*

		Test						
location	γ	RRS	RRD	RRP	S	MB10	MB30	MB50
lower	12.5	0.959	0.957	0.930	0.975	0.939	0.972	0.852
	19	0.753	0.773	0.728	0.723	0.742	0.480	0.469
	36	0.252	0.302	0.294	0.135	0.106	0.059	0.070
mid	12.5	0.911	0.936	0.909	0.914	0.923	0.928	0.257
	19	0.611	0.692	0.640	0.523	0.812	0.397	0.024
	36	0.189	0.261	0.256	0.127	0.242	0.111	0.044

one period of a sine function. In particular,

$$m(x_i) = 5 - 2x_i + \lambda \sin(\gamma x_i) I\{\delta_1 \leq i \leq \delta_2\}, \tag{4.7}$$

where the amplitude, $\lambda = 0.10, \ldots, 0.90$, determines the strength of the lack-of-fit. The period, $\frac{2\pi}{\gamma}$, with $\gamma = 12.5, 16, 19, 24$ or 36, determines the length of the interval where the lack-of-fit occurs, varying from global departures ($\gamma = 12.5$) to local departures ($\gamma = 36$). Finally, δ_1 and δ_2 are the lower- and upper bounds of the interval which depend on the period of the sine function and the location of the interval in the predictor range. This is illustrated in the left panels of Figures 4.7 and 4.8, which show some examples of the generated LOF ($\lambda = 0.9$) for large intervals, $\gamma = 12.5$, for medium-sized intervals, $\gamma = 19$ and for small intervals, $\gamma = 36$, in both the low and the mid range of the predictor variable.

Firstly, we discuss a small simulation study that briefly illustrates the disadvantages of the related tests of Stute (1997) and Lin et al. (2002). Secondly, we present the results of a more extended simulation study that illustrates the performance of the new tests in comparison with some classical LOF tests.

Comparison to related tests

In Table 4.1, the empirical powers of the RRS, RRD and RRP tests are compared with those of the closely related tests of Stute (1997) and Lin et al. (2002). We consider both global and local alternatives and as representatives, we selected a lack-of-fit of size $\lambda = 0.5$, which is introduced for three different lengths of intervals: $\gamma = 12.5$ for global departures, $\gamma = 19$, and $\gamma = 36$ for local departures, situated in the lower and mid range of the predictor variable. All

samples had $n = 50$ observations and all tests were performed at the 5% level of significance. The results of the MB tests clearly show the dependence of the power on the choice of the fixed block size b. Larger values of b will lead to more powerful tests when a lack-of-fit is situated over a larger range of the predictor variable, while smaller values of b are needed to detect more local deviations. Also the inferior performance of the S test when the lack-of-fit is situated in the mid-range instead of the lower range of the predictor space can be observed. This could be expected as the cumulative sums of residuals put larger weights on residuals with low covariate values. If the LOF occurs in the mid range of the predictor variable, it is harder to detect the LOF with this test. On the other hand, the RRS, RRD and RRP tests perform well in all cases.
Similar results were found in all other simulations presented further in this section. Therefore, only the results of the RRS, RRD and RRP tests, together with those of the classical lack-of-fit tests will be shown in the remainder of this chapter.

Comparison to classical tests

The main results of the more extended power study with $n = 50$ are visualized using power curves. Figures 4.7 and 4.8 display the estimated power curves for several alternatives. A distinction is made between large, intermediate and small intervals of lack-of-fit and intervals situated at the start of the predictor range (low-range), in the middle (mid-range) or at the end (high-range). Figure 4.7 shows the plots for the low-range, and Figure 4.8 shows those for the mid-range, while the plots for the high-range are not shown as they are similar to those of the low-range.
When comparing the three new and the three classical tests under different conditions of lack-of-fit, the following conclusions can be made. In case of a rather global lack-of-fit (upper panels of Figure 4.7 and 4.8), all tests have good power characteristics, with a slight advantage for the smoothing based KH test, and a rather bad performance of the cusum-based B test in the mid-range. It may be concluded that for rather global departures from the simple linear regression model the power of the new tests are comparable to those of the classical tests.
For lack-of-fit intervals of medium length (middle panels of Figure 4.7 and 4.8), hardly any difference in performance can be seen between the smoothing based KH test and the regional residual based tests. As the length of the interval decreases, it becomes more difficult to discriminate between systematic deviations and noise. In this case, the regional residual based tests have the best power whatever the location, in particular the RRD test. Notice the complete power breakdown of the cusum B test and the poor performance of the KH test. In

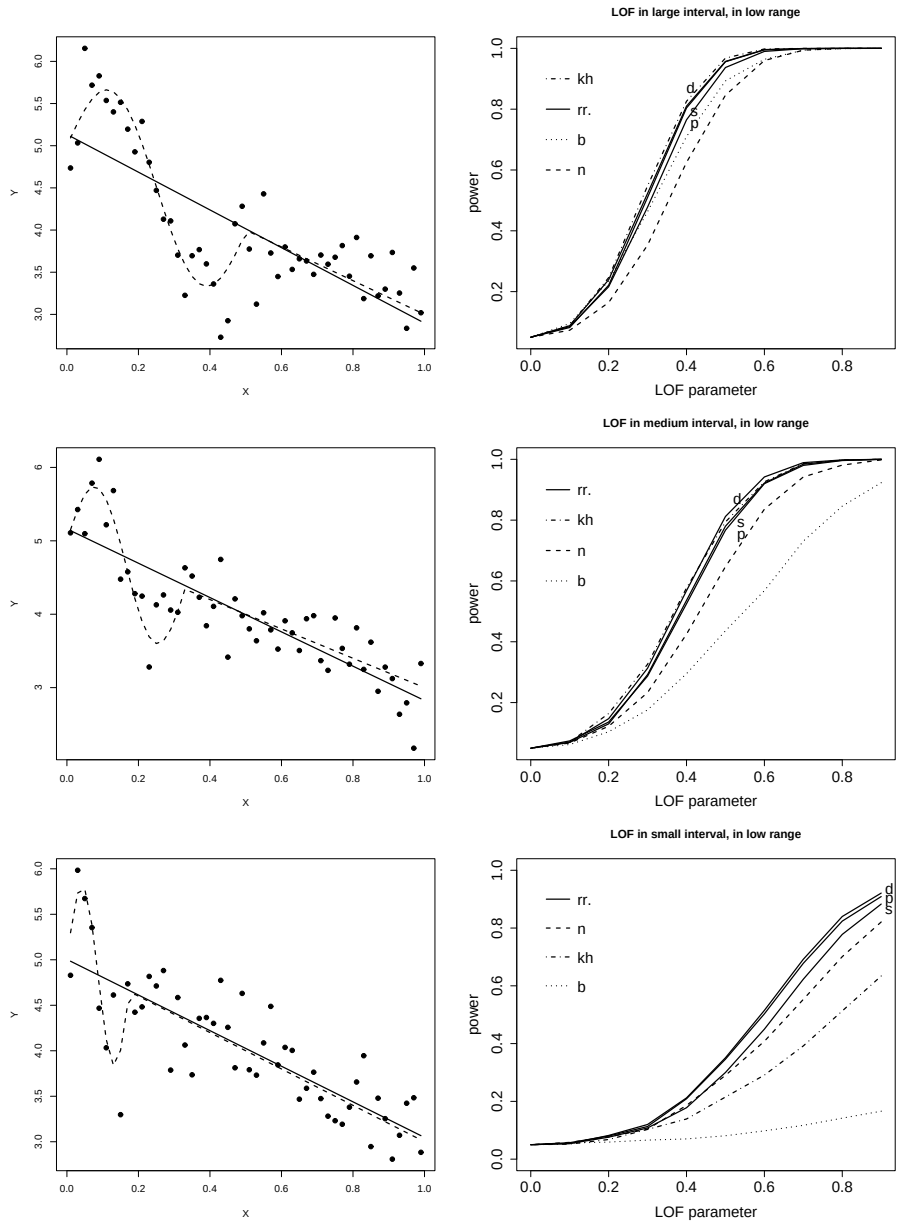

FIGURE 4.7: *(Left panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line, $\lambda = 0.9$) and fitted (solid line) constant mean regression model ($n = 50$); (right panels) Estimated power curves for the N, B, KH tests and the three residual based tests, RRS, RRD and RRP (full lines, last letter is added to the curves to differentiate between them) and for different areas of lack-of-fit; (upper panels) large interval of lack-of-fit situated in the low range ($\gamma = 12.5, \delta_1 = 0.01, \delta_2 = 0.49$); (middle panels) intermediate interval of lack-of-fit situated in the low-range ($\gamma = 19, \delta_1 = 0.01, \delta_2 = 0.31$); (lower panels) small interval of lack-of-fit situated in the low-range ($\gamma = 36, \delta_1 = 0.01, \delta_2 = 0.17$).*

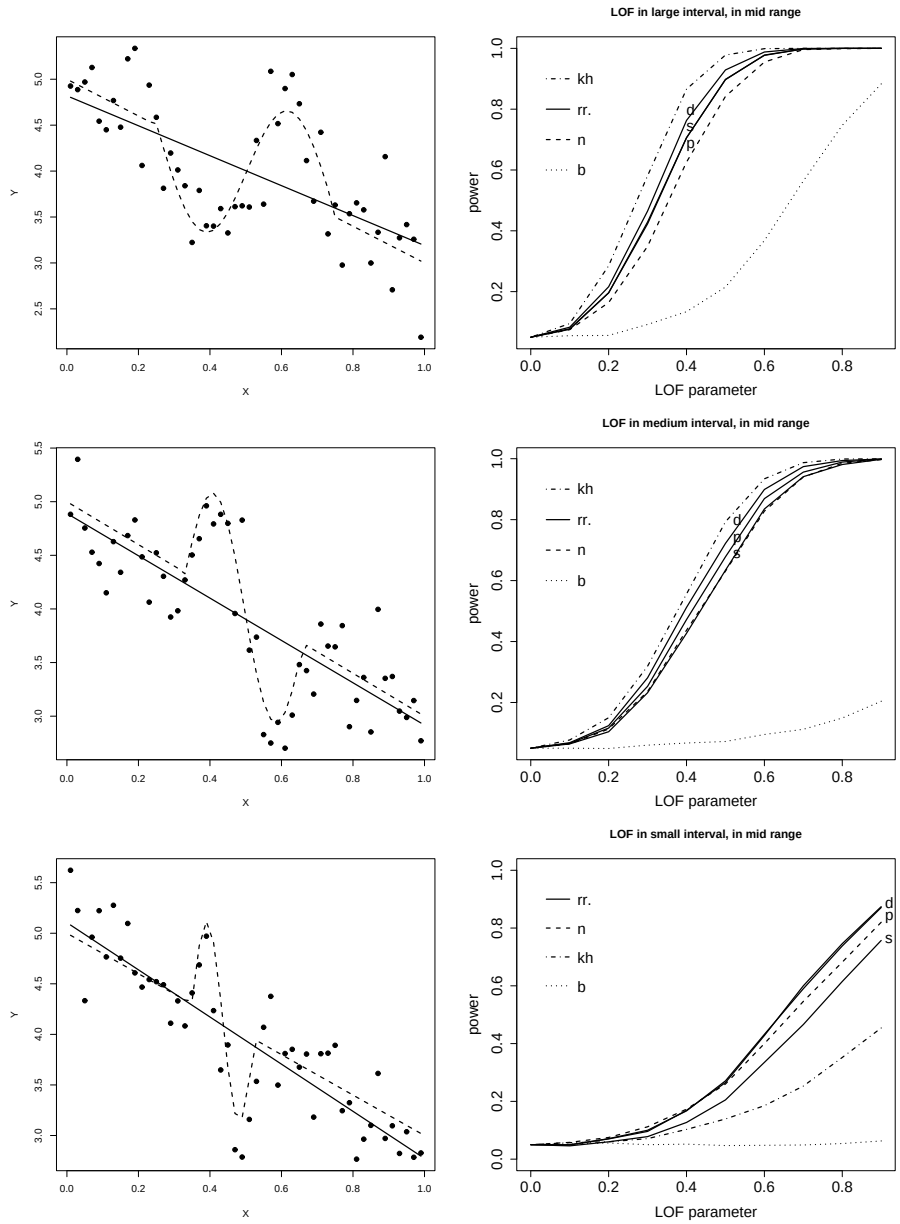

FIGURE 4.8: *(Left panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line, $\lambda = 0.9$) and fitted (solid line) constant mean regression model ($n = 50$); (right panels) Estimated power curves for the N, B, KH tests and the three residual based tests, RRS, RRD and RRP (full lines, last letter is added to the curves to differentiate between them) and for different areas of lack-of-fit; (upper panels) large interval of lack-of-fit situated in the mid-range ($\gamma = 12.5, \delta_1 = 0.25, \delta_2 = 0.73$); (middle panels) intermediate interval of lack-of-fit situated in the mid-range ($\gamma = 19, \delta_1 = 0.33, \delta_2 = 0.65$); (lower panels) small interval of lack-of-fit situated in the mid-range ($\gamma = 36, \delta_1 = 0.35, \delta_2 = 0.51$).*

TABLE 4.2: *Estimated powers with $m(x_i) = 5 - 2x_i + \lambda \sin(\gamma x_i) I\{\delta_1 \leq i \leq \delta_2\}$ for various sample sizes (n=20, 50 and 100), various interval lengths (γ = 12.5, 19 and 36) and in case of no lack-of-fit ($\lambda = 0.0$) and of lack-of-fit ($\lambda = 0.5$) in the low-range.*

			Test					
λ	γ	n	RRS	RRD	RRP	N	B	KH
0.0		20	0.038	0.052	0.046	0.051	0.063	0.063
		50	0.049	0.052	0.052	0.048	0.056	0.050
		100	0.053	0.049	0.051	0.063	0.059	0.066
0.5	12.5	20	0.385	0.539	0.483	0.497	0.371	0.439
		50	0.959	0.957	0.930	0.843	0.894	0.968
		100	1.000	0.999	1.000	0.981	0.996	1.000
0.5	19	20	0.147	0.220	0.238	0.242	0.142	0.201
		50	0.753	0.773	0.728	0.618	0.428	0.759
		100	0.995	0.991	0.990	0.888	0.744	1.000
0.5	36	20	0.038	0.043	0.035	0.037	0.039	0.050
		50	0.252	0.302	0.294	0.255	0.089	0.212
		100	0.726	0.748	0.741	0.546	0.115	0.625

contrast, the power of the three new tests decrease only very slowly with decreasing length of the lack-of-fit interval. This means that for local departures from the simple linear regression model (lower panels of Figure 4.7 and 4.8), our tests perform much better in comparison with the three classical tests.

The general power decrease for LOF that is situated in the mid-range of the predictor variable may be explained by the fact that local deviations around the mean of the covariate x have less influence on the least square fit as compared to deviations near the boundaries of the covariate range, where the design points are high leverage points. This may result in somewhat lower or less extreme residuals, and therefore in somewhat lower power as compared to the same local deviations added in the low- or the high range of the predictor variable.

To study the effect of the sample size, data were simulated with sample sizes 20, 50 and 100. Some results are presented in Table 4.2. The scenario with lack-of-fit strength λ = 0.5 is chosen for this illustration. In general, the previous conclusions seem to remain valid. In particular, the empirical levels are sufficiently close to the nominal significance levels, for all sample sizes. The powers of the three regional residual tests are quite similar, with a minor power advantage of the RRD test, especially in small samples. The power advantage of this test can be explained by the fact that the Rice estimator has the smallest bias in this

particular case (Dette et al., 1998).
In conclusion, the new tests have power similar to that of the KH test and perform better than the N and B tests when rather global lack-of-fit occurs. The powers of regional residual based tests even exceed those of the classical tests in case of local lack-of-fit.

Localization ability of the RR tests

A major advantage of the new procedures, which is the ability of the regional residual plots to formally locate lack-of-fit, is illustrated in Figure 4.9 (left panel). In this graph, each point (i, j) corresponds to a particular interval for which a probability, say P_{ij}, is estimated and plotted. P_{ij} is the probability that the corresponding standardized regional residual is larger than the 5 % critical value of the global test T_{RRD}. This rejection probability is estimated as the ratio of the number of times the standardized regional residual exceeds the simulated critical value of T_{RRD} and the total number of simulation runs (5000). The study was performed under the condition that the lack-of-fit is introduced in two small intervals over the x-range, in $[0.19, 0.35]$ and $[0.79, 0.95]$ with $\lambda = 0.7$ and $\sigma^2 = 0.1$. Figure 4.9 (right panel) shows an example of the local lack-of-fit simulated under these conditions. It is clearly observed in Figure 4.9 (left panel) that mainly the regional residuals calculated over intervals including the area of lack-of-fit, are responsible for the rejection of the null hypothesis. The localization ability of the RR tests is studied more extensively in the next chapter.

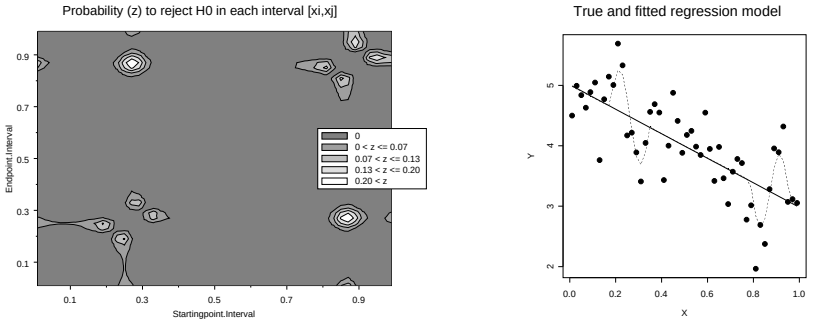

FIGURE 4.9: *(Left panel) Contour plot showing the estimated probability to reject the null hypothesis of no lack-of-fit in each interval $[x_i, x_j]$. (Right panel) True (dashed line) and fitted (solid line) regression model ($n = 50$); local lack-of-fit situated in $[0.19, 0.35]$ and $[0.79, 0.95]$ with $\lambda = 0.7$ and $\sigma^2 = 0.1$.*

4.2.2 Heteroscedasticity and Gaussian error terms

As in the simulation study of Dette and Munk (1998), a simulation study is set up with three different models for the standard deviation of the normal error term, to study the loss of efficiency in using the proposed procedures under heteroscedastic errors,

$$\sigma(x) = \sigma \exp(cx) \quad \text{Monotone, model I} \tag{4.8}$$

$$\sigma(x) = \sigma(1 + c \sin 10x)^2 \quad \text{High frequency, model II} \tag{4.9}$$

$$\sigma(x) = \sigma(1 + cx)^2 \quad \text{Unimodal, model III} \tag{4.10}$$

where different values for c = 0, 0.5 and 1.0 are used, $\sigma^2 = 0.1$.
To deal with heteroscedastic errors, the two wild bootstrap procedures discussed in Section 3.5.3, can be used. When using the popular distribution F_1 as suggested by Mammen (1993) instead of the Rademacher distribution F_2, the size distortion was larger and the power smaller in all cases. We therefore recommend the Rademacher distribution. Table 4.3 shows the empirical sizes with the Rademacher distribution. The empirical power results are presented in Table 4.4. Sufficiently accurate approximations to the nominal level for the bootstrap method are observed in nearly all cases. To make the powers of all tests comparable, all estimated rejection probabilities are based on the wild bootstrap method with distribution F_2. The tests are performed at the 5% level of significance. All possible scenarios of lack-of-fit discussed in Section 4.2.1 are reconsidered here. Only some representative results of rather global and local lack-of-fit are shown in Table 4.4. For all tests, the power clearly decreases with increasing heteroscedasticity. The KH and B tests tend to achieve the best power in case of global lack-of-fit, although the RRD test often performs almost equally well. In case of local lack-of-fit, the RRD and N tests outperform all other tests.

4.2.3 Homoscedasticity and non Gaussian error terms

Finally, the performances of the tests are investigated when dealing with heavy tailed error distributions. To address this issue, the same settings are adopted as in Section 4.2.1, but, as in Dette and Munk (1998), t-distributed error terms with 4 degrees of freedom instead of normally distributed error terms are considered as to obtain the heavy tailed error distribution. The error terms are first rescaled to obtain the same variance as in Section 4.2.1. The estimated powers based on the wild bootstrap are presented in Table 4.5. We conclude that they are similar to those in the homoscedastic case with normal errors, except that some power loss is observed due to the use of the wild bootstrap procedure instead of the

TABLE 4.3: *Empirical sizes for various variance functions I - III for* $\alpha = 0.05$.

		Test					
Model	c	RRS	RRD	RRP	N	B	KH
	0.0	0.049	0.052	0.052	0.048	0.056	0.050
I	0.5	0.053	0.055	0.054	0.054	0.053	0.054
	1.0	0.055	0.054	0.057	0.053	0.053	0.054
II	0.5	0.059	0.068	0.067	0.054	0.048	0.044
	1.0	0.088	0.077	0.072	0.061	0.049	0.043
III	0.5	0.051	0.058	0.059	0.054	0.057	0.058
	1.0	0.056	0.059	0.058	0.057	0.051	0.055

TABLE 4.4: *Estimated powers for various variance functions I - III, in case of global (*$\gamma = 12.5$, $\lambda = 0.5$*) and local (*$\gamma = 36$, $\lambda = 0.9$*) lack-of-fit.*

				Test					
γ	λ	Model	c	RRS	RRD	RRP	N	B	KH
12.5	0.5		0.0	0.970	0.972	0.948	0.853	0.893	0.971
		I	0.5	0.782	0.791	0.740	0.585	0.714	0.793
			1.0	0.375	0.404	0.370	0.317	0.423	0.459
		II	0.5	0.516	0.552	0.492	0.426	0.749	0.665
			1.0	0.211	0.227	0.201	0.177	0.436	0.271
		III	0.5	0.513	0.534	0.496	0.408	0.514	0.579
			1.0	0.155	0.184	0.173	0.179	0.245	0.218
36	0.9		0.0	0.418	0.703	0.661	0.710	0.073	0.600
		I	0.5	0.315	0.497	0.472	0.500	0.071	0.370
			1.0	0.135	0.237	0.234	0.285	0.061	0.188
		II	0.5	0.274	0.344	0.309	0.314	0.058	0.264
			1.0	0.164	0.171	0.159	0.147	0.057	0.130
		III	0.5	0.185	0.302	0.296	0.338	0.069	0.237
			1.0	0.071	0.116	0.126	0.174	0.058	0.114

TABLE 4.5: *Estimated powers for heavy tailed data, using wild F_2 bootstrap, in case of no lack-of-fit ($\lambda = 0.0$) and lack-of-fit ($\gamma = 12.5, 19, 36$, $\lambda = 0.3, 0.6$).*

		Tests					
γ	λ	RRS	RRD	RRP	N	B	KH
	0.0	0.049	0.054	0.056	0.051	0.047	0.053
12.5	0.3	0.490	0.556	0.507	0.408	0.527	0.595
	0.6	0.938	0.969	0.957	0.937	0.950	0.982
19	0.3	0.282	0.364	0.335	0.274	0.201	0.369
	0.6	0.841	0.911	0.869	0.840	0.642	0.907
36	0.3	0.092	0.154	0.147	0.129	0.059	0.111
	0.6	0.271	0.429	0.418	0.458	0.101	0.329

residual based bootstrap. These results thus suggest that the performance of the new tests is quite robust against heavy-tailed error distributions.

4.3 Data examples

The new lack-of-fit tests and corresponding plots are applied to three real data sets from the literature: the windmill, the ice crystal and the Citibase Monthly Indicator data. The use of the regional residual plots to detect and locate the LOF in case of global and local deviations from hypothesized linear and non-linear models is illustrated.

4.3.1 Windmill data

Reconsider the windmill data of Example 4 in Chapter 3. As before, we fit a simple linear regression model to the original, untransformed data. Figure 4.10 shows the scatter plot of the DC output (y) versus the wind velocity x, as well as the formal regional residual plot. The LOF test based on the Rice estimator results in a p-value of $p < 0.0001$, obtained by applying the residual based bootstrap (B=10000), strongly indicating the presence of a lack-of-fit. The formal regional residual plot in the right panel of Figure 4.10 shows a significant ($\alpha = 0.05$) overestimation of the data for small intervals in the low-range of wind velocity. Further, a significant underestimation is found for large intervals, mainly containing design points from the mid-range and even larger intervals, including almost the entire range. It is clear that the overestimation in the low- and high-range, and the underestimation of the data points in the mid-range of the predictor variable are statistically significant. This suggests the presence of a global lack-of-fit.

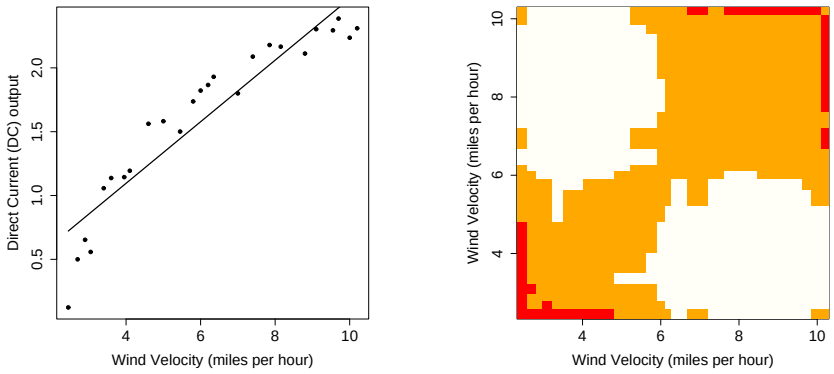

FIGURE 4.10: *(Left panel) Windmill Data (Montgomery and Peck, 1992); y = Direct Current (DC) Output; x = Wind Velocity (miles per hour); (right panel) Formal regional residual plot for the LOF test based on the Rice variance estimator* $(p < 0.00001)$*; red areas indicate an overestimation in the low- and high-range of the wind velocity; white areas an underestimation of the data points in the mid-range.*

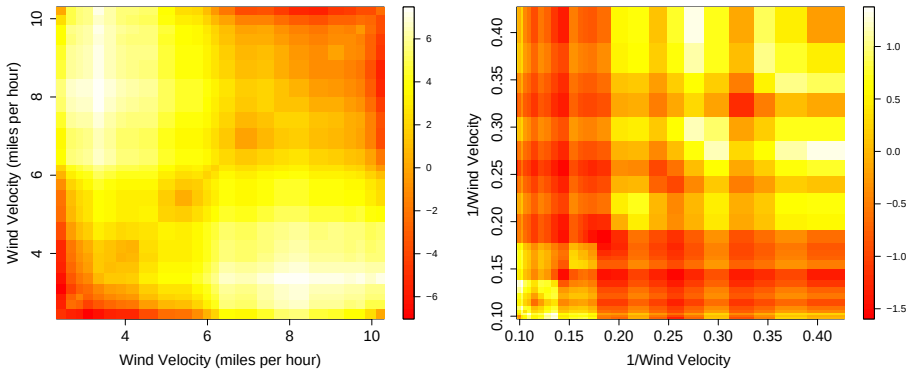

FIGURE 4.11: *Exploratory regional residual plots for windmill data without (left panel) and with (right panel) reciprocal transformation on the wind velocity. White (red) areas correspond to regions of under-(over) estimation of the data when fitting a linear least squares regression model.*

When a linear least squares regression model is fit to the DC Output versus a reciprocal transformation on the wind velocity, the LOF test based on the Rice variance estimator no longer rejects the null hypothesis ($p = 0.85$). This can also

be seen from the exploratory regional residual plot. Figure 4.11 shows these plots for the fitted model based on the original (left panel) and the transformed (right panel) data. In case of LOF, a clear systematic pattern of white and red areas are observed in the regions where the LOF is located, while when no LOF is present, the colours are just scattered around without any systematic pattern as is shown in the right panel. Note that the same colour scheme is used in both plots, but that the scale over which the standardized regional residuals range in the right panel is considerably smaller than that of the left panel. Although this might appear to be confusing, we prefer to use the same colour scheme, as hardly any deviations in colour can be observed when the corresponding colour scheme of the left panel is used. The random colour pattern when no deviations from the null model are present is precisely what we wish to stress.

4.3.2 Ice crystal data

For the ice crystal data set introduced in Example 1 in Chapter 3, we apply the new test to find out whether the local bump that appears in the loess smooth in the left panel of Figure 4.12 corresponds to a significant area of LOF or not. The p-value of the LOF tests based on S_n^2 ($p = 0.149$) do not allow us to conclude a significant local deviation. Although the exploratory regional residual plot shows clear patterns of red and white and light yellow areas, there is not enough evidence in this dataset to conclude that a simple linear regression model for axial length versus times is not appropriate. We would, however, recommend a closer investigation of the model in the highlighted area by the experimenter.

4.3.3 Citibase monthly indicators data

To illustrate the new methodology for nonlinear regression, we assess the fit of two parametric models for the Citibase Monthly Indicators (CITIMON) data set. The data file is available in the SASHELP library of SAS 9.1 and consists of 144 LHUR (unemployment rate) observations from January 1980 to January 1992. The x variable represents the monthly industrial production (IP) index. We could suspect that the unemployment rates are inversely proportional to the industrial production index. Therefore, we assume the following nonlinear parametric regression model, as is done in the SASHELP library

$$m(x, \boldsymbol{\theta}) = \frac{1}{\theta_1 x + \theta_2} + \theta_3.$$

Figure 4.13 shows the scatter plot of the data with the fitted parametric nonlin-

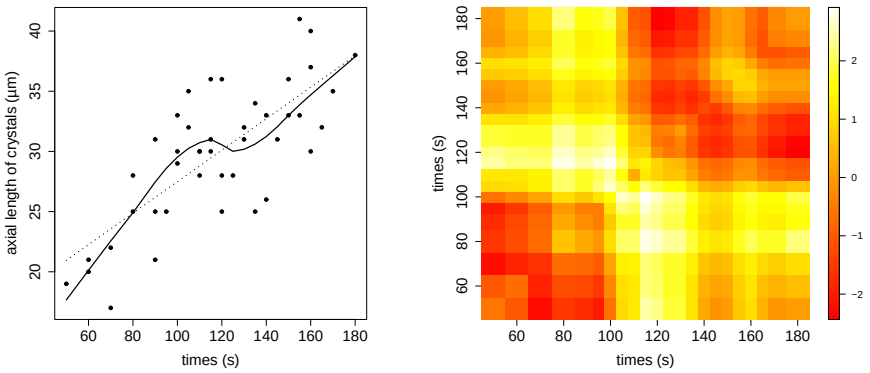

FIGURE 4.12: *(Left panel) Ice crystal data (Ryan et al., 1976); A = axial length of the ice crystal in micrometers; T = times in seconds from the introduction of the crystals. The straight dotted line represents the least squares fit of a linear model, the smoothed line is the fit of a loess smoother to the data (span=0.75). (Right panel) Exploratory regional residual plot for the ice crystal data based on the S_n^2 variance estimator.*

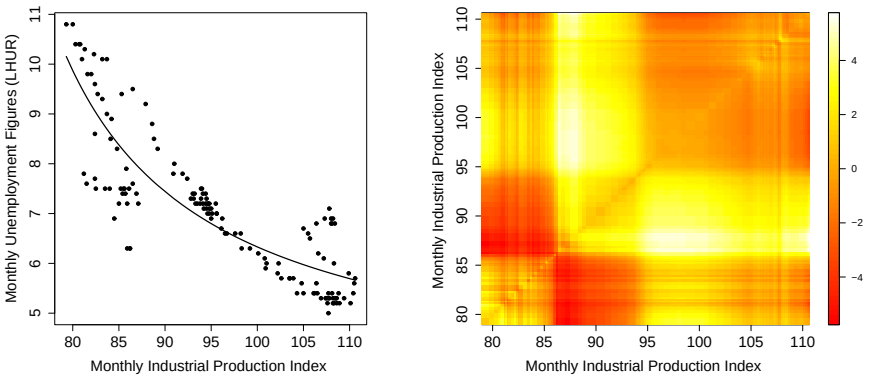

FIGURE 4.13: *(Left panel) CITIMON data (SASHELP library SAS 9.1); y = Monthly Unemployement Figures (LHUR); x = Monthly Industrial Production (IP) index. The solid line is the parametric model fit (Equation 4.11). (Right panel) Exploratory regional residual plot for the Citimon data based on the S_n^2 variance estimator.*

TABLE 4.6: *Empirical levels for the RRS, OS and FH tests for the nominal significance levels ($\alpha = 0.01$, 0.05 and 0.10), obtained by using the wild F_2 bootstrap procedure, for a nonlinear null model (Equation 4.11). The results were obtained by performing 1000 Monte Carlo loops and 100 bootstrap loops for each Monte Carlo loop.*

	Test		
α	RRS	OS	FH
0.01	0.015	0.012	0.012
0.05	0.049	0.048	0.057
0.10	0.092	0.096	0.107

ear model (solid line):

$$m(x, \hat{\boldsymbol{\theta}}_n) = \frac{1}{0.009x - 0.574} + 3.347 \tag{4.11}$$

The nonlinear ordinary least squares estimation method in R is used to estimate the regression parameters. The right panel of this plot shows the values of the standardized regional residuals, based on the S_n^2 residual variance estimate. On this plot we see a systematic pattern in the lower range of the predictor variable suggesting LOF. We obtain a bootstrap p-value of < 0.001 for the RRS test, by approximating the null distribution with the wild F_2 bootstrap procedure (Section 3.5.3) based on 1000 bootstrap replications. Although we have suggested in Section 3.5.4 to use the double bootstrap procedure, we prefer to use the wild F_2 bootstrap procedure instead. In Table 4.6 we present the empirical levels using the wild F_2 bootstrap procedure for a parametric nonlinear null model in Equation 4.11 with normally distributed error terms with mean zero and variance $\sigma^2 = 0.545$ and sample size $n = 144$. The results were obtained by performing 1000 Monte Carlo loops and 100 bootstrap loops for each Monte Carlo loop. Table 4.6 shows a sufficiently accurate approximation of the level of the test when the wild bootstrap is used.

The use of the wild bootstrap procedure, compared to applying the double bootstrap procedure, considerably reduces simulation time for our already computationally intensive LOF procedure. The OS and FH tests (Section 3.2.3) were also applied to the citimon data and also rejected the null hypothesis with both p-values < 0.001, but these tests do not locate the LOF in the x-variable. For the RRS test, the formal regional residual plot is presented in the right panel of Figure 4.14. We find a significant overestimation of the data for rather small intervals that start between an IP value of 79 and 86 and end between 86 and

89. It is clear that the red areas mainly indicate a significant ($\alpha = 0.05$) overestimation in a small area in the low range of the predictor variable around 86, as most red areas intersect in this region. A second region of LOF is identified for larger intervals that start between 86 and 89 and end between 95 and 104 and also end between 108 and 111. The white areas thus indicate a significant underestimation in the mid range, as most white areas intersect in the region between 89 and 95. As the null hypothesis is also rejected for larger intervals in the mid- and upper range of the predictor variable, a small area of LOF occurs in the very upper range of the predictor variable.

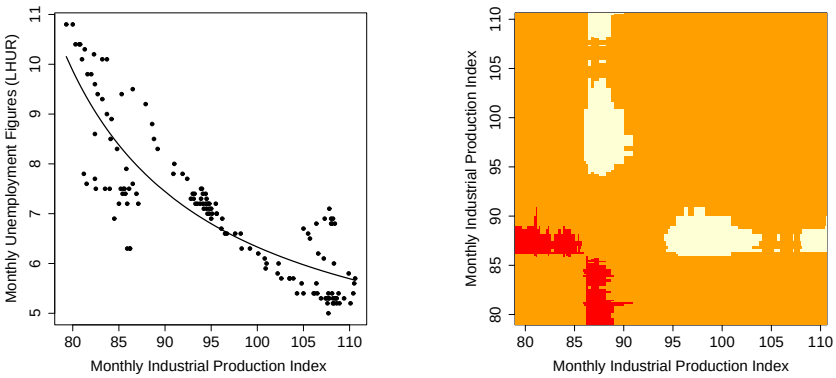

FIGURE 4.14: *(Left panel) CITIMON data; y = Monthly Unemployement Figures (LHUR); x = Monthly Industrial Production (IP) index. The solid line is the parametric model fit (Equation 4.11). (Right panel) Formal regional residual plot (p=0.001) based on the test statistic T_{RRS} and the wild F_2 bootstrap procedure (Section 3.5.3). Red areas indicate a significant ($\alpha = 0.05$) overestimation in the low range and the white areas a significant underestimation in the mid range.*

We continue the discussion with the assessment of a parametric fit that corrects the regions of significant over- and underestimation described above. The x variable, the monthly industrial production (IP) index, is now rescaled such that all points lie within the interval $[0, 1]$. Li (2005) considered the following nonlinear parametric regression model,

$$m(x, \boldsymbol{\theta}) = \exp(\theta_1 + \theta_2 x) + \theta_3 \sin(\pi x) + \theta_4 \sin(2\pi x) + \theta_5 \sin(3\pi x) + \theta_6 \sin(4\pi x),$$

to describe the relationship between the mean LHUR and the IP index.

Figure 4.15 shows the scatter plot of the data with the new fitted model (solid

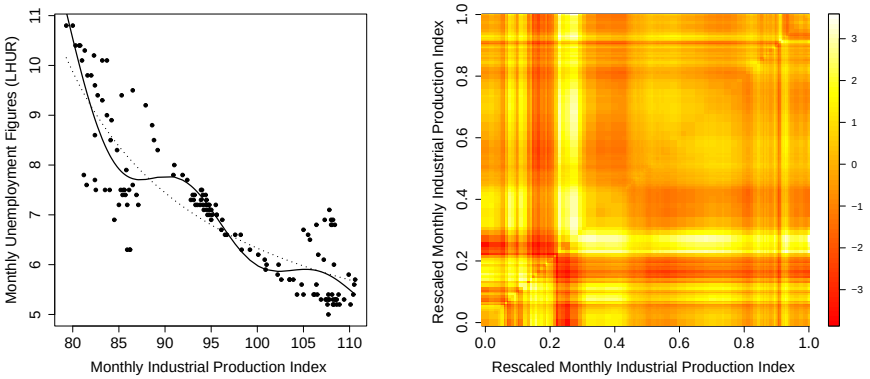

FIGURE 4.15: *(Left panel) CITIMON data; y = Monthly Unemployement Figures (LHUR); x = Monthly Industrial Production (IP) index, rescaled so that all points lie between* $[0, 1]$*. The solid line is the Li model fit (Equation 4.12), the dotted line the SAS model fit (Equation 4.11). (Right panel) Exploratory regional residual plot for the Citimon data based on the* S_n^2 *variance estimator.*

line):

$$m(x, \hat{\boldsymbol{\theta}}_n) = \exp(2.405 - 0.718x) - 1.068\sin(\pi x) - 0.477\sin(2\pi x) \\ - 0.451\sin(3\pi x) - 0.448\sin(4\pi x). \quad (4.12)$$

The fit of the previous model (dotted line) is also added. We see that the new model corrects the regions of significant over- and underestimation described above. The right panel of this plot shows the values of the new standardized regional residuals in an exploratory regional residual plot, based on the S_n^2 residual variance estimate. It still shows a rather systematic pattern in the lower range of the predictor variable. The values suggest first a small area of underestimation around [0.15,0.20], then a small area of overestimation [0.21,0.27], and finally, again a small area of underestimation [0.28,0.43]. However, no significant LOF can be detected at the 10 % significance level or smaller, as the p-value equals 0.136, as approximated by 1000 bootstrap replications. Also, the OS and FH tests have p-values of 0.454 and 0.31 respectively. Although there is not enough evidence in the data to detect a statistically significant LOF, we would recommend the data analyst to investigate more closely the fit in the small areas, highlighted in the exploratory regional residuals plot (Figure 4.15).

4.4 Unstandardized test statistics

Before we conclude this chapter, we add an extra section that actually resulted from findings later in this dissertation. As different variance estimators may influence the performance of the regional residual based tests, and the choice of nonparametric variance estimators is not straightforward in the multiple predictor setting, we may want to consider test statistics based on unstandardized regional residuals. Although it seems unnatural not to standardize the regional residual, most authors prefer this version when tests are based on cumulative sums of residuals, e.g. Stute (1997), Diebolt and Zuber (1999), Lin et al. (2002). One advantage is that the asymptotic theory has a simpler formulation, but secondly, a small simulation study also showed a rather good performance for unstandardized tests. The results of this simulation study are discussed further on in this section. We investigate in this small power study two unstandardized versions. The first one is simply the supremum of the increments of the process $\hat{\mathbb{B}}_n$ studied by Stute (1997) and Diebolt and Zuber (1999). The test statistic is defined as

$$T_{RRUn} = \sup_{i\leq j;i,j=1,\ldots,n} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^{n} e_k I(x_i \leq x_k \leq x_j) \right| = \sup_{i\leq j;i,j=1,\ldots,n} \left| \frac{n_{ij}}{\sqrt{n}} R(A_{ij}) \right|. \tag{4.13}$$

Note that in the test statistic T_{RRUn} the size of the region for which the regional residual is calculated, is not taken into account. This means that residuals that are calculated over large intervals are relatively more important than regional residuals that are calculated over small intervals. As we still want to take the size of the regional residual into account and focus on local LOF, we also consider the test statistic,

$$T_{RRUnij} = \sup_{i\leq j;i,j=1,\ldots,n} \left| \sqrt{n_{ij}} R(A_{ij}) \right|. \tag{4.14}$$

Note that we have to add $\sqrt{n_{ij}}$ in (4.14) to obtain convergence of this test statistic (Chapter 8). Actually this is also necessary for the tests based on standardized regional residuals. However, whether the test statistic would be based on $\frac{\sqrt{n_{ij}}R(A_{ij})}{\mathrm{sd}(\sqrt{n_{ij}}R(A_{ij}))}$ or on $\frac{R(A_{ij})}{\mathrm{sd}(R(A_{ij}))}$, where $\mathrm{sd}(.)$ denotes the standard deviation, does not matter in finite samples as the factor $\sqrt{n_{ij}}$ is canceled out. For weak convergence of the test statistic however, this factor is crucial.

In the next paragraph, we present the result of a small simulation study comparing the small sample performance of the RRS, RRK, RRUn and RRUnij tests when both global and local deviations from the hypothesized model occur. Note that the RRK test refers to the regional residual test with known variance σ^2. We only include the RRS test in this simulation study for its ease of

implementation and its wide applicability.
We consider testing the no-effect hypothesis against an alternative with global, medium-sized and local LOF. The global LOF is represented by

$$m_1(x) = 2.33 + 0.5\lambda \exp\left(\frac{-(x-0.5)^2}{0.06}\right) / \sqrt{2\pi 0.03}, \tag{4.15}$$

where $x_i = (i-0.5)/n, i = 1, \ldots, 72$, and λ is the LOF parameter that ranges from 0 to 1, for which $\lambda = 0$ corresponds to the null hypothesis. To illustrate the performance for the medium-sized deviations from the model, we take

$$m_2(x) = \begin{cases} 2.33 & \text{if } x \notin [0.57, 0.98]; \\ 2.33 + \frac{3}{2}\lambda \sin(4 + 15x) & \text{if } x \in [0.57, 0.98]. \end{cases} \tag{4.16}$$

Finally, a small area of LOF is introduced by adding a period of the sine function to the constant mean model at 2.33. In particular,

$$m_3(x) = \begin{cases} 2.33 & \text{if } x \notin [0.72, 0.88]; \\ 2.33 + 3\lambda \sin(36x) & \text{if } x \in [0.72, 0.88]. \end{cases} \tag{4.17}$$

For each type of LOF, 10000 random data sets of sample size 72 are generated by adding a normally distributed error term with mean zero and standard deviation $\sigma = 0.64$. Scatter plots showing an example of all types of LOF with $\lambda = 1$, are shown in the left panels of Figure 4.16. In the right panels, the estimated powers of the RRS, RRK, RRUn and RRUnij tests are presented. It is remarkable that all four tests perform equally well in case of medium-sized deviations from the hypothesized model, while a distinction in performance is observed between the tests for both global and local LOF. For global LOF, the RRUn test performs better than the RRUnij test. This could be expected, as dividing sums of residuals by $\sqrt{n}$ instead of $\sqrt{n_{ij}}$ results in relatively larger absolute values for large intervals compared to those in small intervals. As global LOF implies large patterns of positive or negative residuals in large intervals, the RRUn test will be more sensitive to global deviations than the RRUnij that puts more weight on regional residuals calculated over small intervals. As a consequence, we find in the lower panel of Figure 4.16 a clear power advantage of the RRUnij test in case of local deviations. The standardized test statistics seem to be a nice compromise between these two unstandardized tests. Also, for the three alternatives studied here, the performance of the test is only weakly influenced by estimating the residual variance. Hardly any power is lost for the RRS test, compared to that of the RRK test. In contrast, we finally show a fourth, global high frequency alternative,

$$m_4(x) = 2.33 + \lambda \sin(36x).$$

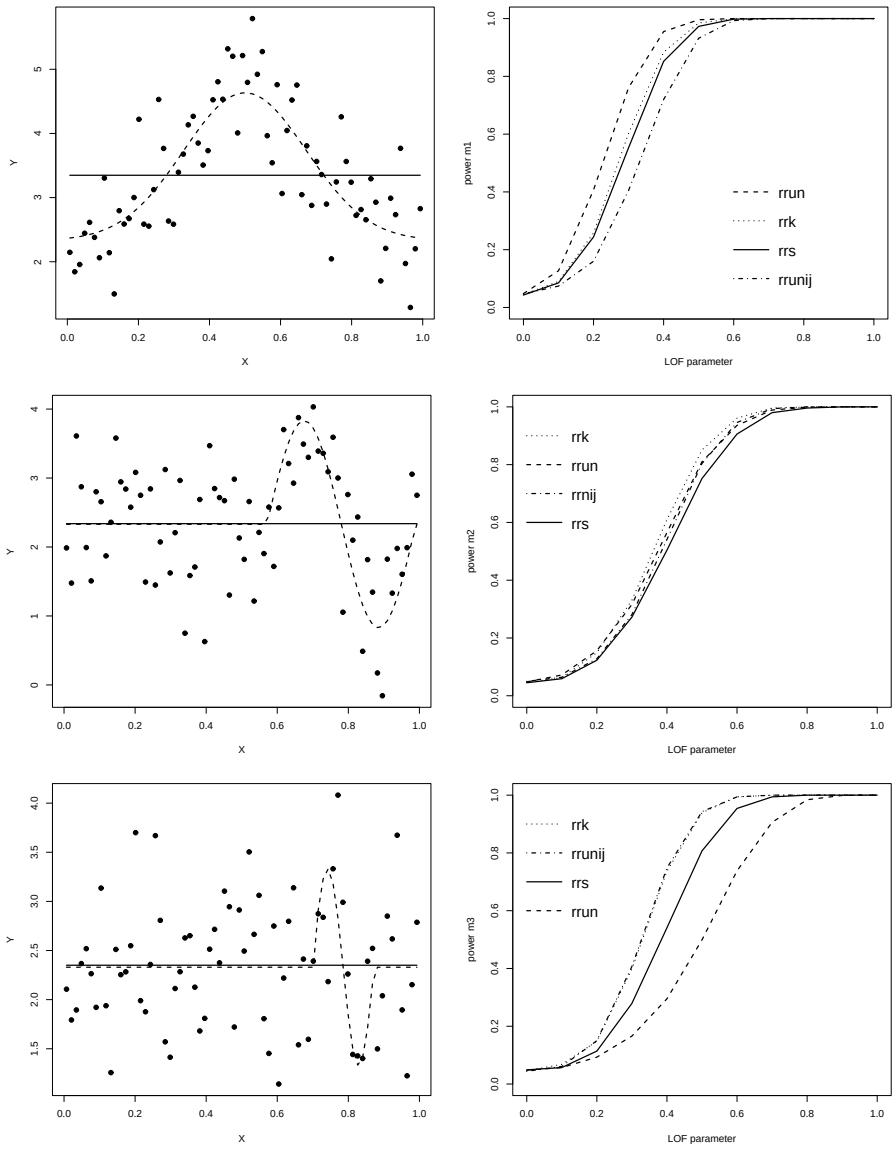

FIGURE 4.16: *(Left panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line, $\lambda = 1$) and fitted (solid line) constant ($n = 72$) mean regression model; (upper panels) global LOF function m_1, (middle panels) medium-sized LOF function m_2 (lower panels) local LOF function m_3. (Right panels) Empirical power curves for the RRS, RRK, RRUn and RRUnij test in function of the LOF parameter λ.*

The LOF is illustrated in the left panel of Figure 4.17 for $\lambda = 1$. Empirical power curves for the RRS, RRK, RRUn and RRUnij tests as functions of the LOF parameter λ are plotted in the right panel. In this case the variance estimator S_n^2 is seriously biased, which results in a RRS test that has no power at all to detect this type of high frequency alternative. As the RRK test performs well, standardizing regional residual using a biased variance estimator has a baleful influence on the performance of the regional residual test. Unstandardized test statistics are most welcome here. As the alternative is periodic in small intervals, a clear power advantage is observed for the RRUnij test in comparison to the RRUn test. Unfortunately, in practice, most likely we do not have an indication of local or global deviations in advance. This small simulation study provides at least some insights into the behaviour of these four tests.

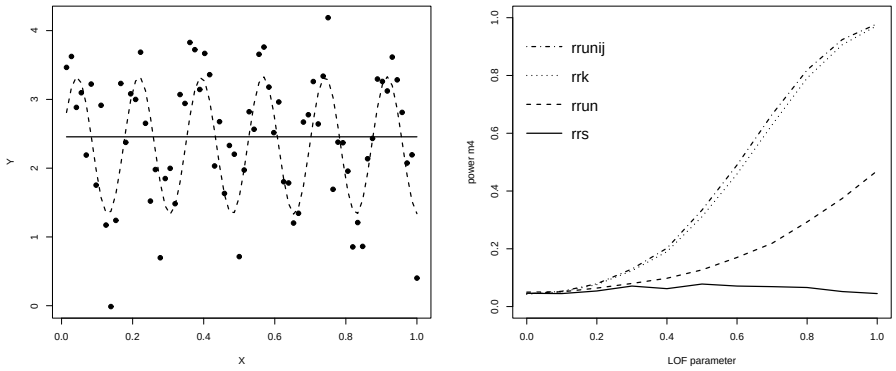

FIGURE 4.17: *(Left panel) Scatter plot showing an example of the simulated high frequency lack-of-fit (dashed line, $\lambda = 1$) and fitted (solid line) mean regression model. (Right panels) Empirical power curves for the RRS, RRK, RRUn and RRUnij test in function of the LOF parameter λ.*

4.5 Conclusions

Lack-of-fit tests and corresponding regional residual plots are proposed to assess the fit of both linear and nonlinear parametric models in a single predictor variable. Simulations suggest that the powers of the proposed testing procedures are at least comparable to the powers of popular classical methods. With the Rice variance estimator good empirical powers are obtained for alternatives with both global and local lack-of-fit. This test seems to behave similarly to the KH test, except for cases with local lack-of-fit, where the proposed tests outperform the classical tests. A major advantage of the new procedures is the ability to locate lack-of-fit in a formal graphical way. Even in situations of violations

of the model assumption of homoscedasticity the new tests still behave well as compared to other classical tests. The use of the wild bootstrap is recommended in practice, as it handles adequately heteroscedasticity and non normality of the error terms.
In the next chapters, extensions to a single circular predictor (Chapter 5), to more than one predictor variable (Chapter 6), and to generalized linear models (Chapter 7) are proposed and investigated. The asymptotic behaviour of the new tests is presented in Chapter 8.

CHAPTER 5

LOF tests and plots for circular-linear regression models

Regression diagnostics and lack-of-fit tests mainly focus on linear-linear regression models, where both the predictor and the response variable have their support on the real line. When the design points are distributed on the circumference of a circle, difficulties arise as there is no natural starting point or origin. Most classical lack-of-fit tests require an arbitrarily chosen origin, but different choices may result in different conclusions. Our methodology in Chapter 4 is easily extended to circular-linear regression models [1], where the predictor variable is measured on a cyclical scale and the response variable on the real line.

5.1 Introduction

In the food industry, micro-encapsulation is used for the isolation of food ingredients, enzymes, cells or other materials, so as to protect them from moisture, heat or other extreme conditions, and thus enhancing their stability and maintaining viability (see e.g. Gibbs et al. (1999)). Ongoing research aims at improving existing techniques to construct a uniform wall around a small sphere, like a food particle. A spray nozzle atomising coating liquid can be used to manufacture microcapsules. The micro-encapsulation data of De Pypere (2005) contains for such an encapsulated food particle, microscopy measurements of the thickness of the coating layer at the circumference of a cross-section of the food particle that were taken at every five degrees. The data are presented in Figure 5.2 (upper panel). The food scientist wants to obtain a quantification of the mean coating thickness and the uniformity of the coating layer around the circumference. He is particularly interested in locating deviations of the mean coating thickness on the circumference of the cross-section of the food particle. To address the research question, the fit of a constant mean regression model, $m(x;\theta) = \theta_0$, for $x \in (0, 360]$ is studied. In the literature, graphical methods and

[1] Most of this chapter is submitted for publication in Deschepper E., Thas O., Ottoy J.P. (2007) *Tests and Diagnostic Plots for Detecting Lack-of-Fit for Linear-Circular Regression Models*. Submitted to Biometrics. In review.

statistical tests used to assess the fit of a parametric regression model mainly refer to cases where the sample space of the predictor is a subset of the real line. However, in the micro-encapsulation data, a random variable is measured on the circumference of a circle, and circular-linear regression analysis is more appropriate. In our example, the response is linear, but one or more predictor variables are angular. Many other examples are available in the literature, e.g. wind direction in relation to the level of a pollutant in an environmental study (Johnson and Wehrly, 1978). Jammalamadaka and Lund (2005) presented an example in which the effect of wind direction on ozone levels was studied. de-Bruyn and Meeuwig (2001) investigated the influence of lunar cycles in marine ecology. The date of birth as a disease indicator was used in Le et al. (2003). Although many case studies are available, the aptness of the model fit has hardly received any attention so far. Maybe this is because classical linear regression analysis can be used to fit these regression models (see e.g. Fisher (1993), and Jammalamadaka and SenGupta (2001)) and therefore ordinary residual analysis is available for the user (see e.g. Johnson and Wehrly (1978)). The use of classical (linear-linear) LOF tests are however often not convenient because they may produce misleading results. Measures of angles depend on the choice of the origin (North, South, etc.) and the sense of rotation (clockwise or counterclockwise). Moreover, angles near zero and near 360 degrees are neighboring directions, so distance measures between angles should be used with care. Statistics for angular data should not depend on such aspects of the data. As this is not the case for the majority of the (linear-linear) LOF tests described in Chapter 3, the p-values of these tests depend on the choice of the origin of the angular variable. Some of the classical LOF tests are appropriate for testing lack-of-fit in circular-linear regression, but, at least to our knowledge, have not been discussed in this context yet. Even more important in the context of this food-industry example, they are not designed for localizing regions in the predictor space where LOF occurs and are thus unable to fully answer the research question.

In this chapter, the methodology developed in Chapter 4 is extended to circular-linear regression. More specifically, a graphical diagnostic tool and a related statistical test to assess the fit of a parametric model in circular-linear regression is proposed, not requiring a natural origin. The method is based on regional residuals which are defined on arcs of a circle instead of on intervals of the real line. The regional residuals plots formally locate and visualize arcs of poorly fitted observations in the circular predictor space. Section 5.2 presents the statistical test. The regional residual plots are constructed and empirically evaluated in Section 5.4. The plots are illustrated on the micro-encapsulation data. For this particular example, the typical problems with many conventional

lack-of-fit tests are demonstrated in Section 5.5. A simulation study shows the performance of the new tests as compared to some classical LOF tests. Finally, some concluding remarks are given in Section 5.6.

5.2 A lack-of-fit test based on regional residuals

In circular-linear regression, the purpose is to fit a regression model to predict the mean of the linear random response variable given a circular predictor variable. Consider n independent pairs (x_i, y_i), with x on a cyclical scale and y a response variable with support on the real line, and a regression model

$$y_i = m(x_i; \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $m(x_i; \boldsymbol{\theta})$ is the conditional mean function, which usually includes the sine and cosine of the angular predictor, instead of the angular variable itself, and $\boldsymbol{\theta}$ is a p-dimensional parameter vector. The error terms ε_i are assumed to be i.i.d. with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. In fact, this is a typical regression model that can easily be fitted by any statistical software package. We assume that $\boldsymbol{\theta}$ is estimated by a consistent estimator, say $\hat{\boldsymbol{\theta}}_n$, e.g. the least squares estimator. Classical residuals, say $e_i = y_i - m(x_i; \hat{\boldsymbol{\theta}}_n)$, are defined as usual.
The methodology developed in Chapter 4 is extended to circular-linear regression by defining the regional residuals on arcs of the circle instead of on intervals of the real line. In particular, we define regional residuals as weighted sums of classical residuals,

$$R(A_{ij}) = \frac{\sum_{k=1}^{n} I(x_k \in A_{ij}) e_k}{\sum_{k=1}^{n} I(x_k \in A_{ij})} = \frac{1}{n_{ij}} \sum_{k=1}^{n} e_k I(x_k \in A_{ij}),$$

calculated over all possible arcs,

$$A_{ij} = \left\{ x \in (0, 2\pi] : x \in \text{ arc } [x_i, x_j] \right\}, \quad (i, j = 1, \ldots, n),$$

where A_{ij} includes the design points x_i and x_j and n_{ij} denotes the number of elements of A_{ij}. Since the sets A_{ij} are defined over all arcs, the collection of regional residuals $R(A_{ij})$ is origin independent.
Large absolute values of standardized regional residuals suggest a possible lack-of-fit of the hypothesized model, located in the corresponding arc on the circle. The generalization of the test statistic of Equation (4.5) is straightforward. To obtain an overall measure of deviation from the hypothesized model, the supremum norm of the standardized regional residuals is again taken as a test statistic. In particular,

$$T_{RRC} = \sup_{i,j} \left| \frac{R(A_{ij})}{\mathrm{sd}(R(A_{ij}))} \right|, \tag{5.1}$$

where sd(.) denotes the standard deviation, or a consistent estimator, and RRC is used to refer to the test based on standardized regional residuals with a circular predictor. This statistic is sensitive to both *global* and *local* deviations from the hypothesized model (Section 5.5.2). Here, global and local refer to large and small arcs in the predictor space, respectively.
As before, the standard deviation in (5.1) may be obtained by straightforward calculations. Let $\mathbf{I}_{A_{ij}}$ denote a $n \times 1$ inclusion matrix, with $\mathbf{I}_{A_{ij},k} = 1$ if $x_k \in A_{ij}$, else 0, and $\mathbf{I}_n$ is the $n \times n$ identity matrix, and let $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ denote the hat matrix. Then, $\text{sd}(R(A_{ij})) = n_{ij}^{-1/2}\sigma h_{ij}$, where $h_{ij}^2 = (\mathbf{I}_{A_{ij}}^t\mathbf{I}_{A_{ij}})^{-1}\mathbf{I}_{A_{ij}}^t(\mathbf{I}_n - \mathbf{H})\mathbf{I}_{A_{ij}}$. Note that the standard deviation depends on the complete design through the hat matrix $\mathbf{H}$. In practice, however, the residual variance σ^2 is unknown. In this chapter, we only consider the estimator $S_n^2 = (n-p)^{-1}\sum_{i=1}^n (y_i - m(x_i;\hat{\boldsymbol{\theta}}_n))^2$, for its ease of implementation and its wide applicability. In the particular case of normally distributed error terms, under the null hypothesis of no lack-of-fit, the standardized regional residuals are again t-distributed with $n-p$ degrees of freedom (Lemma 1 in Chapter 4).

The asymptotic null distribution of T_{RRC} under the no-effects null hypothesis, $m(x;\theta) = \theta_0$, follows immediately from Theorem 2 in Chapter 4. The proof is given in Chapter 8. However, since the convergence is slow, the asymptotic approximation may not be appropriate for small sample sizes. We therefore recommend a bootstrap procedure to obtain approximate p-values. As the model assumptions include homoscedasticity, the ordinary residual based bootstrap (Section 3.5.2) is performed. If this assumption is relaxed, we suggest applying the wild bootstrap procedure (Section 3.5.3).

TABLE 5.1: *Empirical levels obtained by using the residual based bootstrap procedure for several sample sizes (n=24, 36, 60 and 72) and nominal significance levels (α =0.01, 0.05 and 0.10)*

	Sample Size (n)			
α	24	36	60	72
0.01	0.004	0.009	0.009	0.010
0.05	0.036	0.039	0.038	0.042
0.10	0.082	0.078	0.087	0.087

This bootstrap procedure is evaluated for circular-linear regression models in a small simulation study. For sample sizes of $n = 24$, 36, 60 and 72, and nominal

significance levels of α=0.01, 0.05 and 0.10, we have estimated the type I error rate based on 1000 Monte Carlo runs and 1000 bootstrap runs. For the no-effect hypothesis, the results are presented in Table 5.1. We obtain rather conservative empirical levels, but sufficiently close to the nominal significance levels.

5.3 Micro-encapsulation data

We illustrate the new LOF test on the micro-encapsulation data presented in the introduction. Note that for the new methodology to be applicable, independent observations have to be assumed. Microscopy-measurements at five degree intervals were taken to ensure that arcs between subsequent design points are large enough to obtain independent observations. This assumption is confirmed for the micro-encapsulation data by an autocorrelation plot which is shown in Figure 5.1. One lag in the x-axis of this figure corresponds to five degrees on the circle. Although a significant correlation is found between the observations at lag two, we believe that it is reasonable to consider the microscopy-measurements as being approximately independent.

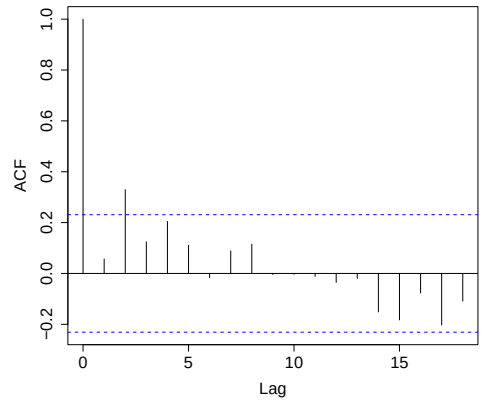

FIGURE 5.1: *Autocorrelogram for the micro-encapsulation data.*

To address the research question, the constant mean regression model, $m(x;\theta) = \theta_0$, for $x \in (0, 360]$ is assessed. For all possible arcs, standardized regional residuals are calculated. We find T = 4.829, corresponding to a bootstrap p-value of 0.001, which clearly demonstrates that the mean thickness of the coating layer varies around the sphere. One of the natural questions raised by the food scientist, is where on the sphere large deviations from the constant mean model are observed. To assist the food scientist in finding these regions,

we have developed regional residual plots. These are presented in the next section.

5.4 Regional residual plots

5.4.1 Construction

The extension of the regional residual plots of Chapter 4 to the circular-linear regression context is immediate. The standardized regional residuals for the micro-encapsulation data can be visualized by plotting them in a heat map (Figure 5.2, left panel). The x-axis (y-axis) of the heat map shows the starting point (end point) of the arc for which the standardized regional residual is calculated. The plot shows white areas for regional residuals calculated over rather small arcs, starting between 200 and 250 and ending between 250 and 300 degrees, which indicates a possible underestimation in this region of the circular predictor variable. Red areas are observed for regional residuals calculated over the complementary arcs, suggesting possible regions of overestimation. As before, the interpretation of this regional residual plot has only a pointwise nature and no formal conclusion can be inferred from them. Instead, a formal regional residual plot, which takes the multiplicity into account, can be constructed by only colouring the arcs for which the absolute value of the standardized regional residual exceeds the bootstrap $\alpha-$level critical value of the test statistic T_{RRC}. This plot is shown in Figure 5.2 (right panel) for the micro-encapsulation data and $\alpha = 5\%$. It formally confirms the conclusion that the mean thickness of the coating layer is significantly larger for small arcs, starting between 200 and 250 and ending between 250 and 300 degrees (white areas), and significantly smaller in the complementary arcs (red areas).

5.4.2 Simulation study

In this section we present the results of an empirical simulation study that aims at illustrating the localization ability of the formal regional residual plots. In a Monte Carlo study we have simulated data under a particular model showing lack-of-fit in a well specified arc, say arc $[a, b]$. The null hypothesis is the no-effect hypothesis, $m(x, \theta) = \theta_0$. For each simulated data set of sample size 72, our test is applied. At rejection we recorded for which arcs A_{ij} the standardized regional residuals exceeded the $\alpha = 0.05$ critical value of the test statistic T_{RRC}. From these simulations, we estimated the rejection probabilities at each arc A_{ij} based on 10000 Monte Carlo and 10000 bootstrap loops. The results are presented in graphs where for each point (x_i, x_j) the estimated rejection probability P_{ij} is plotted. The study was performed under several situations of LOF. The upper panels of Figure 5.3 show the rejection probability plot and the scat-

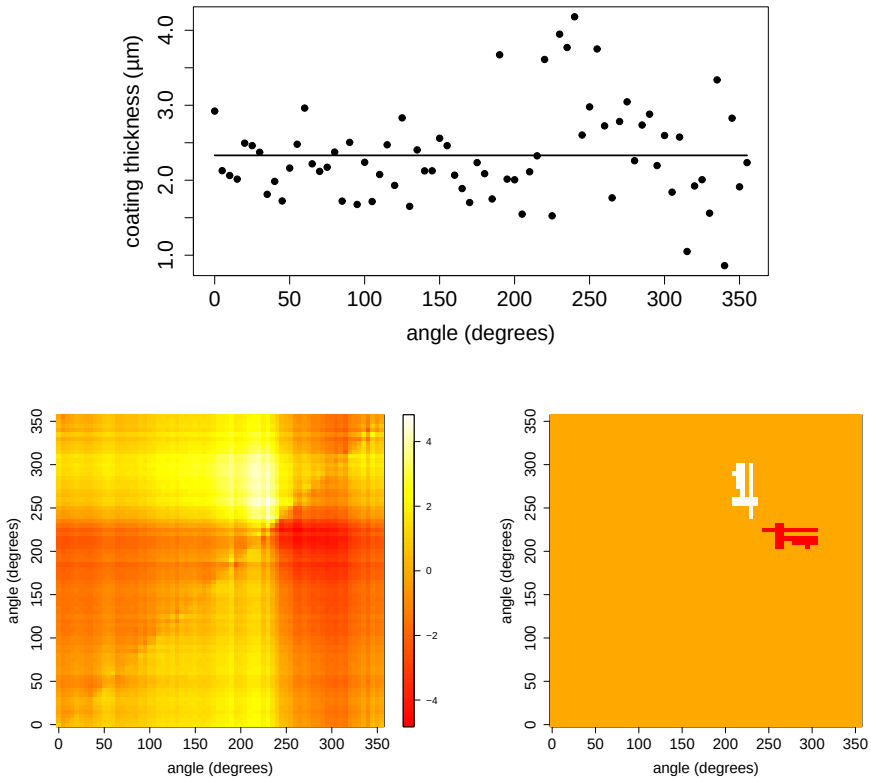

FIGURE 5.2: *(Upper panel) Microscopy measurements of the thickness of the coating layer at every five degrees on the circumference of a cross-section of the food particle. The solid line is the fit of the constant mean regression model. (Lower left panel) Exploratory regional residual plot. (Lower right panel) Formal regional residual plot (p=0.001). White areas correspond to a significant ($\alpha = 0.05$) underestimation of the data, red areas to a significant overestimation. For orange areas no significant deviations are detected.*

ter plot of a random simulated data set, where the true model is the constant mean regression model with mean 0 and σ = 0.6. The estimated power for this artificial data example with no LOF is 6.1%, close to the nominal α-level. The upper left rejection probability plot in Figure 5.3 shows that in case of no LOF all regional residuals could lead to a rare rejection of the null hypothesis. Note that all rejection probabilities are very small ($P_{ij} < 0.0007$). Another study is performed under the condition that the lack-of-fit is introduced in two small intervals over the x-range: in $[40, 95]$ and $[280, 335]$. In these intervals, the function $3\cos(6x) - 3\sin(6x)$ was added to a constant zero mean model. Figure 5.3 (middle right panel) shows an example of the local lack-of-fit simulated under these conditions. The power for this extreme artificial data example is 100%. It is clearly observed in Figure 5.3 (middle left panel) that mainly the regional residuals calculated over intervals including the area of lack-of-fit, are responsible for the rejection of the null hypothesis. The four regional residuals that have the largest probability to reject the null hypothesis correspond to arc $[75, 90]$ ($P_{ij} = 0.865$), arc $[285, 300]$ ($P_{ij} = 0.863$), arc $[315, 330]$ ($P_{ij} = 0.862$), and arc $[45, 60]$ ($P_{ij} = 0.855$) and to their four complementary arcs. Note that for both the regional residuals plots and this rejection probability plot, the graph is symmetric when the null hypothesis corresponds to the constant mean model, but this does not hold for more complex null models, as the standardization of the arcs and their complements may then be different. An example is given in Figure 6.12 (right panel). The largest probabilities of rejection correspond thus to the four arcs where the under- or overestimation is situated. Further, the rejection probabilities are studied in case of global lack-of-fit in the lower panels of Figure 5.3. Many more arcs now have standardized regional residuals that exceed the supremum of the bootstrap critical value. The large dark spot in the lower right corner of the rejection probability plot corresponds to rather large intervals that mainly include the region of overestimation in arc $[260, 80]$. The left upper dark spot includes the complementary arcs.

In Figure 5.4 the localization ability of the formal regional residual plots is investigated in three more situations of LOF. The rejection probabilities are shown in case of local underestimation, and local overestimation, and in case of the combination of both. Local refers here to a small arc $[295, 20]$ (upper panel) and $[205, 290]$ (middle panel) where the function $(\cos(2x) - \sin(2x))$ is added to the true constant mean model. Note the shift in location of the rejection probabilities in the upper and middle panel, corresponding to the shift of the LOF in the predictor range. In the rejection probabilities plots no difference is found between regions of over- or underestimation. Note that this information is available in the regional residual plots by means of the colour scheme. Finally, the lower panel of Figure 5.4 shows the combination of both the local under- and

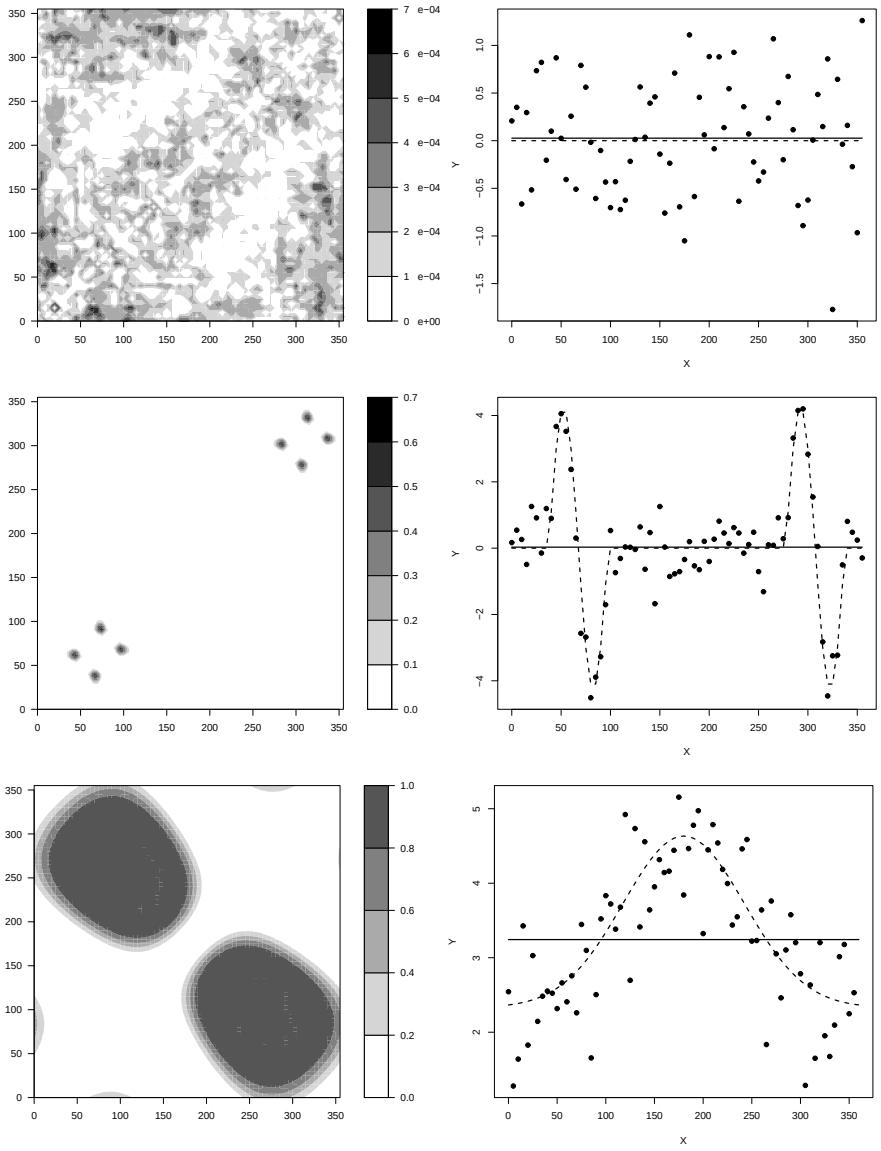

FIGURE 5.3: *(Left panels) Contour plots showing the estimated probabilities to reject the null hypothesis of no lack-of-fit in each arc $[x_i, x_j]$. (Right panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line) and fitted constant mean (solid line) regression model ($n = 72$); (upper panels) no LOF, (middle panels) local lack-of-fit situated in $[40, 95]$ and $[280, 335]$ where $(3\cos(6x) - 3\sin(6x))$ is added to the true mean model, (lower panels) global lack-of-fit; the function $2.33 + \exp(-((x/360 - 0.5)^2)/(2 * 0.03))/\sqrt{2\pi 0.03}$ is added to the constant mean model in the entire x range.*

overestimation, resulting in a larger region of LOF. Arcs that include the LOF region lead to a rejection of the null hypothesis, except for arcs containing both many negative and positive residuals.
This small simulation study convincingly illustrates that the regional residual plots succeed in localizing a lack-of-fit. To complete the discussion we would like to note that by comparing all individual regional residuals with the critical value of the null distribution of the supremum test statistic T_{RRC}, which is the maximum of all standardized regional residuals, we show slightly too conservative plots.

5.5 Evaluation of LOF tests in circular-linear regression

Although ordinary least squares regression can be used to fit circular-linear regression models, classical LOF tests for linear-linear regression models may fail to detect lack-of-fit properly. To learn about the differences in performance and applicability between classical LOF tests and the regional residual based test, we have first applied several tests to the micro-encapsulation data (Section 5.5.1). In Section 5.5.2 empirical powers are compared in a simulation study.

5.5.1 Applicability of LOF tests in circular-linear regression

We first describe some "classical" LOF tests from Chapter 3 in the context of circular-linear regression. We include the supremum test of Stute (1997) and Lin et al. (2002), the S test, and Buckley's (B) test, which are all based on cumulative sums of residuals. When the predictor is angular, these cumulative sums of residuals are sums within arcs with starting point equal to the origin. A third test is the generalization of the von Neumann test, described by Hart (1997) (the N test). Finally, two smoothing based LOF tests are considered. The first one is Hart's order selection test with sine series, rather than a cosine series as it is presented in e.g. Hart (1997). The sine series should make the test more origin-independent. This test is referred to as the OS test. Finally, we mention the data-driven Neyman smooth test of Fan and Huang (2001), which uses both a sine and cosine series estimator. We refer to this test as the FH test. The combination of sines and cosines makes their test origin independent, though this was not recognized by Fan and Huang as they only considered linear-linear regression in their paper. The sines and cosines combination also appears in the components of the Watson (1961) goodness-of-fit test for circular uniformity (Shorack and Wellner, 1986).

For the micro-encapsulation data, Figure 5.5 illustrates the dependence of the p-values for all tests on the choice of the origin. The asymptotic null distribution

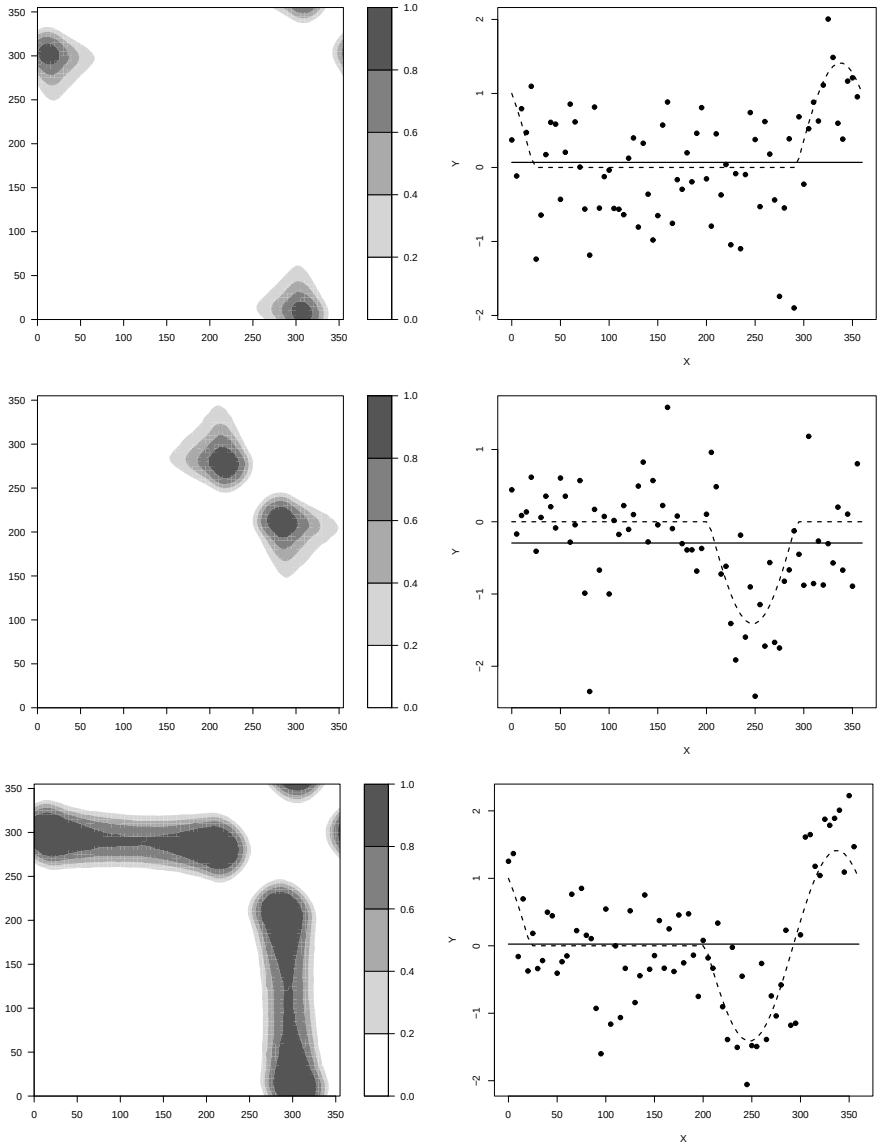

FIGURE 5.4: *(Left panels) Contour plots showing the estimated probabilities to reject the null hypothesis of no lack-of-fit in each arc $[x_i, x_j]$. (Right panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line) and fitted (solid line) constant mean regression model ($n = 72$); the function $(\cos(2x) - \sin(2x))$ is added to the constant mean model in $[295, 20]$ (upper panels), in $[205, 290]$ (middle panels) and in $[205, 20]$ (lower panels).*

is used for the N and B tests, while the bootstrap procedure described in Section 3.5.2 is used for all other tests. The horizontal line at $p = 0.001$ connects the p-values of the regional residual test, confirming that the RRC test is origin independent. Tests based on cumulative sums of residuals (the B and S tests), on the other hand, show a strong dependence on the choice of the origin. The S test, for example, only considers the supremum of cumulative sums with respect to the origin. As a consequence, some origins result in a failure to reject the null hypothesis, whereas others result in very small p-values corresponding to significance at the 5% level of significance. The N test also shows varying p-values, because the variance estimators vary with changing starting points. For this particular data set, the test fails to reject the null hypothesis. The OS test seems to be more or less origin independent, as only small variations occur in the p-values of this test. The horizontal line at $p = 0.014$ connects the p-values of the origin independent FH test.
In conclusion, this figure clearly illustrates the drawback of using classical LOF tests in circular-linear regression and the need for specific solutions. Although not discussed in this context yet, the FH and OS tests are also suitable for testing LOF on the circle, but they are not designed for localizing LOF.

5.5.2 Power study

To fully appreciate the performance of the RRC, FH and OS tests in case of both global and local deviations from the null model, a simulation study is set up. We have chosen a null model under the conditions of the food industry example so that the reader gets an idea of how much of an effect would be needed to be reliably detected by the RRC, FH and OS tests. As an example for global lack-of-fit, we used the regression function

$$m_1(x) = 2.33 + 0.5\lambda \exp\left(\frac{-(x/360 - 0.5)^2}{0.06}\right) / \sqrt{2\pi 0.03},$$

where x is in $(0, 360]$ and λ is the LOF parameter that ranges from 0 to 1, for which $\lambda = 0$ corresponds to the null hypothesis. To illustrate the performance for medium-sized and local deviations from the model, we consider three functions. The first regression function shows a LOF in a medium-sized interval that ranges over more or less half of the predictor space. One period of a sine function is added to the constant mean model at 2.33. Observations are generated with the regression function

$$m_2(x) = \begin{cases} 2.33 & \text{if } x \notin [205, 350]; \\ 2.33 + \frac{3}{2}\lambda \sin(4 + 15\frac{x}{360}) & \text{if } x \in [205, 350]. \end{cases}$$

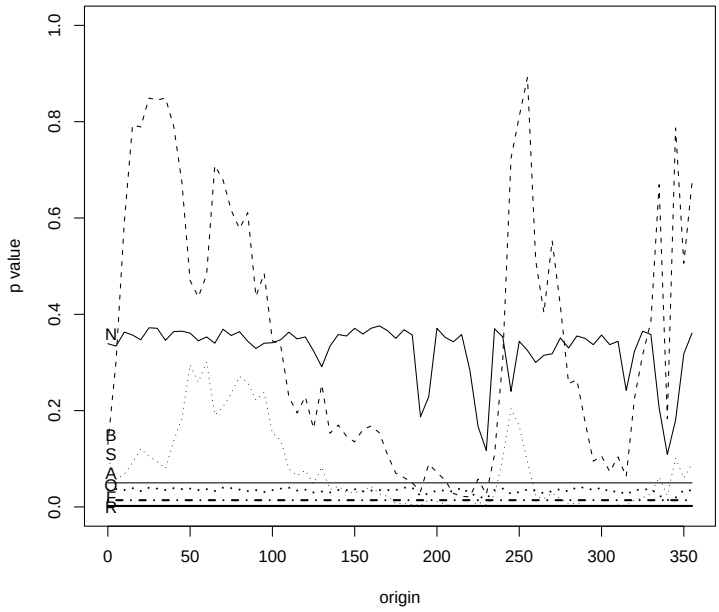

FIGURE 5.5: *p-value plot for the micro-encapsulation data. Tests that heavily depend on the choice of the origin are plotted in thin lines (N = solid line, B = dashed line, S = dotted line). The more or less origin independent OS test is plotted in thick lines (OS = dotted (O) line). The horizontal lines refer to* $\alpha = 0.05$ *(label A), the p-value of the R = Regional Residual test (*$p = 0.001$*), and the p-value of the FH test,* $p = 0.014$ *(dashed-dotted (F) line).*

In a next example, a small deviation occurs over an interval that has about half the width of the one in the previous example. We consider

$$m_3(x) = \begin{cases} 2.33 & \text{if } x \notin [255, 310]; \\ 2.33 + 3\lambda \sin(36\frac{x}{360}) & \text{if } x \in [255, 310]. \end{cases}$$

Finally, a small area of LOF is introduced by adding half a period of a sine function to the constant mean model at 2.33. The local LOF thus only includes an area of underestimation of the true regression function. In particular,

$$m_4(x) = \begin{cases} 2.33 & \text{if } x \notin [210, 280]; \\ 2.33 + 2\lambda \sin(4 + 15\frac{x}{360}) & \text{if } x \in [210, 280]. \end{cases}$$

For each type of LOF, 5000 random data sets of sample size 72 are generated by adding a normally distributed error term with mean zero and standard devia-

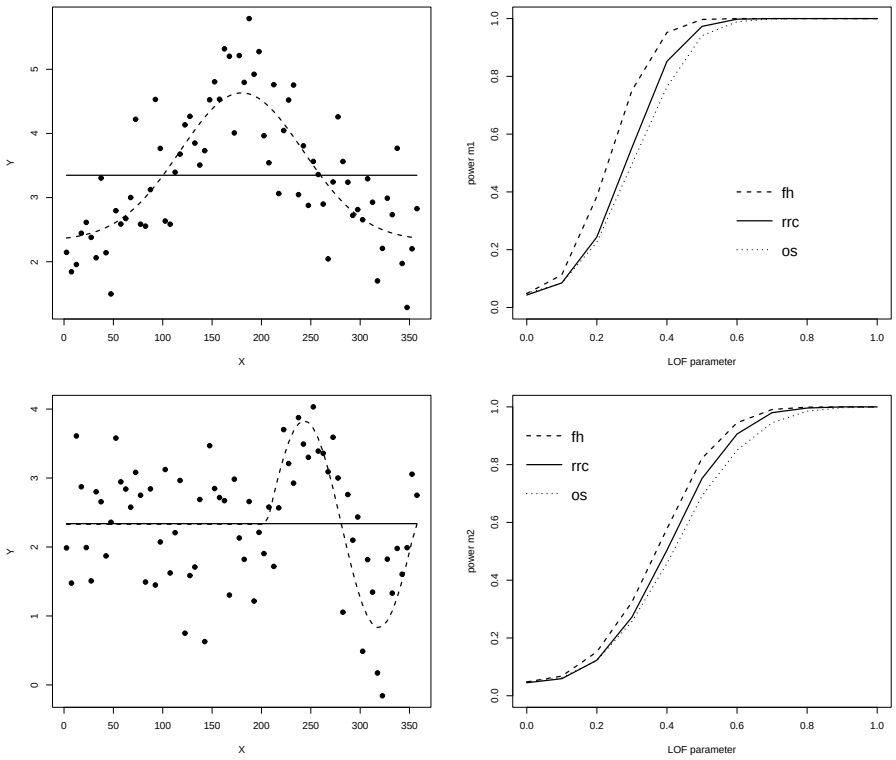

FIGURE 5.6: *(Left panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line) and fitted (solid line) constant mean regression model; (upper panels) global LOF function m_1, (lower panels) local LOF function m_2. (Right panels) Power curves for the RRC, FH and OS test in function of the LOF parameter λ.*

tion $\sigma = 0.64$. Scatter plots showing an example of all types of LOF with $\lambda = 1$, are shown in the left panels of Figures 5.6 and 5.7. All tests are performed at the 5% level of significance. The estimated powers are shown in the right panels. We conclude that the performance of all three tests is good for both global and local lack-of-fit. None of the tests is uniformly better than the others. For example, the FH test is more powerful in the global LOF case and the RRC test in the local LOF case.

5.6 Conclusions

Although ordinary least squares regression can be used to fit circular-linear regression models, classical LOF tests for linear-linear regression models often

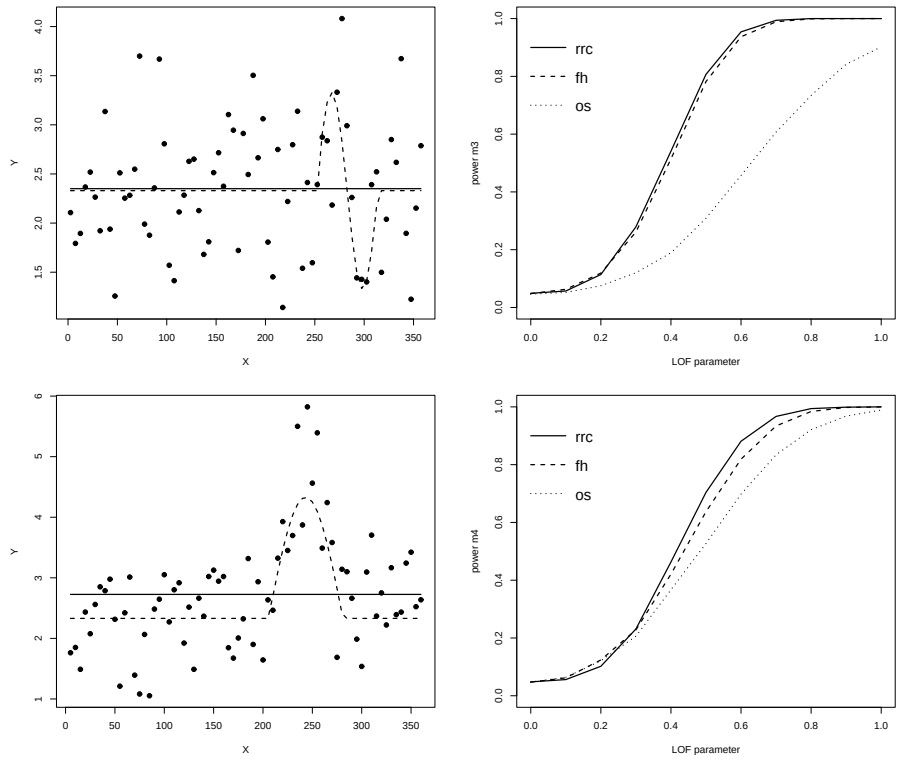

FIGURE 5.7: *(Left panels) Scatter plots showing an example of the simulated lack-of-fit (dashed line) and fitted (solid line) constant mean regression model; (upper panels) local LOF function m_3, (lower panels) local LOF function m_4. (Right panels) Power curves for the RRC, FH and OS test in function of the LOF parameter λ.*

fail to detect deviations from the hypothesized model because their p-values strongly depend on the choice of the origin of the circular variate. We have proposed the regional residual test to properly detect lack-of-fit on the circle. This test is origin independent. We have also illustrated that regional residuals can be used to construct a regional residual plot. Combined with the testing procedure, this graphical diagnostic tool allows both global and local deviations to be detected and localized in the predictor space.
We have also observed good powers for the smooth test of Fan and Huang (2001), which is also origin independent. This latter feature, however, has not been recognized before.

CHAPTER 6

Regional residuals for multiple regression models

As most lack-of-fit tests, the proposed regional residual tests of Chapter 4 depend on an order relation of the residuals. In the univariate case, such an order is obvious, but in the case of two or more covariates, it is not straightforward to order a multivariate vector. The discussions on this problem in the literature are limited (Barnett (1976), Kuchibhatla and Hart (1996), Fan and Huang (2001), Lin et al. (2002), among others). For $\mathbf{x} \in \mathbb{R}^d$, the order relation may be defined as $\mathbf{x}_i \leq \mathbf{x}_j$ if

- all components of $\mathbf{x}_i$ are smaller than or equal to those of $\mathbf{x}_j$, this means $x_{ik} \leq x_{jk}$, for all $k = 1, \ldots, d$,
- the k^{th} component of $\mathbf{x}_i$ is smaller than or equal to that of $\mathbf{x}_j$, thus $x_{ik} \leq x_{jk}$ for a specified k,
- $s_i \leq s_j$, where s_i is the score of a specified function of $\mathbf{x}_i$, e.g. the first principal component.
- $\hat{y}_i \leq \hat{y}_j$, where $\hat{y}$ denotes the predicted values of the fitted regression model.

In what follows, we will discuss two possible extensions of the proposed tests and plots to multiple regression. Firstly, we construct marginal test statistics in Section 6.1 [1] by applying the previous tests with respect to each of the k predictor variables separately, taking the second definition of the order relation into account. We consider a global test statistic based on the supremum of all marginal test statistics. Marginal regional residual plots for each variable allow detection of lack-of-fit and in which variables, and where the lack-of-fit occurs. In a second approach, we adapt the definition of the regional residuals by

[1] Most of this section is published in Deschepper E., Thas O., Ottoy J.P. (2006) *Regional Residual Plots for Assessing the Fit of Linear Regression Models.* Computational Statistics and Data Analysis, 50, 1995-2013.

considering a distance measure in the predictor space. This procedure avoids choosing an order of the residuals in advance, or choosing a smoothing parameter (Section 6.2). This test has nice power properties, but is computationally rather heavy. We construct a new type of regional residual plot based on the adapted definition of the regional residuals. It keeps its formal interpretation and its ability to locate lack-of-fit in the predictor space.
Finally, in Section 6.4 the extension of the tests in Chapter 5 is discussed when one or more variables are angular. We end this chapter by summarizing some conclusions.

6.1 Marginal lack-of-fit tests and plots

6.1.1 Multiple regression

Consider the multiple predictor variable $\mathbf{x} \in \mathbb{R}^d$, and let $m(\mathbf{x})$ again denote the parametric regression model for the mean of the response variable y,

$$y_i = m(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_i's are i.i.d. random variables with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. Recall the null hypothesis,

$$H_0 : m \in \mathcal{M} = \{m(., \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\},$$

where Θ is a p-dimensional proper parameter set in $\mathbb{R}^p$. The residuals are defined by $e_i = y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)$, where $\hat{\boldsymbol{\theta}}_n$ is assumed to be a consistent estimator of $\boldsymbol{\theta}$, e.g. the least squares estimator.

6.1.2 Marginal regional residuals

Marginal regional residuals with respect to the k^{th} covariate x_k are defined as the average of residuals in the subset $A_{kij} = [x_{ki}, x_{kj}]$, $i \leq j$; $i, j = 1, \ldots, n$, (see for example rectangle in Figure 6.1),

$$R(A_{kij}) = \frac{\sum_{l=1}^{n} e_l I(x_{ki} \leq x_{kl} \leq x_{kj})}{\sum_{l=1}^{n} I(x_{ki} \leq x_{kl} \leq x_{kj})} = \frac{1}{n_{kij}} \sum_{l=1}^{n} e_l I(x_{ki} \leq x_{kl} \leq x_{kj})$$

where n_{kij} is the number of observations in the subset A_{kij}, and the design points are ordered with respect to the k^{th} covariate x_k. Of course, other directions can be investigated in the same way, e.g. principal components or fitted values.
Under the null hypothesis of no lack-of-fit, these regional residuals have zero mean. The expression for the variance is similar to that for simple regression

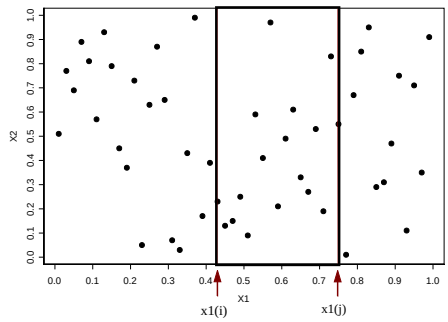

FIGURE 6.1: *Example of the subset A_{kij} when the order is defined according to one covariate at a time.*

analysis. Let $\mathbf{H}$ denote the hat matrix, and $\mathbf{I}_{A_{kij}}$ is a $n \times 1$ inclusion matrix, with $\mathbf{I}_{A_{kij},l} = 1$ if $x_{kl} \in A_{kij}$, else 0, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The variance of $R(A_{kij})$ under the null hypothesis is then given by $n_{kij}^{-1}\sigma^2 h_{kij}^2$, where $h_{kij}^2 = (\mathbf{I}_{A_{kij}}^t \mathbf{I}_{A_{kij}})^{-1}\mathbf{I}_{A_{kij}}^t(\mathbf{I}_n - \mathbf{H})\mathbf{I}_{A_{kij}}$. For a linear model, the expression of the hat matrix is given by Equation 4.3, for a nonlinear model by Equation 4.4.

Standardized marginal regional residuals are obtained by replacing the unknown residual variance σ^2 by the natural estimator $S_n^2 = (n-p)^{-1}\sum_{i=1}^n (y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n))^2$, resulting in

$$R_{S_n^2}(A_{kij}) = \sqrt{n_{kij}}\frac{R(A_{kij})}{Sh_{kij}}.$$

Nonparametric variance estimators are available in the literature, but will often heavily depend on the order relation for the observations or the choice of subsets in the predictor space. The interested reader is referred to Hall et al. (1991), Kulasekera and Gallagher (2002), Munk et al. (2005), and Tong and Wang (2005), among others, for more details. However, in later simulation studies and in the discussion of data examples, only the natural estimator is considered for its ease in computation and for a fair comparison among tests.

6.1.3 A lack-of-fit test

For all possible intervals of the k^{th} covariate x_k, $A_{kij} = [x_{ki}, x_{kj}]$, $i \leq j$; $i, j = 1, \ldots, n$, the standardized marginal regional residuals are calculated. Large absolute values of these standardized regional residuals indicate a possible lack-of-fit. To overcome the problem of multiplicity and to obtain a global measure

of lack-of-fit, the supremum norm of all the standardized regional residuals is proposed as a test statistic,

$$T_{k,S_n^2} = \sup_{i \leq j} \left| R_{S_n^2}(A_{kij}) \right|.$$

This test statistic only contains marginal information on lack-of-fit with respect to the k^{th} covariate x_k, but they can be further combined into one global test statistic T_{gl}, defined as the supremum of the d marginal statistics $T_{k,S_n^2} (k = 1, \ldots, d)$,

$$T_{gl} = \sup_{k=1,\ldots,d} (T_{k,S_n^2}).$$

If one is specifically interested in one covariate or if one has prior information that LOF can be expected in a certain direction, one could base the test statistic only on that one direction to obtain a more powerful test. However, in practice, such information is rarely available. In what follows, we always consider the global test statistic T_{gl}. The derivation of the asymptotic null distribution is beyond the scope of this thesis, but hypothesis testing may again be based on bootstrap p-values. In what follows this test is called the RRGL test.

6.1.4 Marginal regional residual plots

In case of more than one predictor variable, marginal regional residual plots are considered for each component of the multiple predictor vector $\mathbf{x}$. Standardized marginal regional residuals are plotted in each point of the (i, j) plane of the selected covariate x_k. As before, a light yellow to white colour is assigned to very large standardized regional residuals, and a red colour to very small values. Formal marginal regional residual plots are obtained by colouring regions for which the standardized regional residual exceeds the α-level critical value of the global test statistic T_{gl}. So, whenever one white or red spot appears in any marginal regional residual plot, the global null hypothesis of no lack-of-fit is rejected at the α significance level. In addition, the marginal plots show in which variables a region of lack-of-fit occurs and where this area is located. These marginal plots include a lack-of-fit test itself and thus allows one to conclude in a formal way where the multiple linear regression model is appropriate or not. The usefulness of these marginal plots in localizing lack-of-fit is illustrated in the next subsection for the US temperatures data example. Especially in case of more than two predictor variables, where graphical display of the regression model and the observed data is hardly possible, the marginal regional residual plots can be very helpful.

6.1.5 US temperatures data

To illustrate the tests and corresponding marginal regional residual plots, the US temperatures data, introduced in Chapter 2, is discussed. Recall that the normal average January minimum temperature, y, in degrees Fahrenheit (1931-1960) of 56 U.S. cities is studied in relation to longitude (in degrees), x_1, and latitude (in degrees), x_2. As in Example 3 in Chapter 3, a linear regression model in longitude, x_1, and latitude, x_2,

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon, \tag{6.1}$$

is tested for adequacy. The calculated values of the test statistics T_{1,S_n^2} and T_{2,S_n^2} from the data sample are 5.82 and 3.60, respectively. Thus, $T_{gl} = \max(T_{1,S_n^2}, T_{2,S_n^2}) = 5.82$, which corresponds to a bootstrap p-value < 0.00001. The percentiles of the test statistic T_{gl} were approximated using 100000 bootstrap samples drawn from the classical residuals, resulting in a critical value of 3.94 at the $\alpha = 0.05$ significance level.

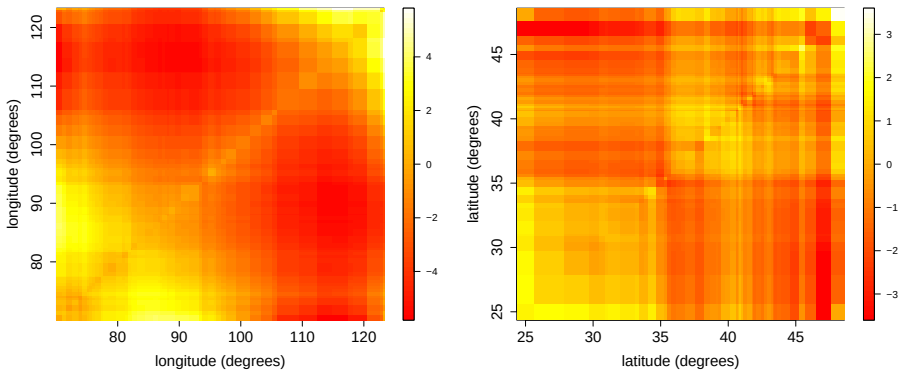

FIGURE 6.2: *Exploratory marginal regional residuals plots for longitude (left panel) and for latitude (right panel).*

A clear lack-of-fit is suggested in the exploratory marginal regional residual plot for longitude (Figure 6.2) and formally detected in the formal marginal regional residual plots (Figure 6.3). These plots can be used to localize the lack-of-fit. No significant lack-of-fit is found in the marginal regional residual plot of latitude, which confirms the earlier stated linear relationship between the mean January minimum temperature and latitude (Chapter 2). However, there is a clear lack-of-fit detected for the variable longitude. Figure 6.3 (left panel) shows that the underestimation of the data in the low and high-range of longitude is

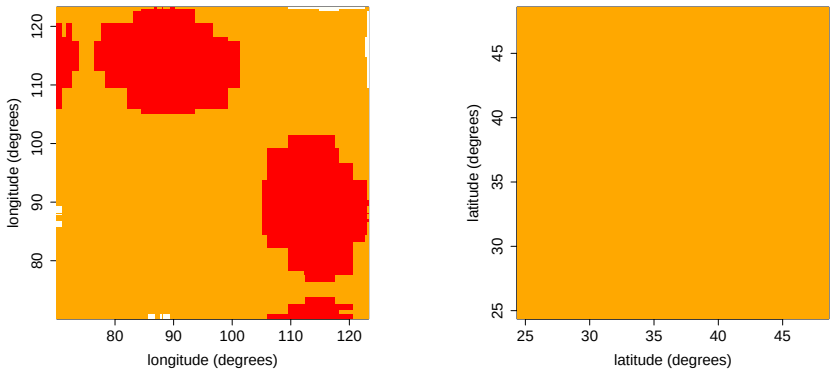

FIGURE 6.3: *Formal marginal regional residual plots for the US Temperature data ($p < 0.00001$) for longitude (left panel) and latitude (right panel). The red areas in the left panel show that the overestimation of the data in the high-range of longitude is statistically significant at the 5% level, while for latitude no regions of lack-of-fit are found.*

statistically significant, as well as a statistically significant overestimation for larger areas. The large amount of large areas indicates the presence of a global LOF. We formally conclude that the relationship between the mean January minimum temperature and longitude is not linear.
The solution proposed by Peixoto (1990), a cubic polynomial in longitude,

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_1^2 x_2 + \theta_6 x_1^3 + \theta_7 x_1^3 x_2 + \epsilon, \qquad (6.2)$$

results in marginal regional residuals plots that display no lack-of-fit. Both values of the fitted marginal statistics, $T_{1,S_n^2} = 3.23$ and $T_{2,S_n^2} = 3.07$, are smaller than the critical value 3.75 (p=0.332). This confirms that the second model is a major improvement as compared to the first model. No evidence is found that this model does not accurately predict the average January minimum temperature.

6.2 Spherical regional residuals

le Cessie and van Houwelingen (1995) pointed out that if the model does not fit, in some areas predictions will be too small as compared with the observed values, while in other regions, they will be too large. In any event, observations that are close to one another with respect to some distance measure in the predictor space will deviate from the model in the same directions and will be positively correlated. This thought is the underlying motivation for us to construct a test statistic based on regional residuals calculated over spherical

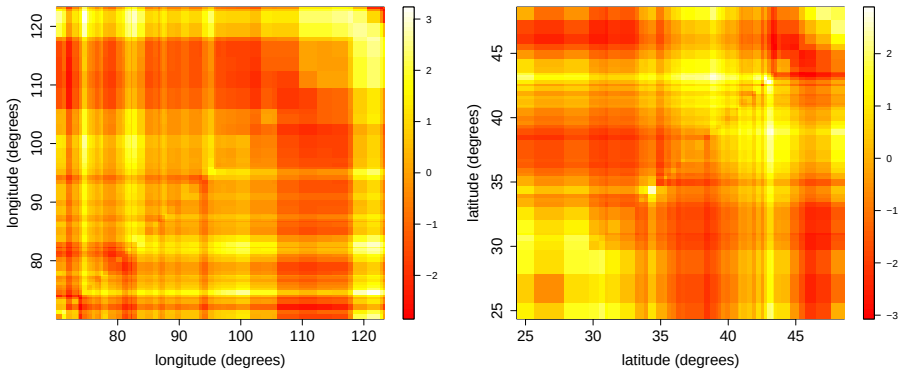

FIGURE 6.4: *Exploratory marginal regional residuals plots for the US Temperature data with the parametric model fit (Equation 6.2) suggested by Peixoto (1990) for longitude (left panel) and for latitude (right panel).*

subsets based on Euclidean distance measures on the standardized covariates. In what follows, we thus explicitly account for the multivariate nature of the predictor space by considering d-dimensional spheres instead of intervals for each predictor variable separately. We believe more powerful test statistics for LOF are constructed based on averages of residuals in a certain higher dimensional neighbourhood, rather than choosing a univariate direction. One could consider using a multivariate kernel and end up with smoothing based test statistics, but we prefer to be independent of any choice of type of smoother and smoothing parameter. The computational cost of considering all spherical neighbourhoods is the price that we are willing to pay. The definition of the regional residuals has to be adapted, so that it is based on a distance measure in the predictor space which avoids choosing an order of the residuals in advance, or choosing a smoothing parameter. Proper standardization of all covariates is crucial for the test to have power in all directions of the predictor space. If we would not standardize the covariates before applying the distance measure to the covariates, those variables with large variances would dominate the choice of the spherical subsets and the resulting test would only be powerful in those directions. Finally, a corresponding new type of regional residual plot is introduced as well.

6.2.1 Construction of spherical regional residuals

From now on we suppose that all predictor variables are standardized, so as they all have standard deviation one. Let $B_{i,r}$ denote the $d-$dimensional sphere

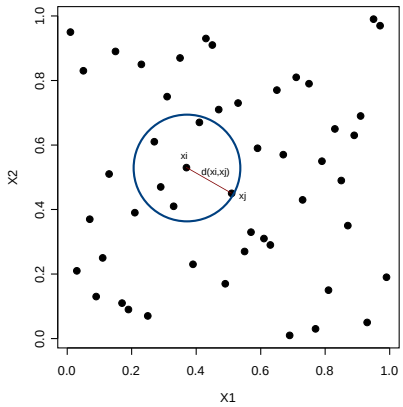

FIGURE 6.5: *Example of a 2 dimensional sphere $B_{i,r}$ in the case of two predictor variables x_1 and x_2.*

$B_{i,r} = \{\mathbf{x}_k \mid d(\mathbf{x}_i, \mathbf{x}_k) \leq r, k = 1, \ldots, n\}$ with the i^{th} design point as its center and radius r, and $d(\mathbf{x}_i, \mathbf{x}_k)$ is the Euclidean distance between design points $\mathbf{x}_i$ and $\mathbf{x}_k$. Spherical regional residuals $R(B_{i,r})$ are defined as the average of classical residuals, $e_k = (y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\theta}}_n))$, inside the $d-$dimensional sphere $B_{i,r}$, $(i = 1, \ldots, n)$, i.e.

$$R(B_{i,r}) = \frac{\sum_{k=1}^{n} e_k I(\mathbf{x}_k \in B_{i,r})}{\sum_{l=1}^{n} I(\mathbf{x}_k \in B_{i,r})} = \frac{1}{n_{B_{i,r}}} \sum_{k=1}^{n} e_k I(\mathbf{x}_k \in B_{i,r})$$

where $n_{B_{i,r}}$ is equal to the number of design points in $B_{i,r}$. Figure 6.5 shows an example of a 2 dimensional sphere $B_{i,r}$ in the case of two predictor variables x_1 and x_2.

When the radius $r = 0$, the regional residuals are equal to the classical residuals at each design point. In case of multiple measurements, a regional residual is defined as the average of the classical residuals at each design point. When the radius $r = \max_j d(\mathbf{x}_i, \mathbf{x}_j)$, $(j = 1, \ldots, n)$, the sphere $B_{i,r}$ contains all the design points and the corresponding regional residual is exactly 0. In what follows we calculate the regional residuals $R(B_{i,r})$ for all design points $\mathbf{x}_i$ and for all radii $r = d(\mathbf{x}_i, \mathbf{x}_j)$, $(i, j = 1, \ldots, n)$.

When no lack-of-fit is present, spherical regional residuals have mean zero and variance $n_{i,r}^{-1}\sigma^2 h_{i,r}^2$, where $h_{i,r}^2 = (\mathbf{I}_{i,r}^t \mathbf{I}_{i,r})^{-1} \mathbf{I}_{i,r}^t (\mathbf{I}_n - \mathbf{H}) \mathbf{I}_{i,r}$. The expression is

similar to the one for ordinary or marginal regional residuals, only $\mathbf{I}_{i,r}$ denotes now the $n \times 1$ inclusion matrix, where the k^{th} element of $\mathbf{I}_{i,r}$ equals 1 if $\mathbf{x}_k \in B_{i,r}$, otherwise it equals 0. If the residual variance σ^2 is unknown, we replace it by the natural estimator $S_n^2 = (n-p)^{-1} \sum_{i=1}^{n} (y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n))^2$ and we obtain standardized spherical regional (SSR) residuals

$$R_{S_n^2}(B_{i,r}) = \frac{\sqrt{n_{i,r}}}{S_n h_{i,r}} R(B_{i,r}).$$

6.2.2 A lack-of-fit test

As before, we consider the supremum norm, but now of the standardized spherical regional residuals, so as to obtain a global measure of LOF

$$T_{SRRS} = \sup_{i,r} \mid R_{S_n^2}(B_{i,r}) \mid . \tag{6.3}$$

The derivation of the asymptotic null distribution is out of the scope of this thesis. We refer to Chapter 9 for a brief discussion and suggest bootstrapping the null distribution for hypothesis testing.

6.2.3 Exploratory spherical regional residual plots

For ordinary regional residuals, the starting- and end points of the interval completely specify the region in the predictor space over which the regional residual is calculated. For spherical regional residuals, this role is taken over by the center $\mathbf{x}_i$ and radius d. We therefore generalize the formal regional residual plots as plots that are constructed by plotting the SSR residuals for each design point or center $\mathbf{x}_i$ and all radii $r = d(\mathbf{x}_i, \mathbf{x}_j), (j = 1, \ldots, n)$ in a bubble color plot. An example of this plot is shown in Figure 6.6 for the US temperature data set, for the assessment of the first order polynomial model fit (Equation 6.1). The x-axis represents the center of the SSR residuals, the y-axis represents the radius. The size of a bubble corresponds to the absolute value of the SSR residual. Large absolute values of these SSR residuals, and thus large bubbles, may indicate a possible lack-of-fit.
Note that by considering all design points as a center and all Euclidean distances between design points as radii, some SSR residuals will be duplicated. Therefore, to reduce calculation time, only unique SSR residuals are plotted, only for the center with the smallest index in the data set. SSR residuals with small radii correspond to small areas in the predictor space, close to the specific center. SSR residuals with large radii correspond to large areas in the predictor space. Both plots in Figure 6.6 represent actually the same SSR residuals, only the centers on the x-axis are ordered differently. By choosing different directions to order the centers, one might get an indication of which predictor

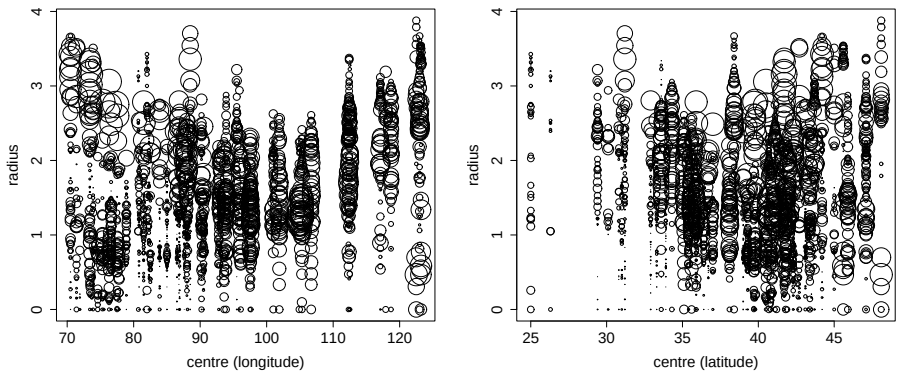

FIGURE 6.6: *Exploratory spherical regional residual plots for US temperature data. The size of a bubble corresponds to the absolute value of the SSR residual, the centers of the spherical regional residuals are ordered with respect to longitude (left panel) and latitude (right panel).*

causes the LOF. In this example, we find large absolute values of SSR residuals for large spherical subsets of the predictor space. In the exploratory SSR residual plot where the centers are ordered with respect to longitude, we also find a cluster of large absolute values of SSR residuals, corresponding to small regions in the very upper range of the longitude variable. To find out whether the observed deviations from the null model are statistically significant, we construct formal SSR residual plots in the next section.

6.2.4 Formal spherical regional residual plots

A colour scheme can be added to the exploratory SSR residual plots to formally locate in which regions the absolute value of the SSR residuals results in the rejection of the null hypothesis. In analogy to previous formal regional residual plots, we give a red colour to all negative SSR residuals for which their absolute values exceed the bootstrap 5%-level critical value. This percentile is approximated using 100 000 bootstrap samples drawn from the classical residuals (Section 3.5.2). Light yellow areas indicate all positive SSR residuals that exceed the bootstrap critical value, detecting a significant underestimation in the corresponding regions. For the US temperature data when the first order polynomial regression model is fit, the value of the test statistic T_{SRRS} is 5.82, which corresponds to a bootstrap p-value < 0.00001. A clear lack-of-fit is detected and the formal regional residual plot (Figure 6.7) can be used to localize this lack-of-fit. There is a significant underestimation of the data in small areas

in the high range of longitude and a clear overestimation for large areas, not having these design points as a center. To have a better idea of the location of the detected LOF, we plot the geographical map of the US and show which areas correspond to the largest negative and positive SSR residual (Figure 6.8).

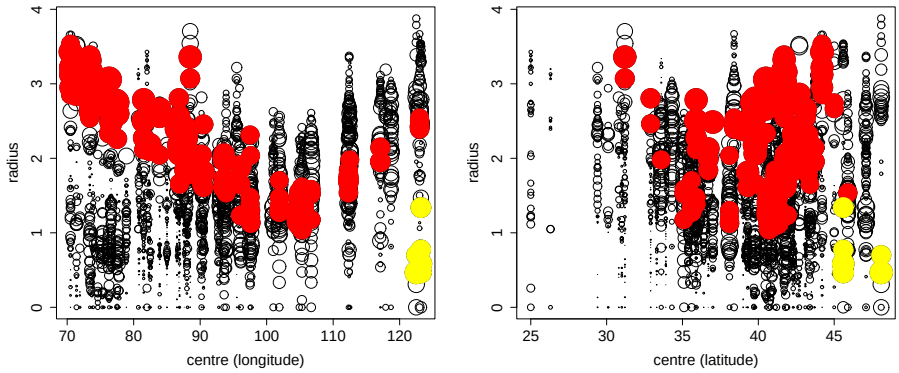

FIGURE 6.7: *Formal SSR residual plot ($p < 0.00001$) for US temperature data, locating the lack-of-fit ($\alpha = 0.05$). Yellow (resp. red) areas identify areas of under- (resp. over-) estimation of the data when fitting the first order polynomial regression model (Equation 6.1).*

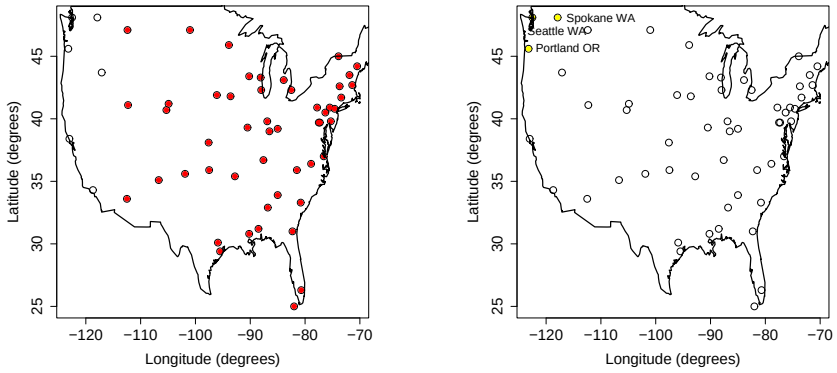

FIGURE 6.8: *Coloured dots in the geographical map of the US show the areas that correspond to the largest negative (red dots in the left panel) and positive (yellow dots in the right panel) SSR residuals for the US temperature data when fitting the first order polynomial regression model (Equation 6.1) .*

The largest negative SSR residual corresponds to a significant overestimation

of the data in a large area that almost covers the entire US, except for the west coast (left panel). The largest positive SSR residual corresponds to a significant underestimation of the data in a very small area around Seattle WA, Spokane WA and Portland OR (right panel). The large amount of negative SSR residuals in large areas suggests a global LOF. We can conclude that Model (6.1) is not appropriate for these data.
We further investigate the solution proposed by Peixoto (1990), a cubic polynomial in longitude (Equation 6.2). Although in Sections 3.1 and 6.1, we have concluded that this model is a considerable improvement over the first order polynomial model (Equation 6.1), we still detect a significant LOF when using the spherical regional residuals ($p = 0.006$). Figure 6.9 displays the formal SSR residual plots that display a significant local LOF in a small area in the low range of longitude (left panel) or in the high range of latitude (right panel). It corresponds to significant overestimation of the data in a small region around Burlington VM, Portland ME, Concord NH and Albany NY (Figure 6.10). We would advise the data analyst to further investigate the model in this neighbourhood and to be very careful if this model is used for predictions in this specified area.

6.3 Comparison to classical lack-of-fit tests

For the multiple regression setting, only a limited number of tests is available. In addition, all tests discussed in Chapter 3 that can be extended to multiple regression, suffer some disadvantages in this setting. For example, the classical F-test is easily extended to multiple predictor variables, but requires exact replicates. The reduction method requires a specific alternative model in advance. The nonparametric tests based on smoothers and applied to residuals like in Section 3.2 depend highly on the order relation chosen for the residuals when a univariate smoother is used. When a multivariate smoother is chosen to solve this problem, its dependence on the type of smoother and the smoothing parameter remains. The tests based on cumulative sums or averages of residuals are also in the multiple setting dominated by the residuals for which the predictor variables have low covariate values. Our tests do not suffer any of the above shortcomings as they can handle both exact replicates and no-replicates. They are omnibus in the sense that they are able to detect a wide range of alternatives without specifying an alternative in advance. They do not depend on the choice of a univariate direction, since the multivariate structure is taken into account by considering d dimensional spheres in the predictor space. However, the computational cost is heavy as we consider all possible spheres around each of the design points.

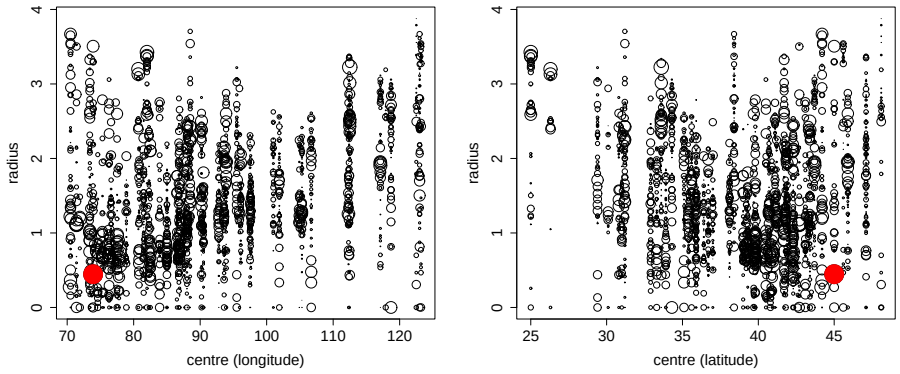

FIGURE 6.9: *Formal SSR residual plot ($p = 0.006$) for US temperature data, locating the lack-of-fit ($\alpha = 0.05$). The red area identifies a local area of overestimation of the data when fitting the third order polynomial regression model (Equation 6.2).*

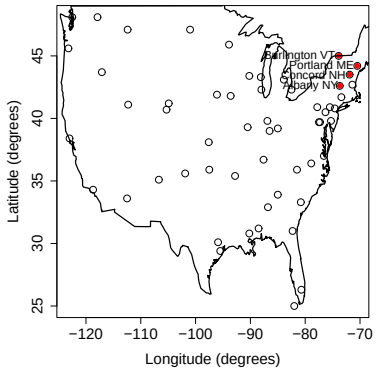

FIGURE 6.10: *Coloured dots in the geographical map of the US show the area that correspond to the only SSR residual that exceeds the 5% level critical bootstrap value when fitting the regression model proposed by Peixoto (1990) (Equation 6.2).*

6.3.1 Simulation study

To study the small sample power properties of our test based on spherical regional residuals, we compare our SRRS test to two smoothing based tests in case of both global and local LOF. The first test is the adaptive Neyman Test introduced by Kuchibhatla and Hart (1996), abbreviated as the KH test, and the second is the adaptive Neyman test proposed by Fan and Huang (2001), say the FH test. The third test in the study is our SRRS test. For all tests, bootstrap

p-values were generated using the residual based bootstrap procedure (Section 3.5.2).
We generate a design by considering three normally distributed covariates x_1, x_2 and x_3, with mean 0, variance 1 and bivariate correlations 0.5 and one binary covariate, x_4, which is independent of x_1, x_2 and x_3 with $P(x_4 = 1) = 0.4$ and $P(x_4 = 0) = 0.6$. The null model under study is the multiple linear regression model $m(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$, and the sample size is 128. All tests are performed at the 5% level of significance.
Two alternatives, a global and local LOF are included in this study. The first alternative represents a global LOF by generating 5000 random data sets from the quadratic regression model,

$$m_1 = x_1 + \lambda x_2^2 + 2x_4,$$

where λ is the LOF parameter that ranges from 0 to 1. The local LOF is introduced by generating 5000 random data sets from model m_2,

$$m_2 = \begin{cases} x_1 + 2x_4 & \text{if } \mathbf{x} \notin B_{84,d(84,86)} \\ x_1 + \lambda x_2^2 + 2x_4 & \text{if } \mathbf{x} \in B_{84,d(84,86)}, \end{cases}$$

where $B_{84,d(84,86)}$ denotes the sphere that has the 84^{th} design point as its center, and the Euclidean distance between the 84^{th} and 86^{th} design point as its radius. In both cases, standard normally distributed error terms were added to the model.
As already discussed in the introduction of this section, the order of the residuals determines the performance of the smoothing based tests. Therefore, Figure 6.11 shows the estimated power curves in case of the global LOF, m_1, when the residuals are ordered according to the direction where the LOF occurs, x_2 (upper left panel), according to the first principal direction (upper right panel) and when they are not ordered in a specific direction (lower left panel). As could be expected, the smoothing based tests perform better when the direction of LOF is known and the residuals are ordered according to this direction. However, when the direction is unknown, as is usually the case in practice, the best choice is to consider the first principal direction, though this seriously reduces the power of the smoothing based tests. When the residuals are unsorted, they have virtually no power left. On the other hand, our SRRS test does not suffer from this disadvantage, and performs equally well in all three situations and has even good power properties as compared to the smoothing based tests when the direction is known. Finally, in the lower right panel of Figure 6.11 the power curves of the three tests are shown in case of local LOF, m_2, when the residuals are ordered according to the first principal direction, as

would be done for the smoothing based tests in practice. A clear advantage in performance is found for the SRRS test.

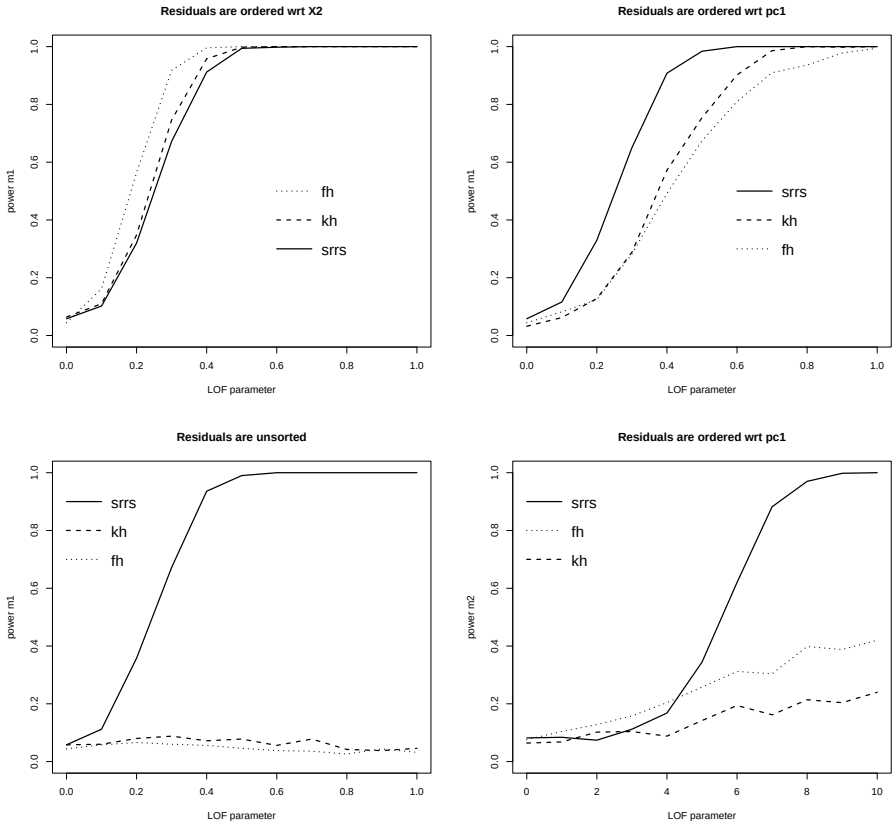

FIGURE 6.11: *Empirical powers of the FH, KH and SRRS test in case of global LOF when the residuals are ordered according to x_2, the LOF direction (upper left panel), according to the first principal direction (upper right panel), when the residuals are unsorted (lower left panel) and in case of local LOF when the residuals are ordered according to the first principal direction (lower right panel).*

6.4 One or more angular predictor variables

So far in this chapter, we studied possible extensions to multiple regression in case of predictor variables defined on the real line. Also in circular-linear regression, more than one predictor variable can be important in the prediction of the response variable, and these predictors can be both linear or circular. In this

section, we first describe the extensions to multiple circular-linear regression and we end this chapter with an illustration of the methodology on an environmental study where an air quality index is predicted by both temperature and wind direction.

6.5 Construction of marginal regional residual tests

The extension of the methodology in Section 6.1 is immediate if we combine the results of Chapter 5 and Section 6.1. Let d denote the number of predictor variables in the model, and let x_{ki} denote the k^{th} predictor variable for the i^{th} observation. For each predictor variable x_k we calculate the test statistic

$$T_k = \sup_{ij} \left| \frac{R(A_{kij})}{\text{sd}(R(A_{kij}))} \right|, \quad k = 1, \dots d$$

where for a circular predictor variable x_k, the A_{kij} refers to the arcs, and for a linear predictor variable to intervals.
These k test statistics only contain marginal information on lack-of-fit with respect to the k^{th} covariate, but they can be further combined into one global test statistic T, defined as the supremum of the p marginal statistics T_k ($k = 1, \dots, d$), i.e.

$$T_{gl} = \sup_{1 \le k \le d} (T_k).$$

The derivation of the asymptotic null distribution is beyond the scope of this thesis, but hypothesis testing may again be based on bootstrap p-values.
To localize the lack-of-fit in the predictor space, d formal regional residuals plots are constructed. Only areas for which the absolute value of the standardized regional residual exceeds the bootstrap $\alpha-$level critical value of the overall test statistic T_{gl} are now coloured white or red to indicate under- or overestimation, respectively. Our procedure is illustrated in the next section.

6.6 Air quality data

Johnson and Wehrly (1978) discussed the use of a conditional distribution for circular-linear regression with linear and circular predictor variables. To illustrate our LOF test, we reanalysed the regression model they obtained for the air quality index data (De Wiest and Della Fiorentina, 1975). The air quality index, y, is predicted as a function of the temperature in $^\circ C$, x_1, and the sine and cosine of wind direction in degrees, x_2. The resulting least squares regression equation is

$$\hat{y} = 0.306 + 0.028x_1 - 0.179\cos(x_2) + 0.216\sin(x_2). \tag{6.4}$$

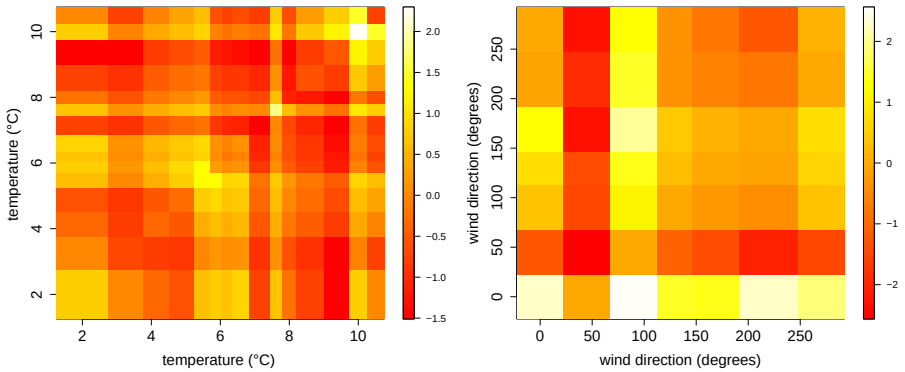

FIGURE 6.12: *Exploratory regional residual plots for the air quality data set with respect to temperature (left panel) and wind direction (right panel). The colour scheme of the standardized regional residuals ranges from red (large negative) to white (large positive values).*

We find $T_{gl} = \max(T_1, T_2) = \max(2.299, 2.565) = 2.565$ with p-value $p = 0.3185$. The exploratory regional residual plots, one for temperature (Figure 6.12, left panel), and one for wind direction (Figure 6.12, right panel) do not show any suspicious regions. Since no significant lack-of-fit was found, no formal regional residuals plots are shown.
To conclude this subsection, we would like to remark that the exploratory regional residuals plot for a circular predictor variable is not necessarily symmetric. This is illustrated in Figure 6.12 as for more complex null models, the standardization of the arcs and their complements may be different (Section 5.4).

6.7 Conclusions

Two possible extensions to the multiple linear regression setting for both linear and circular predictor variables are discussed in this chapter. The first one, based on marginal information for each predictor variable, is mainly useful to detect deviations from the null model in univariate directions. The second approach, however, takes the multivariate structure of the design space into account. In this way, a more powerful test for local deviations in the higher dimensional predictor space is constructed and allows the detection of a broader class of alternatives. The advantage of this approach is that no order relation of the residuals has to be specified in advance, neither the choice of a smoothing parameter. In addition, corresponding spherical regional

residual plots include a formal LOF test and locate the area of LOF in the predictor space. The test statistic is simple and intuitively appealing, though computationally demanding.

CHAPTER 7

Lack-of-fit in generalized linear regression models

In this chapter, possible extensions of the methodology to the complete class of Generalized Linear Models (GLM) are investigated. The GLM extends the linear model of Chapters 4 and 6 in several ways. We start with the special case of logistic regression models in Sections 7.1 and 7.2, and discuss the more general class of generalized linear models in Section 7.3.

7.1 Regional residuals in logistic regression analysis

When the outcome variable is binary, like the presence or absence of a certain disease, survival or death of patients, the occurrence of low birth weight of a newborn or not, ..., linear regression analysis is no longer appropriate. To illustrate this, assume n independent pairs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{m}(\mathbf{x}_i)^t$ is the $p-$dimensional vector of the functional forms of d fixed covariates, and the response is binomial distributed, $y_i \sim \text{Bin}(n_i, \pi(\mathbf{x}_i))$. This means that y_i is not coded as 0 or 1, but represents the number of successes or 1's for the i^th covariate pattern $\mathbf{x}_i$ with n_i replications. In linear regression analysis, the conditional mean $E(y\,|\mathbf{x}_i) = \sum_{j=0}^{p-1} m_j(\mathbf{x}_i)\theta_j$ of the linear regression model could take any value between $-\infty$ and $+\infty$. As the response is binomial, it should be formulated so as to be bounded between 0 and n_i. Therefore, a link function between the conditional mean and the linear regression model is introduced. Let μ_i denote the conditional mean of y given $\mathbf{x}_i$, i.e. $\mu_i = E(y\,|\mathbf{x}_i) = n_i\pi(\mathbf{x}_i)$. The logit transformation links this conditional mean to a linear predictor which is given by

$$g(\pi(\mathbf{x}_i)) = \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \sum_{j=0}^{p-1} m_j(\mathbf{x}_i)\theta_j.$$

The latter expression is linear in the parameters, and ranges from $-\infty$ to $+\infty$. Further, the error term ϵ which expresses the deviation between an observation and its conditional mean, $\epsilon_i = y_i - n_i\pi(\mathbf{x}_i)$, is no longer normally distributed.

It has conditional mean, $E(\epsilon\,|\mathbf{x}_i) = 0$, and conditional variance $\text{var}(\epsilon\,|\mathbf{x}_i)) = \text{var}(y\,|\mathbf{x}_i)) = n_i\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))$.
One important issue about logistic regression, and about GLM in general, is the existence of many different definitions of the residuals. Possible definitions are

$$e_{r,i} = y_i - n_i\hat{\pi}(\mathbf{x}_i)$$
$$e_{P,i} = \frac{y_i - n_i\hat{\pi}(\mathbf{x}_i)}{\sqrt{n_i\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))}}$$
$$e_{d,i} = \text{sign}(y_i - n_i\hat{\pi}(\mathbf{x}_i))\sqrt{2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}(\mathbf{x}_i)}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}(\mathbf{x}_i)}\right)\right]}$$

where $\hat{\pi}$ denotes the weighted least squares estimate of π, $e_{r,i}$ denotes the raw residuals, $e_{P,i}$ the Pearson residuals and, finally, $e_{d,i}$, the deviance residual. As described in Section 3.3, the squared Pearson and squared deviance residuals are the individual contributions to the Pearson χ^2 statistic and to the deviance, respectively.
As we only have the intension to briefly introduce logistic regression analysis here, we refer the reader to, e.g., Hosmer and Lemeshow (2000) for more details. In what follows, we discuss possible extensions of regional residuals for logistic regression models. We illustrate the methodology with some real data examples and investigate the performance of the tests in a small power study.

7.1.1 Regional residuals in logistic regression analysis

To generalize the notation for the regional residuals, let $C_{\alpha,\beta}$ denote the specific region over which the regional residual is calculated. For intervals, α and β denote the begin- and endpoint of the interval, respectively, while for spherical regions α and β denote the center and the radius. For simplicity, we define the regional residuals as the average of the raw residuals in the region $C_{\alpha,\beta}$,

$$R(C_{\alpha,\beta}) = \frac{\sum_{k=1}^{n} e_{r,k} I(\mathbf{x}_k \in C_{\alpha,\beta})}{\sum_{k=1}^{n} I(\mathbf{x}_k \in C_{\alpha,\beta})} = \frac{1}{n_{C_{\alpha,\beta}}} \sum_{k=1}^{n} e_{r,k} I(\mathbf{x}_k \in C_{\alpha,\beta}),$$

where $n_{C_{\alpha,\beta}}$ represents the number of observations in this specific region. Under the null hypothesis of no lack-of-fit, we expect the regional residuals to have zero mean. The expression for the variance is approximated by applying Pregibon's (1981) linear regression-like approximation for the residual at the i^{th} covariate pattern. In particular, let $\mathbf{X}$ denote the design matrix with i^th row $\mathbf{m}(\mathbf{x}_i)^t$, $\mathbf{Y}$ is the $n \times 1$ response matrix, and $\hat{\mathbf{V}}$ is the $n \times n$ diagonal variance covariance matrix of y with i^{th}-element $n_i\hat{\pi}_i(1-\hat{\pi}_i)$. Then,

$$y_i - n_i\hat{\pi}(\mathbf{x}_i) \approx ((\mathbf{I}_n - \mathbf{H})\mathbf{Y})_i,$$

where $\mathbf{H}$ represents the logistic regression version of the hat matrix, $\mathbf{H} = \hat{\mathbf{V}}^{1/2}\mathbf{X}(\mathbf{X}^t\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}^t\hat{\mathbf{V}}^{1/2}$. The regional residuals can then be written as

$$R(C_{\alpha,\beta}) = (\mathbf{I}^t_{C_{\alpha,\beta}}\mathbf{I}_{C_{\alpha,\beta}})^{-1}\mathbf{I}^t_{C_{\alpha,\beta}}(\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

where $\mathbf{I}_{C_{\alpha,\beta}}$ is the $n \times 1$ column matrix for which the i^{th} element is 1 if $\mathbf{x}_i \in C_{\alpha,\beta}$ and 0 otherwise. Their variance is approximated by

$$\text{Var}(R(C_{\alpha,\beta})) \approx (\mathbf{I}^t_{C_{\alpha,\beta}}\mathbf{I}_{C_{\alpha,\beta}})^{-2}\mathbf{I}^t_{C_{\alpha,\beta}}(\mathbf{I}_n - \mathbf{H})\hat{\mathbf{V}}(\mathbf{I}_n - \mathbf{H})^t\mathbf{I}_{C_{\alpha,\beta}}.$$

7.1.2 Tests and plots

As before, large values of regional residuals may indicate a possible lack-of-fit of the logistic regression model. To obtain a global measure of lack-of-fit, we take again the supremum norm of the regional residuals as a test statistic,

$$T_{RRLR} = \sup_{C_{\alpha,\beta}} \left| \frac{R(C_{\alpha,\beta})}{\text{sd}(R(C_{\alpha,\beta}))} \right|,$$

where sd(.) denotes the standard deviation. Large sample properties of the test statistic are discussed in Chapter 8, but we recommend the parametric bootstrap scheme of Section 3.5.1 to obtain approximate p-values. Regional residual plots can be constructed as explained in Chapter 4 and Chapter 6. These are illustrated in the next subsection.

7.1.3 Illustration

To illustrate the use of the statistical test, together with the corresponding plots, we discuss the use of the extended methodology in an artificial data example. We consider an equally spaced design, $x_j = \frac{j-0.5}{n} - 0.5$, $j = 1, \ldots, n$ with $n = 50$ design points and one replicate on each design point. Figure 7.1 (upper panel) shows the generated data, together with the logit of the true conditional mean (dotted line),

$$g(\pi(x)) = \begin{cases} 3x & \text{if } x \notin [-0.49, -0.19] \\ 3x + 3.5\sin(19x) & \text{if } x \in [-0.49, -0.19] \end{cases}$$

and the weighted least squares fit of the simple linear logistic regression model (full line), $g(\hat{\pi}(x)) = 0.282 + 2.923x$.
The value of the test statistic T_{RRLR} is 4.021 (p=0.0465) and thus detects the local LOF at the 5% significance level. Also the Hosmer - Lemeshow deciles of risk test (Section 3.3.1) rejects the null hypothesis of no lack-of-fit at the

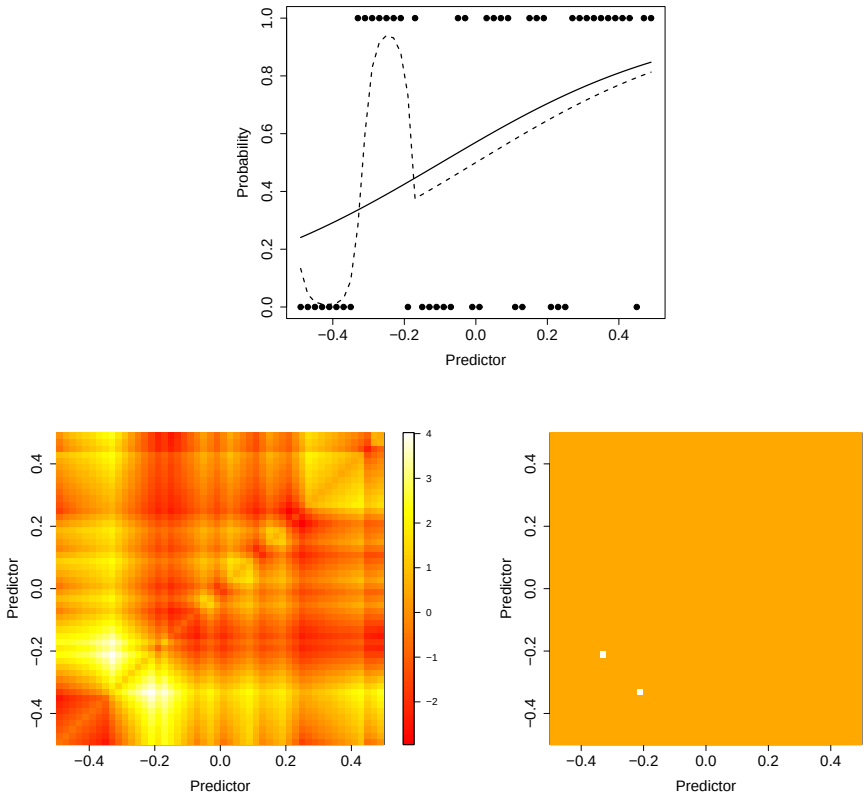

FIGURE 7.1: *(Upper panel) Artificial data example with a clear local LOF in the lower range of the predictor variable x. The full line represents the least squares fit of a simple linear logistic regression model, the dashed line the logit of the true underlying linear predictor. (Left panel) Exploratory regional residual plot based on raw residuals. (Right panel) Formal regional residual plot based on raw residuals ($p = 0.0465$).*

5% significance level (p=0.0343), but the Pearson χ^2 test (p=0.7336) and the unweighted residual sum of squares test (p=0.2876) do not. If we want to know whether there is a global or local LOF present, and where it is located in the predictor space in case of local deviations, we construct the regional residual plots. Figure 7.1 shows the exploratory regional residual plot in the left panel, which already suggests a local LOF in a small area in the lower range of the predictor variable. White and light yellow areas suggest areas of underestimation, red areas of overestimation. The formal regional residual plot in the right panel confirms that there is a significant underestimation in the area [-0.33;-0.21], which exactly corresponds to the local area where the positive part of the sine function was added to the simple linear logistic regression model.

For an illustration of the extension of spherical regional residual tests and plots we refer to Section 7.2.2 where the methodology is illustrated on a real data example.

7.1.4 Alternative test statistics

Instead of using raw residuals, we could use Pearson or deviance residuals as well. Although the Pearson residuals appear to be standardized, leverage adjustments should be taken into account to compensate for estimation of the parameters in the linear predictor (e.g. Hosmer and Lemeshow, 2000, and Williams, 1987). The standardized Pearson residuals are defined as

$$e_{P,i} = \frac{y_i - n_i\hat{\pi}(\mathbf{x}_i)}{\sqrt{n_i\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))(1-h_i)}}, \tag{7.1}$$

where h_i is the i^{th} diagonal element of the hat matrix $\mathbf{H}$. In large samples, we expect $e_{P,i}$ to have mean zero and variance approximately 1. Similarly, the leverage adjustment is also applied to deviance residuals,

$$e_{d,i} = \frac{\text{sign}(y_i - n_i\hat{\pi}(\mathbf{x}_i))\sqrt{2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}(\mathbf{x}_i)}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}(\mathbf{x}_i)}\right)\right]}}{\sqrt{1-h_i}}. \tag{7.2}$$

We could define regional residuals now as the average of the standardized Pearson or the standardized deviance residuals in a specific area $C_{\alpha,\beta}$ in the predictor space. As standardizing regional residuals based on deviance residuals is not straightforward and as we obtained nice results for the performance of the RRUnij test in Chapter 4, we now prefer to use the unstandardized version of our test statistic (Chapter 4) to create a corresponding test statistic here.

As alternative global measures of model deviations we consider the test statistics

$$T_{RRLP} = \sup_{C_{\alpha,\beta}} \left| \frac{1}{\sqrt{n_{C_{\alpha,\beta}}}} \sum_{k=1}^{n} e_{P,k} I(\mathbf{x}_k \in C_{\alpha,\beta}) \right|, \tag{7.3}$$

$$T_{RRLD} = \sup_{C_{\alpha,\beta}} \left| \frac{1}{\sqrt{n_{C_{\alpha,\beta}}}} \sum_{k=1}^{n} e_{d,k} I(\mathbf{x}_k \in C_{\alpha,\beta}) \right|, \tag{7.4}$$

The derivations of the asymptotic distributions of test statistics T_{RRLP} and T_{RRLD} are out of the scope of this dissertation. We suggest to bootstrap the null distribution for hypothesis testing.

7.1.5 Small sample behaviour

We perform a small simulation study to investigate the performance of the three new tests in logistic regression in comparison with classical tests discussed in Chapter 3. We included the Pearson χ^2 goodness-of-fit test, denoted as X, the Hosmer-Lemeshow decile of risk test, C (Hosmer and Lemeshow, 1980), and the unweighted residual sum-of-squares test, S (Copas, 1989). We also included two smooth tests of le Cessie and van Houwelingen (1991): the uniform kernel smooth, SRU, and the cubic weight smooth, SRC, as described in Hosmer et al. (1997). Finally, the three new tests, the $RRLR$ test based on raw residuals, the $RRLP$ based on standardized Pearson residuals, and the $RRLD$ test based on standardized deviance residuals, were included as well. In what follows, $RRL.$ refers to all regional residual tests for logistic regression.
For the classical tests, the asymptotic null distribution is used, except for the Pearson χ^2 goodness-of-fit test. For this test, using the χ^2_{n-p} distribution as null distribution is inappropriate, because it is based on a contingency table whose expected cell frequencies are too small. Also, for the S test the asymptotic null distribution is not appropriate when replicates are available. For these two classical tests, and for the three regional residual tests the parametric bootstrap (Section 3.5.1) was used for approximating the empirical powers.
In the next simulation study, we focus on LOF that occurs due to a misspecified linear predictor, and again we consider both global and local LOF. For the global LOF, we reconsider a simulation study in Hosmer et al. (1997). The distribution of the continuous predictor variable is $x \sim U(-3, 3)$. The outcome variable y is generated using the logistic regression model with $g_1(\pi(x)) = \theta_0 + \theta_1 x + \theta_2 x^2$ where we chose the values of the three parameters such that $\pi(-1.5) = 0.05$, $\pi(3) = 0.95$ and $\pi(-3) = \gamma$, where γ ranges between 0.01 and 0.5. The parameter γ is thus a LOF parameter and indicates the strength

of LOF. For larger values of γ the lack of linearity in the logit function becomes progressively more pronounced. We have generated 1000 data sets of 50 design points with only one replicate for each design point, $n_i = 1$, and 1000 data sets of 50 design points with 5 replicates, $n_i = 5$.
For the local LOF, we take an equally spaced fixed design, $x_j = \frac{j-0.5}{n} - 0.5$, $j = 1, \ldots, n = 50$ and consider a local misspecification of the linear predictor in the lower, mid and upper range of x by adding one period of a sine function. More specifically, the three linear predictors are

$$g_2(\pi(x)) = \begin{cases} -4x & \text{if } x \notin [-0.49, -0.19] \\ -4x + \lambda \sin(19x) & \text{if } x \in [-0.49, -0.19], \end{cases}$$

$$g_3(\pi(x)) = \begin{cases} -4x & \text{if } x \notin [-0.17, 0.15] \\ -4x + \lambda \sin(19x) & \text{if } x \in [-0.17, 0.15], \end{cases}$$

$$g_4(\pi(x)) = \begin{cases} -4x & \text{if } x \notin [0.19, 0.49] \\ -4x + \lambda \sin(19x) & \text{if } x \in [0.19, 0.49], \end{cases}$$

where λ represents the LOF parameter and ranges from 0 to 6, for the 1000 data sets of 50 design points with only one replicate for each design point, i.e. $n_i = 1$. For the 1000 data sets with 5 replicates, $n_i = 5$, λ ranges from 0 to 3.

The upper panels in Figure 7.2 show an example of the global LOF generated by $g_1(\pi(x))$ for one replicate at each design point (left panel) and for five replicates (right panel). The empirical power curves of *X*, *C*, *S*, *SRU*, *SRC*, *RRLR*, *RRLP* and *RRLD* tests are shown in the middle panels. The same line type is used for the two smooth tests and for the three regional residual tests. To distinguish between these curves in the plots, the last letter of the abbreviation of the test is added to the curve. The classical tests perform better than the regional residual tests for both designs, with and without replicates. The performance of the regional residual test based on raw residuals comes very close to those of the classical tests, particularly when replicates are available. When the LOF is detected by the *RRL*. tests, we find in the regional residual plot that it concerns a global lack-of-fit. The classical tests do not provide this information. A scatter plot of the smooth residuals that correspond to the two smooth tests, *SRU* and *SRC*, may provide this information as well, though not in a formal way. Comparing the three regional residual tests for this type of global LOF, we find a power advantage for the test based on raw residuals. For data sets without replicates, no clear distinction can be made between the *RRLD* and the *RRLP* tests, while for data sets with replicates the *RRLP* test clearly performs better.
As was shown in Section 4.4, tests based on unstandardized regional residuals are powerful to detect global deviations when the factor $\sqrt{n_{ij}}$ is replaced

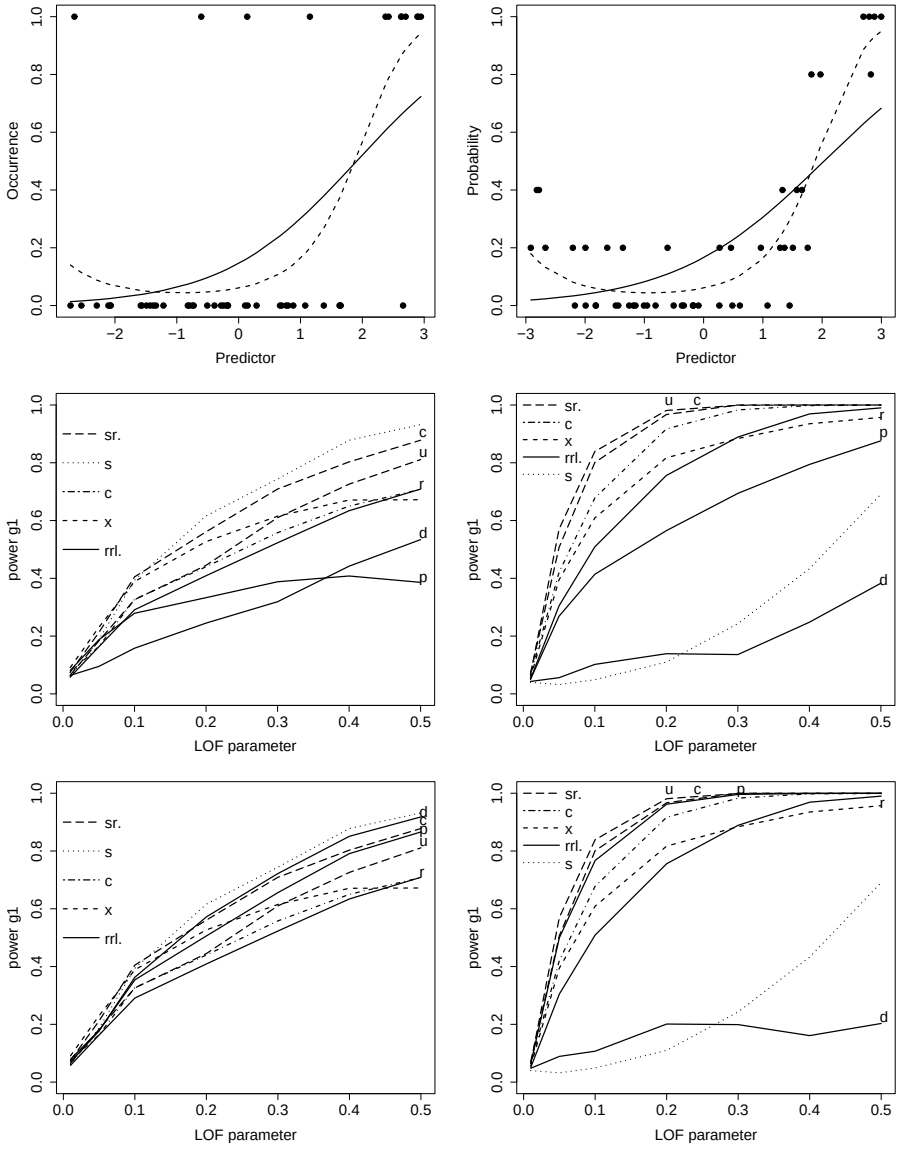

FIGURE 7.2: *(Upper panels) Illustration of global LOF with $\gamma = 0.05$. (Middle panels) Empirical powers of X, C, S, SRU, SRC, RRLR, RRLP and RRLD tests based on 1000 data sets of 50 design points. The same line type is used for the two smooth tests and for the three regional residual tests. To distinguish between these curves in the plots, the last letter of the abbreviation of the test is added to the curve. (Lower panels) Identical to the middle panels, except that for the RRLP and RRLD tests $\sqrt{n_{ij}}$ is replaced by $\sqrt{n}$. (Left panels) Data are generated using the linear predictor function g_1, with only one replicate available for each design point, $n_i = 1$. (Right panels) Data are generated using g_1 with 5 replicates for each design point, $n_i = 5$.*

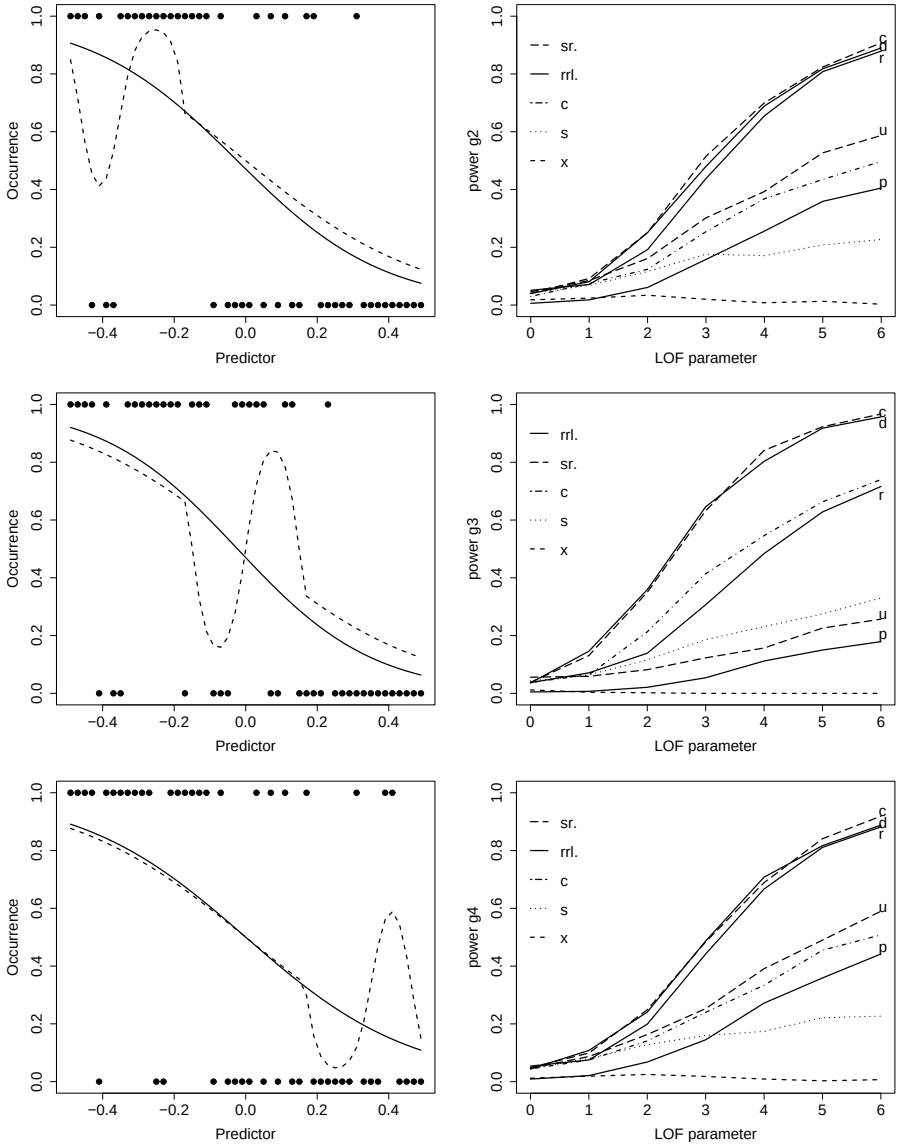

FIGURE 7.3: *(Left panels) Illustration of type of local LOF with $\lambda = 2$. (Right panels) Empirical powers of X, C, S, SRU, SRC, RRLR, RRLP and RRLD tests based on 1000 data sets of 50 design points. The same line type is used for the two smooth tests and for the three regional residual tests. To distinguish between these curves in the plots, the last letter of the abbreviation of the test is added to the curve. Data are generated using the linear predictor function g_2 (upper panels), g_3 (middle panels), and g_4 (lower panels) with only one replicate for each design point, $n_i = 1$.*

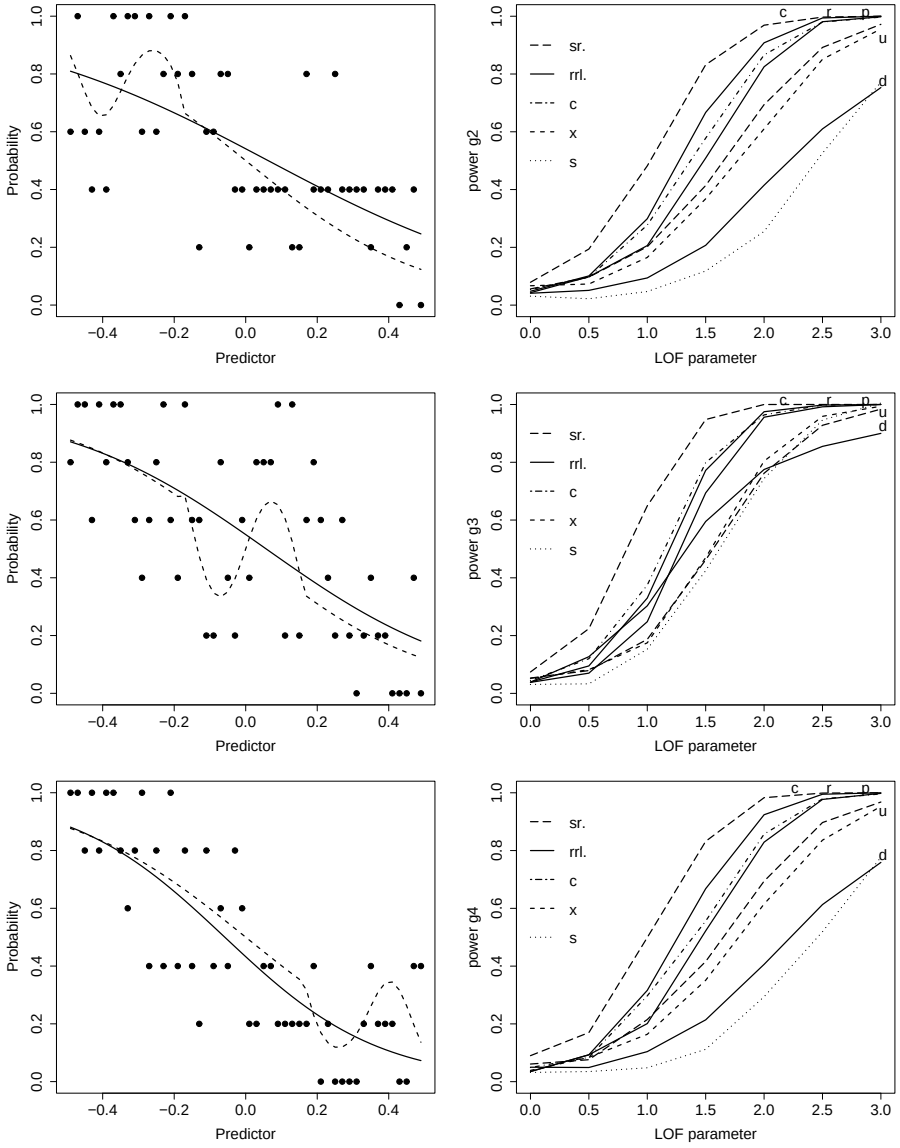

FIGURE 7.4: *(Left panels) Illustration of type of local LOF with $\lambda = 1$. (Right panels) Empirical powers of X, C, S, SRU, SRC, RRLR, RRLP and RRLD tests based on 1000 data sets of 50 design points. The same line type is used for the two smooth tests and for the three regional residual tests. To distinguish between these curves in the plots, the last letter of the abbreviation of the test is added to the curve. Data are generated using the linear predictor function g_2 (upper panels), g_3 (middle panels), and g_4 (lower panels) with five replicates for each design point, $n_i = 5$.*

with $\sqrt{n}$, so that regional residuals that are calculated over large intervals become relatively more important. This is illustrated in the lower panels of Figure 7.2. Both the *RRLD* and *RRLP* tests gain in power and become very competitive with the classical LOF tests. Only when replicates are available, no power improvement is found for the *RRLD* test, which already has an inferior performance. However, as we focus on detecting and localizing local LOF in particular, we only consider the *RRLP* and *RRLD* tests as defined in Equation 7.3 and Equation 7.4 in what follows.

Figure 7.3 shows the empirical power curves in case of local LOF and for data sets when only one replicate is available for each design point. In this case the cubic weight smooth test, *SRC*, and the regional residual test based on standardized deviance residuals perform best. When the local LOF is situated in the lower or the higher range of the predictor variable, the regional residual tests based on the raw regional residuals also perform very well, but they loose considerable power when the local LOF is situated in the mid range (as does the uniform kernel smooth test). The Pearson χ^2 test cannot detect any local deviations at all, and the *S* and *RRLP* tests perform insufficiently in case of local LOF without replicates.

When replicates are available, all tests gain power and all tests have rather good performance. The regional residual tests based on the raw and standardized Pearson residuals, and the Hosmer-Lemeshow decile of risk test, *C*, perform best, even slightly surpassed by the cubic weight smooth test, though this one seems to reject too often in case of no lack-of-fit. The regional residual test based on standardized deviance residuals clearly has an inferior performance in this case.

In summary, we recognize the strong performance of the smooth tests, especially the *SRC* test, in nearly all cases. The performance of the regional residual based tests is equally good in case of local lack-of-fit. To obtain the same performance in global lack-of-fit, we should adapt the *RRLP* and *RRLD* tests. Both procedures come with a graphical tool to locate LOF in the predictor space, but the regional residuals plots do this in a formal way. Both procedures also have their drawbacks. The performance of the smooth tests depends on the choice of the smoother and, even more important, the bandwidth (le Cessie and van Houwelingen (1991), Hosmer et al. (1997)), while the regional residual tests are computationally intensive. The simulation study indicates that the performance depends on the type of residuals used to calculate the lack-of-fit test, but to the best of our knowledge no discussion in this context is available. In practice, when applying the regional residual tests for a single predictor variable, we recommend the test based on standardized deviance residuals when replicates are not available, and the regional residual test based on raw residuals

when they are available.

7.2 Data examples

To illustrate the methodology in the context of logistic regression models, we discuss three data examples from the literature. Firstly, we consider the dose-response data with a single covariate, discussed by Bedrick and Hill (1990). For the multiple setting we apply our methodology to the well known vasoconstriction data of Finney (1947). Finally, we include the POPS data of Verloove and Verwey (1988) to illustrate that our tests are also applicable in large datasets.

7.2.1 Dose - response data

In this dose response experiment the predictor variable $logDose$ gives the log dose of benzepyrene administered to mice. The response variable y is the number of mice that are affected with a tumour. Note that replicates are available, as the same log dose is administered to several mice. As recommended in the previous section, we perform a regional residual test based on raw residuals to determine whether the simple linear logistic regression model with logit link is appropriate for these data. Regional residuals are calculated over all possible intervals in the design space, as only one predictor variable is present. Figure 7.5 shows the data together with the weighted least squares fit of a simple linear logistic regression model, and the exploratory and formal regional residual plots based on raw residuals. The bootstrap p-value equals p=0.0069 and is based on 10000 bootstrap samples. We conclude at the 5% level of significance that the simple linear logistic regression model is not appropriate for these data. The formal regional residual plot (right panel) shows a significant underestimation of the data in the first design point and a significant overestimation of the data in some larger intervals. Further model building will be necessary to obtain a more appropriate model for the data at hand.

7.2.2 Vasoconstriction data

The vasoconstriction data (Finney, 1947) comes from a carefully controlled study of the effect of the rate (liters per second) and the volume (litres) of air inspired on a transient vasoconstriction in the skin of digits. The response variable y is the occurrence or nonoccurence of vasoconstriction in the skin of digits. The linear logistic regression model in log rate and log volume with logit link is suggested as an appropriate model for these data. As we have two predictor variables and no replicates, spherical regional residuals based on the standardized deviance residuals are used to assess the fit of the model. 10000 bootstrap samples are used to obtain the bootstrap p-value of 0.0071. Thus, at the 5%

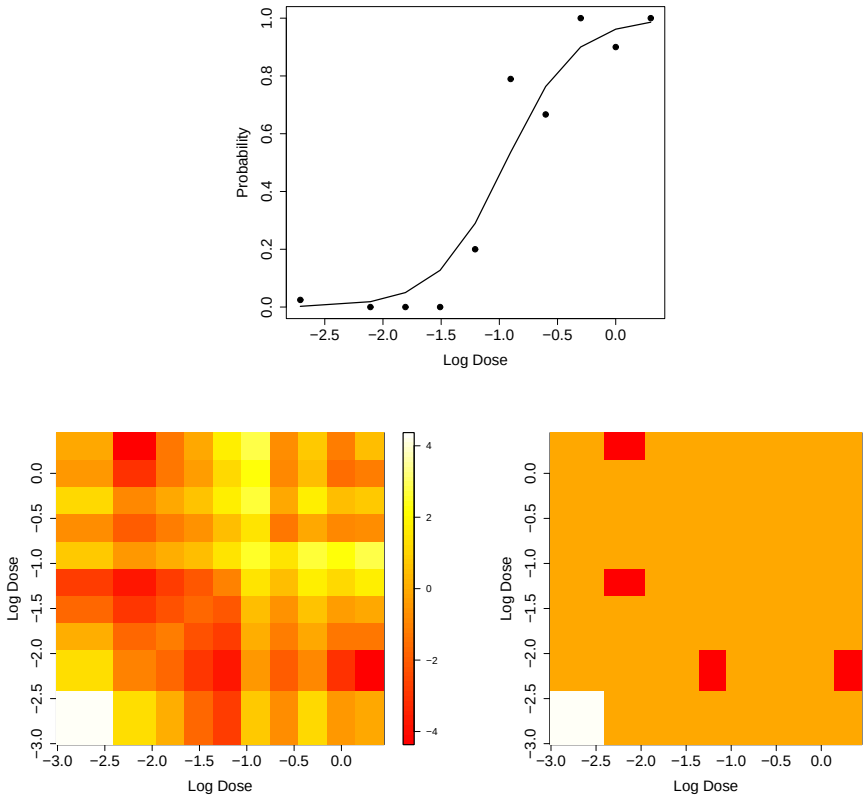

FIGURE 7.5: *(Upper panel) Dose-response data; y = observed number of affected mice, x = log dose of injected benzepyrene. The full line represents the weighted least squares fit of a simple linear logistic regression model. (Left panel) Exploratory regional residual plot for the dose-response data based on raw residuals. (Right panel) Formal regional residual plot for the dose-response data based on raw residuals ($p = 0.0069$).*

level of significance a local LOF is found in the formal regional residual plots in Figure 7.6 for a very small area in the mid range of the predictor variable rate and the low range of predictor variable volume. We locate the area in the predictor space by colouring the design points over which the regional residual that exceeds the bootstrap α-level critical value, is calculated (Figure 7.6, lower panel). In this plot, full dots represent subjects with occurrence of vasoconstriction. When both volume and rate are small, no response occurred, but when either was large (unless the other was very small) the response occurred. For the two subjects that are included in the regional residual that exceeds the critical bootstrap value, the occurrence seems to be unexpected according to the model. These two design points are also recognized in the literature on diagnostics and outlier detection (e.g. Finney (1947) and Pregibon (1981)). Unlike lack-of-fit tests such as those of Su and Wei (1991) and Cheng and Wu (1994) who do not find any evidence against the null hypothesis, our methodology detects one or a small group of outlying observations.

7.2.3 POPS data

The POPS data set originates from the project on preterm and small for gestational age infants in the Netherlands, a Dutch follow-up study on preterm infants by Verloove and Verwey (1988). The study collected information on 1338 infants born in the Netherlands in 1983, having gestational age less than 32 weeks and/or birthweight less than 1500 g. The outcome of interest is a binary variable that indicates whether or not the infant has died within 2 years or survived with a major handicap. After deletion of observations with missing data, a data set of 1310 infants remains. In particular, we include this data set to illustrate our methodology for large data sets.
We first examine whether a logistic regression model, linear in gestational age (in weeks), x_1, and birthweight (in 100g), x_2, fits the data well,

$$g(\pi(x_1, x_2)) = \theta_0 + \theta_1 x_1 + \theta_2 x_2. \tag{7.5}$$

A spherical regional residual based test is used to assess the model fit. Calculating all regional residuals over all possible spheres would require a huge simulation time. As lots of regions contain the same information, we perform the regional residual tests here on a random selection out of all possible regional residuals. To empirically validate this procedure, we selected several samples of 15000 regional residuals, and obtained very similar p-values and regional residual plots in all cases.

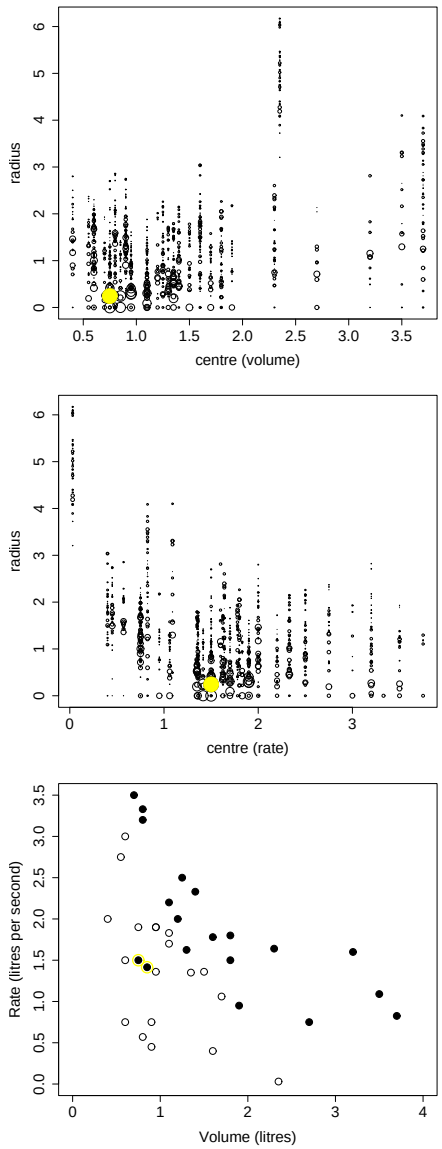

FIGURE 7.6: *Formal regional residual plots for the vasoconstriction data where the centers of the regional residuals are ordered with respect to volume (upper panel) and rate (middle panel) for standardized deviance residuals ($p = 0.0071$). (Lower panel) Location of the regional residual that exceeds the bootstrap α-level critical value in the predictor space for standardized deviance residuals. Full dots represent subjects with occurrence of vasoconstriction. A yellow circle around the subject marks the subjects that were used to compute the regional residual that exceeds the critical bootstrap value at the 5% significance level.*

Note that the simulation study performed in Section 7.1.5 does not involve multiple predictor variables. Therefore, we present in what follows, the results of the spherical regional residual test, based on raw residuals, as was done for the supremum tests based on cumulative sums of residuals in Lin et al. (2002). The RRLR test is performed on a randomly selected group of 15000 regional residuals. 5000 bootstrap samples are used to obtain the approximate bootstrap p-value of < 0.001. Thus, at the 5% level of significance a clear LOF is found in the formal regional residual plots in the upper panels of Figure 7.7. We observe large areas in the mid range of both predictor variables, representing an area of overestimation. Further, we observe two groups of smaller regions of underestimation. Some representative areas in the predictor space that correspond to regional residuals with the smallest negative and with two positive values, that exceed the bootstrap 5%-level critical value, are shown in the middle and lowers panels of Figure 7.7. From all these plots, we conclude that too high risks are predicted for observations in the center, as areas of significant overestimation are found in larger areas in the mid range of both predictor variables. The model predicts too low a risk for the infants with both the smallest gestational ages and smallest birthweights, and for the infants with a larger gestational age, as areas of significant underestimation are found in the low range of both predictor variables and, in addition, in a second area of significant underestimation, located in the high range of gestational age and the mid range of birthweight.

le Cessie and van Houwelingen (1991) reported similar results and suggested to ameliorate the model by including quadratic terms,

$$g(\pi(x_1, x_2)) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 (x_1 - 30)^2 + \theta_4 (x_2 - 12)^2. \tag{7.6}$$

We perform again the spherical regional residual test on raw residuals and obtain an approximate p-value of 0.045 based on 5000 bootstrap samples, indicating borderline significance at the 5% level. Figure 7.8 shows the formal regional residual plots, when the centers are ordered according to gestational age and birthweight. The lower panel in Figure 7.8 shows the corresponding area in the predictor space of the smallest negative regional residual. We find some evidence against model (7.6) in a small area of overestimation in the low to mid range for both predictor variables. Note that Aerts et al. (2000) also found some evidence against this model. They suggested a model based on Legendre polynomials instead.

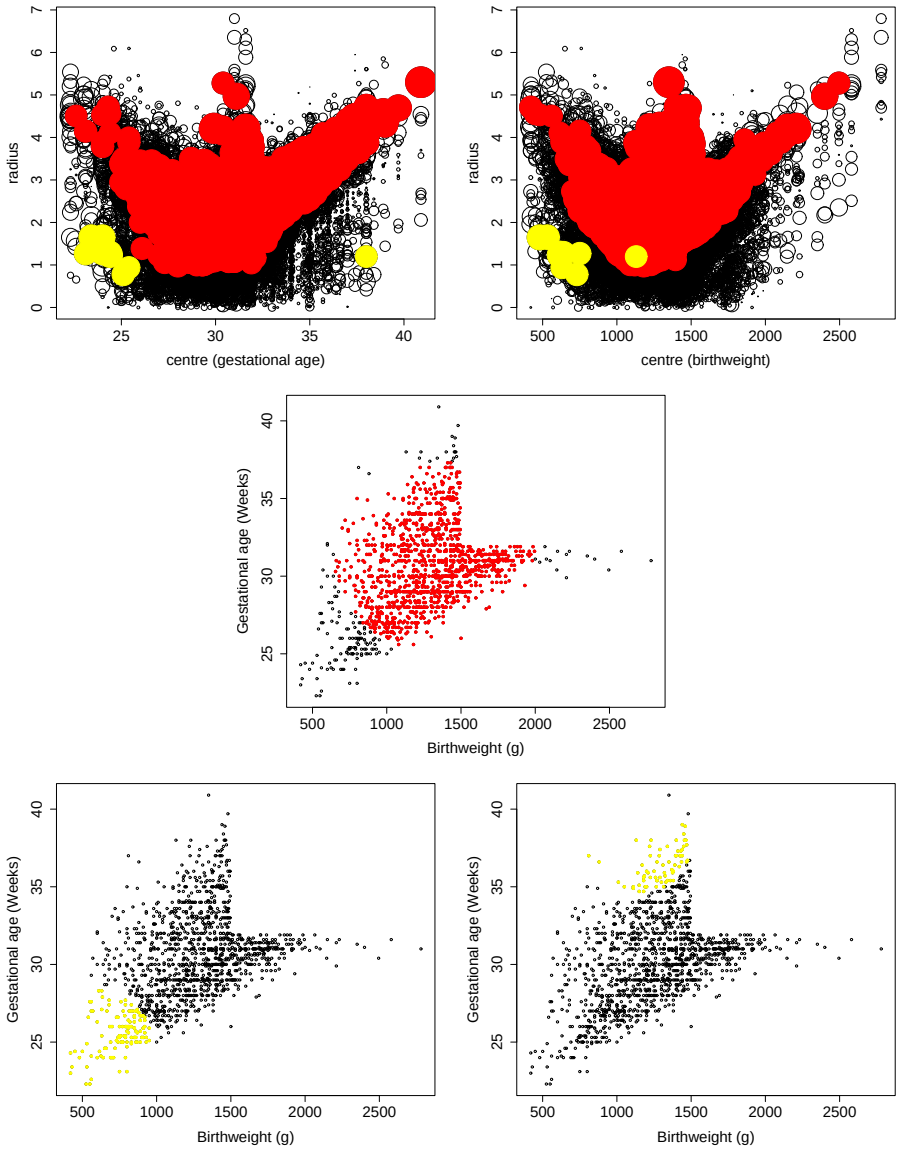

FIGURE 7.7: *Formal regional residual plots for the POPS data for model (7.5), where the centers of the regional residuals are ordered with respect to gestational age (left upper panel) and birthweight (right upper panel) for raw residuals ($p < 0.001$). Location of the smallest negative (middle panel) and two positive regional residuals (lower panels) whose absolute values exceed the bootstrap 5%-level critical value in the predictor space for raw residuals. Yellow dots correspond to an area of underestimation, red dots to an area of overestimation at the 5% significance level.*

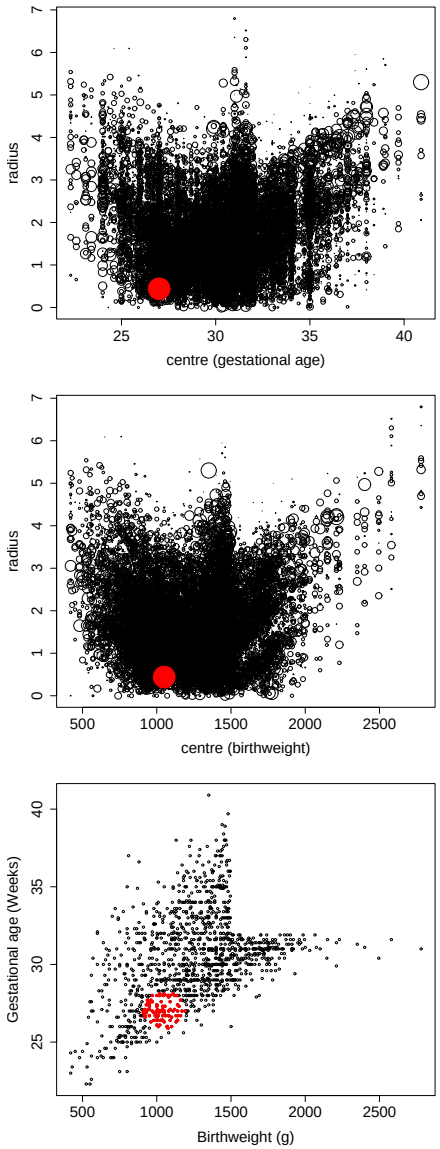

FIGURE 7.8: *Formal regional residual plots for the POPS data for model (7.6), where the centers of the regional residuals are ordered with respect to gestational age (upper panel) and birthweight (middle panel) for raw residuals ($p < 0.001$). (Lower panel) Location of the smallest negative regional residual whose absolute value exceeds the bootstrap 5%-level critical value in the predictor space for raw residuals. Red dots correspond with an area of overestimation at the 5% significance level.*

7.3 Extensions to the more general class of generalized linear models

The methodology of the previous sections can be extended to the complete class of Generalized Linear Models (GLM) (see e.g. McCullagh and Nelder (1989), and Fahrmeir and Tutz (1994) for an overview), which allows distributions in the exponential family, like the normal and the binomial distribution, but also the Poisson, gamma and Inverse Gaussian distribution, among others. The variance function is expressed as an explicit function of the mean $\mu = E(y\,|\mathbf{x})$,

$$\text{var}(y\,|\mathbf{x}) = \frac{\phi v(\mu)}{\omega},$$

where v is the variance function, which depends on the type of exponential distribution of the response y, and ϕ denotes the dispersion parameter, possibly unknown. The parameter ω is a prior known weight, that may vary from observation to observation. For binomial data the weights are $\omega_i = n_i$ and the constant $\phi = 1$. Other link functions that relate the mean to the linear predictor are $g(\mu) = \mathbf{m}(\mathbf{x})^t\boldsymbol{\theta}$, the log for a Poisson distributed response, or the reciprocal for a gamma distributed response, etc. Note that for the special case of normally distributed responses, with the identity link function, the weights equal 1, and the constant $\phi = \sigma^2$. The GLM reduces to the linear regression models considered in Chapters 4 and 6.
When $n_i > 1$, the actual variance of binary or count data is often larger than that associated with the binomial or Poisson model. This extra binomial variation is also called overdispersion and might be due to, for example, unobserved heterogeneity not taken into account by the covariates in the linear predictor, or due to a positive correlation between individual binary responses, e.g. experimental units that belong to the same cluster (e.g. like litter, family, etc). Overdispersion can be taken into account by allowing ϕ to be a free overdispersion parameter that has to be estimated. For more details on overdispersion, the reader is referred to the references in McCullagh and Nelder (1989) and Fahrmeir and Tutz (1994). In our context, however, we will assume that no overdispersion is present so that the $\varphi-$parameter always equals 1.

In the more general context, possible definitions of the residuals are

$$\begin{aligned} y_i - \hat{\mu}_i & \qquad \text{(raw residuals)} \\ \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}} & \qquad \text{(Pearson residuals)} \\ \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\hat{\phi}(l_i(y_i) - l_i(\hat{\mu}_i))} & \qquad \text{(Deviance residuals)} \end{aligned}$$

where $l_i(\hat{\mu}_i)$ is the contribution of the i^{th} covariate pattern to the overall log likelihood $l(\mu)$, in terms of the estimated mean $\hat{\mu}_i$. For example, for gamma distributed responses, the deviance residuals are given by

$$e_{d,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{-2(\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i)}.$$

All previous definitions of regional residuals and corresponding tests and plots can now immediately be reformulated for the raw, the Pearson and the deviance residuals in the more general context of GLM. We illustrate this with a gamma distribution example from McCullagh and Nelder (1989) pp. 300-302.

7.3.1 Clotting times of blood

Hurn et al. (1945) published data on clotting time of blood, giving clotting time in seconds, y, for normal plasma diluted to nine different percentage concentrations with prothrombin-free plasma, u. Clotting was induced by two lots of thromboplastin, L. Both lots are analysed using a GLM with inverse link function and the gamma distribution. The linear predictor includes the main effects and interaction term of $\log(u)$ and the factor lots L, allowing for different intercepts and slopes for the two lots,

$$\mu_i^{-1} = \theta_0 + \theta_1 \log(u_i) + \theta_2 L_i + \theta_3 \log(u_i) L_i.$$

Figure 7.9 (upper panel) shows the data and the weighted linear least squares fit. In the middle panel, a raw residual plot is shown and an unsatisfactory fit is observed for residuals with low percentage concentrations with prothrombin-free plasma. The regional residual test based on raw residuals is applied and a bootstrap p-value of $p = 0.0004$ is found. The p-value is approximated based on 10000 parametric bootstrap samples from a gamma distribution with parameters equal to those estimated by the GLM model. The lower panel shows the formal regional residual plot, where centers are ordered with respect to $\log(u)$. The regional residual with the largest absolute value corresponds to the experimental unit with $u = 5$ in lot 1 (red dot). At the 5% level of significance, a significant overestimation of the data is thus found for this design point. The yellow dots in the formal regional residual plot correspond to large subsets including almost all design points, except for $u = 5$, for both lots or for lot 1 only. It turns out that the observed values are not consistent with the recorded concentration $u = 5$, but they are entirely consistent with $u = 6$ (McCullagh and Nelder, 1989). When the regional residual test is applied after correcting the design, the bootstrap p-value is p=0.3766, and no evidence is found against the postulated hypothesis.

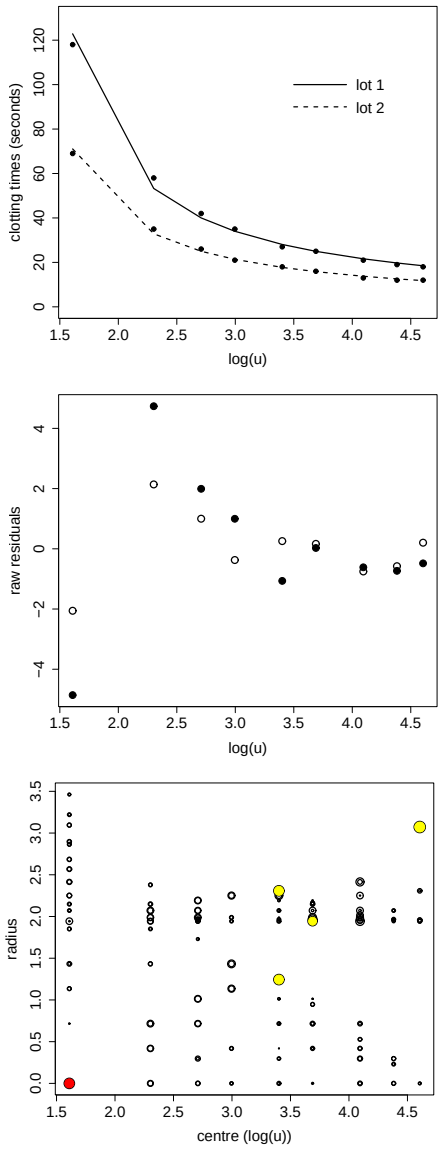

FIGURE 7.9: *(Upper panel) Clotting data and GLM fit with inverse link function and gamma distribution. (Middle panel) Raw residuals; full dots for lot 1, circles for lot 2. (Lower panel) Formal regional residual plot for the clotting data based on raw residuals ($p = 0.0004$).*

7.4 Conclusions

The lack-of-fit tests based on regional residuals and corresponding regional residual plots are extended to the complete class of generalized linear models. Simulations in the logistic regression context strongly suggest that the power of the proposed testing procedures are at least comparable to the power of popular classical methods. As before, our methods are particularly sensitive to local LOF. Regional residual plots again formally locate the LOF in the predictor space.

CHAPTER 8

Large sample properties

The asymptotic behaviour of several test statistics from previous chapters is investigated based on results of the marked empirical process of residuals, which is studied by, e.g., Su and Wei (1991), Diebolt (1995), Stute (1997) and Diebolt and Zuber (1999). To keep the asymptotics as lucid as possible, we start in Section 8.1 with the deduction of the limiting distribution in case of the no-effects hypothesis of the regional residual tests, the RRUnij (Equation 4.14), RRS (Equation 4.5), RRD and RRP (Equation 4.6) tests and the RRC (Equation 5.1) test, where regional residuals are calculated over intervals on the real line, or over arcs on the circle. A numerical example illustrates the rate of convergence of the empirical to the asymptotic null distribution in small sample sizes. The rather slow convergence suggests the use of the bootstrap throughout this dissertation. For more complex models, the standardization can take very complicated expressions. Therefore, we start in Section 8.2 with the deduction of the limiting distribution of the unstandardized test, the RRU test, for more general regression models. We only provide some thoughts on how to obtain the asymptotic null distribution of the standardized tests. Finally, the consistency of the supremum based test is shown in Section 8.3.

8.1 Limiting distribution of RR test statistics under the no-effect hypothesis

The asymptotic behaviour of the RRUnij, RRS, RRR, RRG and the RRC tests, where regional residuals are calculated over intervals on the real line, or over arcs on the circle, are investigated by considering the regional residuals as a function of increments of the marked empirical process of residuals described by Stute (1997) and Diebolt and Zuber (1999). In particular, the marked empirical process is defined as

$$\hat{\mathbb{B}}_n(x) = n^{-1/2} \sum_{i=1}^{n} (y_i - m(x_i; \hat{\boldsymbol{\theta}}_n)) I(x_i \leq x), \quad x \in \mathbb{R}, \tag{8.1}$$

where I is the indicator function, and $\{\hat{\boldsymbol{\theta}}_n\}$ denotes a sequence of $n^{1/2}$-consistent estimators of $\boldsymbol{\theta}$, e.g. $\hat{\boldsymbol{\theta}}_n$ is the least-squares estimator (LSE).

To keep the asymptotics as lucid as possible, we will follow in this section the outline of Zuber (1996) and only consider the no-effect null hypothesis, $H_0 : m(x;\theta) \equiv \theta_0$ for $x \in \mathbb{R}$, with a fixed, uniform design, and homoscedastic error terms. The mean θ_0 is consistently estimated by the sample mean, i.e. $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. However, the asymptotic behaviour under the more general regression model $y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n$, where m belongs to a given parametric family of functions, $H_0 : m \in \mathcal{M} = \{m(.,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$, can be established along the same lines as in Stute (1997) or Diebolt and Zuber (1999), and is deferred to Section 8.2.

8.1.1 Linear-linear regression

The regional residual in any interval A_{st}, $s < t$ so that at least one $x_i \in (s,t]$, can be written as a function of increments of the process $\hat{\mathbb{B}}_n$,

$$\begin{aligned}\sqrt{n}R(A_{st}) &= \frac{\sqrt{n}\sum_{i=1}^{n} I(x_i \in A_{st})(y_i - \hat{\theta}_n)}{\sum_{i=1}^{n} I(x_i \in A_{st})} \\ &\equiv \hat{\mathbb{H}}_n(s,t) = \frac{n}{\sum_{i=1}^{n} I(x_i \in A_{st})}\left(\hat{\mathbb{B}}_n(t) - \hat{\mathbb{B}}_n(s)\right).\end{aligned}$$

Without loss of generality, we assume further that the variable x is restricted to the unit interval $[0,1]$. By considering a fixed, uniform design, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(x_i \in A_{st}) = t - s$. For notational convenience, we rewrite the process $\hat{\mathbb{H}}_n(s,t)$ as $\hat{\mathbb{H}}_n(s,t) = \frac{1}{t-s}\left(\hat{\mathbb{B}}_n(s,t)\right) = \frac{1}{t-s}\left(\hat{\mathbb{B}}_n(t) - \hat{\mathbb{B}}_n(s)\right)$. Both representations result in the same asymptotic properties.

Theorem 3 *Let $0 < c < 1$ denote a small nonzero constant, and define $\mathcal{S} = \{(s,t) \in [0,1]^2 : c < t - s\}$. Then, under the no-effect null hypothesis, the stochastic process $\hat{\mathbb{H}}_n(s,t)$ converges weakly to $\frac{1}{t-s}\sigma\left(\mathbb{Z}(t) - \mathbb{Z}(s)\right)$ over $\mathcal{S}$, with $\mathbb{Z}$ a standard Brownian Bridge on $[0,1]$.*

Proof. Let $\mathbb{Z}(s,t) = \mathbb{Z}(t) - \mathbb{Z}(s)$, for $s,t \in [0,1]$. Theorem 2 in Zuber (1996) establishes the weak convergence of $\hat{\mathbb{B}}_n(t) = \mathbb{B}_n(t) - \mathbb{B}_n(1)t$, to $\sigma\mathbb{Z}$, with $\mathbb{Z}$ a standard Brownian Bridge on [0,1]. Therefore,

$$\begin{aligned}\sup_{s<t}\left|\hat{\mathbb{B}}_n(s,t) - \sigma\mathbb{Z}(s,t)\right| &= \sup_{s<t}\left|\hat{\mathbb{B}}_n(t) - \sigma\mathbb{Z}(t) - (\hat{\mathbb{B}}_n(s) - \sigma\mathbb{Z}(s))\right| \\ &\leq \sup_{t}\left|\hat{\mathbb{B}}_n(t) - \sigma\mathbb{Z}(t)\right| + \sup_{s}\left|\hat{\mathbb{B}}_n(s) - \sigma\mathbb{Z}(s)\right| \\ &\xrightarrow{p} 0 \text{ as } n \to \infty,\end{aligned}$$

by the Skorokhod construction. Since $\hat{\mathbb{H}}_n(s,t) = \frac{1}{t-s}\left(\hat{\mathbb{B}}_n(s,t)\right)$, Theorem 3 follows when s is bounded away from t. □

The following theorem states the asymptotic null distribution of the RRUnij test statistic in case of the no-effect null hypothesis.

Theorem 4 *Let $\mathbb{Z}$ denote a standard Brownian Bridge on [0,1], let $0 < c < 1$ denote a small nonzero constant, and define $\mathcal{S} = \{(s,t) \in [0,1]^2 : c < t - s\}$. Then, under the no-effect null hypothesis, the test statistic T_{RRUnij} converges in distribution to the supremum norm of $\frac{1}{(t-s)}(\mathbb{Z}(t) - \mathbb{Z}(s))$ over $\mathcal{S}$.*

The proof of Theorem 4 follows by Theorem 3 and the continuous mapping theorem. The condition $c < t - s$ is necessary to let T_{RRUnij} have a proper limiting distribution. The reason is that the weight function $1/(t-s)$ gets too large for small $t - s$. A more formal argument is given in the proof of Theorem 17.2.1 of Shorack and Wellner (1986). Note that in fact, the definition of the T_{RRUnij} needs a slight modification. We additionally assume that $n_{ij} > cn$. However, in practice this assumption always holds, since the test statistic is defined over the design points and even when i equals j, there exists such a constant c.

For the regional residual tests based on standardized regional residuals, we proceed as follows. Since $\hat{\theta}_n$ is a consistent estimator, and by Theorem 3, it follows that $\hat{\mathbb{H}}_n(s,t)$ has asymptotically mean zero and variance $\frac{\sigma^2(1-(t-s))}{t-s}$. Straightforward algebraic calculations show that the standard deviation of the regional residual in any interval A_{st} equals $\frac{\sigma\sqrt{1-(t-s)}}{\sqrt{t-s}}$ when the no-effect hypothesis holds. Therefore, the standardized regional residual corresponds to the standardized process $\frac{\sqrt{t-s}\,\hat{\mathbb{H}}_n(s,t)}{\sigma\sqrt{1-(t-s)}}$. The next theorem provides its limiting process.

Theorem 5 *Let $0 < c < 1$ denote a small nonzero constant, and define $\mathcal{S} = \{(s,t) \in [0,1]^2 : c < t - s < 1 - c\}$. Then, under the no-effect null hypothesis, the stochastic process $\frac{\sqrt{t-s}\,\hat{\mathbb{H}}_n(s,t)}{\sigma\sqrt{1-(t-s)}}$, converges weakly to $\frac{1}{\sqrt{(t-s)(1-(t-s))}}(\mathbb{Z}(t) - \mathbb{Z}(s))$ over $\mathcal{S}$, with $\mathbb{Z}$ a standard Brownian Bridge on $[0,1]$.*

Proof. Theorem 5 immediately follows by applying Theorem 2 in Zuber (1996). □

In practice, σ^2 is usually not known, and has to be replaced by a consistent estimator. This does not affect the convergence of the process, as is shown in the next theorem.

Theorem 6 *Let $\hat{\sigma}^2$ denote a consistent estimator of σ^2 under the null hypothesis. Let $0 < c < 1$ denote a small nonzero constant, and define $\mathcal{S} = \{(s,t) \in [0,1]^2$:*

$c < t - s < 1 - c\}$. *Then, under the no-effect null hypothesis, the stochastic process* $\frac{\sqrt{t-s}\hat{\mathbb{H}}_n(s,t)}{\hat{\sigma}\sqrt{1-(t-s)}}$, *converges weakly to* $\frac{1}{\sqrt{(t-s)(1-(t-s))}}(\mathbb{Z}(t) - \mathbb{Z}(s))$ *over* $\mathcal{S}$, *with* $\mathbb{Z}$ *a standard Brownian Bridge on* $[0,1]$.

Proof. Theorem 6 immediately follows by Theorem 5 and Slutsky's Lemma. □

In Theorem 7, we establish the convergence of the test statistics T_{RRS}, T_{RRD} and T_{RRP} under the no-effect null hypothesis.

Theorem 7 *Let* $\mathbb{Z}$ *denote a standard Brownian Bridge on [0,1], and let* $0 < c < 1$ *denote a small nonzero constant, and define* $\mathcal{S} = \{(s,t) \in [0,1]^2 : c < t - s < 1 - c\}$. *Then, under the no-effect null hypothesis, the test statistic* T_{RRS} *converges in distribution to the supremum norm of* $\frac{1}{\sqrt{(t-s)(1-(t-s))}}(\mathbb{Z}(t) - \mathbb{Z}(s))$ *over* $\mathcal{S}$.

Theorem 6 holds for the consistent estimators S_n^2, $\hat{\sigma}_D^2$, and $\hat{\sigma}_P^2$ of σ^2 (e.g. Van Der Vaart (1998), Eubank and Hart (1992), Gasser et al. (1986)). The proof of Theorem 7 follows by Theorem 6 and the continuous mapping theorem. The condition $c < t - s < 1 - c$ is necessary to let T_{RRS}, T_{RRD} and T_{RRP} have a proper limiting distribution. Note that in fact, the definition of the T_{RRS}, T_{RRD} and T_{RRP} need a slight modification. We additionally assume that $n_{ij} > cn$.

8.1.2 Circular-linear regression

The proof of Theorem 7 in case of linear-circular regression analysis follows immediately. The regional residual in any arc A_{st}, s, t so that at least one $x_i \in (s,t]$, can be written as a function of increments of the process $\hat{\mathbb{B}}_n$,

$$\begin{aligned}\sqrt{n}R(A_{st}) &= \frac{\sqrt{n}\sum_{i=1}^n I(x_i \in A_{st})(y_i - \hat{\theta}_n)}{\sum_{i=1}^n I(x_i \in A_{st})} \\ &\equiv \hat{\mathbb{H}}_n(s,t) = \begin{cases} \frac{n}{\sum_{i=1}^n I(x_i \in A_{st})}\left(\hat{\mathbb{B}}_n(t) - \hat{\mathbb{B}}_n(s)\right) & \text{if } s < t; \\ -\frac{n}{\sum_{i=1}^n I(x_i \in A_{st})}\left(\hat{\mathbb{B}}_n(s) - \hat{\mathbb{B}}_n(t)\right) & \text{if } s > t. \end{cases}\end{aligned}$$

The last equality is obtained since all residuals sum to zero. Therefore, it suffices to consider only the case $s < t$ to investigate the asymptotic behaviour, and everything reduces to the linear-linear regression case.

8.1.3 Speed of convergence

In this section the speed of convergence is investigated empirically in a simulation study for small sample sizes. Consider the no-effect regression model with $m(x_i;\theta) = 1$ and an equidistant design $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$. The errors are independent, random normal variables with mean 0 and common variance

$\sigma^2 = 0.2$. The variance σ^2 is treated both as a known and unknown parameter and is estimated by the natural estimator, $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$. The Brownian Bridge in Theorem 2 is approximated based on 1000 time steps.
1000 samples of sizes 20, 50 and 100 were generated from the null model $m(x_i;\theta) = 1$ and $\sigma^2 = 0.2$. If the variance is known, the QQ-plots in the left panels in Figure 8.1 show a rather slow convergence of the test statistic to its asymptotic distribution. When using a consistent estimator of the residual variance, an even slower convergence is noticed (Figure 8.1, right panels). We conclude that for normally distributed error terms, the convergence is slow. This conclusion suggests the use of the bootstrap throughout this dissertation.

8.2 More general regression models

In this section, we consider the more general regression model $y_i = m(x_i) + \epsilon_i, \quad i = 1,\ldots,n$, where m belongs to a given parametric family of functions, $H_0 : m \in \mathcal{M} = \{m(.,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$. In what follows, we will refer to this null hypothesis as the parametric null hypothesis, to have a clear distinction between this and the no-effect null hypothesis in the previous section.
The asymptotic properties are established along the same lines as in Stute (1997) or Diebolt and Zuber (1999). The form of the limiting centered Gaussian process $\mathbb{H}$ of $\hat{\mathbb{H}}_n$ under H_0 is established. In what follows, the error distribution is even allowed to be heteroscedastic, although it was not considered as such in this dissertation. As this section is mainly an extension of the work of Diebolt and Zuber (1999), we report the results as generally as possible.
We start again with the limiting distribution of the RRUnij statistic and end this section with some thoughts concerning the asymptotic null distributions of the regional residual tests based on standardized regional residuals calculated over intervals. Under H_0, and the assumptions 1 - 5 listed in Chapter 3, Theorem 1 establishes the limiting centered Gaussian process $\mathbb{B}$ of $\hat{\mathbb{B}}_n$. The remainder of this section is based on this theorem.

8.2.1 RR test statistics based on unstandardized regional residuals

For convenience, we denote any interval on $[0,1]$ as a subset $C = (s,t] \subseteq [0,1]$ with length $|C| = t - s$, and ζ denotes the collection of all these subsets. Theorem 8 gives the limiting process of $\hat{\mathbb{H}}_n$ under the parametric null hypothesis. Note that $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\varphi}_0$ are not explicitly defined in this theorem. Under Assumption 5, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges, as $n \to \infty$, to a p-dimensional normal random variable with zero mean and variance matrix $\boldsymbol{\Gamma}_0 = \int_{-\infty}^{\infty} \boldsymbol{\varphi}_0 \boldsymbol{\varphi}_0^T dF$. Corollary 1 in Diebolt and Zuber (1999) shows under some additional technical assumptions, that the LSE sequence $\hat{\boldsymbol{\theta}}_n$ is a sequence of $n^{1/2}$-consistent esti-

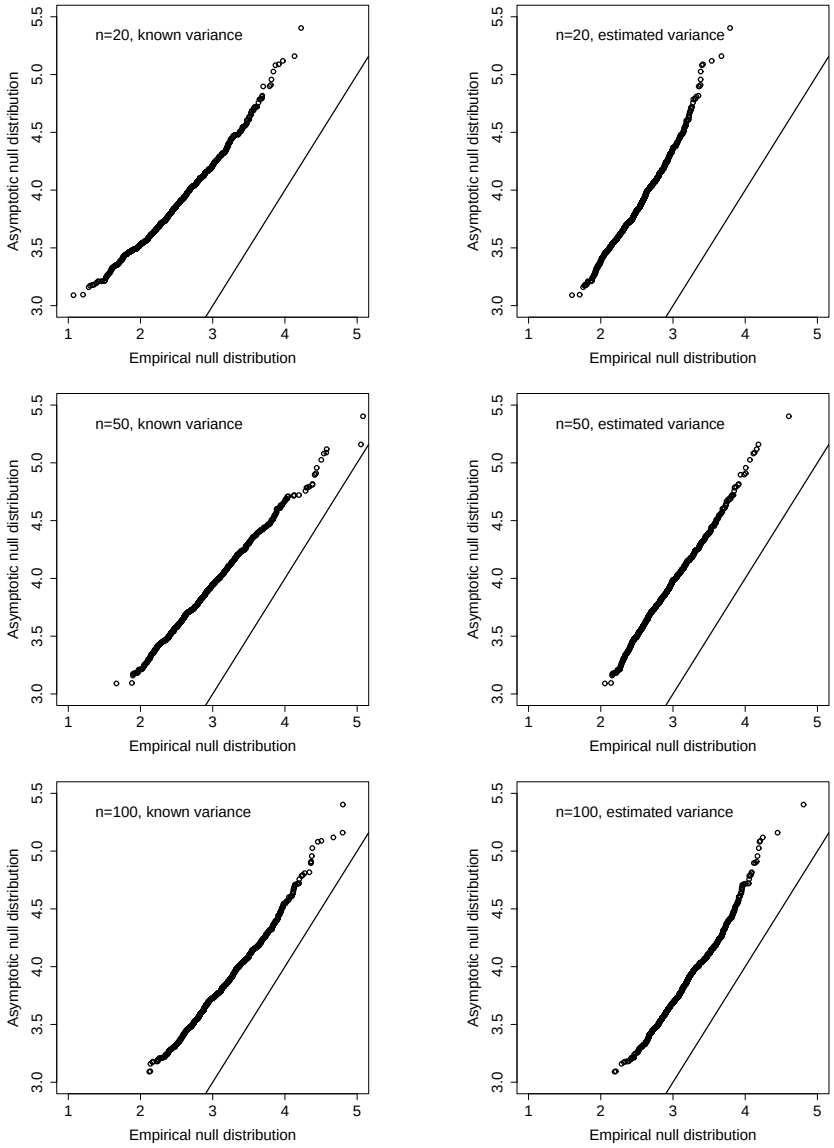

FIGURE 8.1: *QQ-plots of the empirical and the asymptotic null distribution of the TRRS regional residual test for the no-effect null hypothesis with constant variance. This is done for three different small sample sizes $n = 20$ (upper panels), $n = 50$ (middle panels), and $n = 100$ (lower panels), and for known (left panels) and estimated (right panels) variances.*

mators of $\boldsymbol{\theta}$ and satisfies the asymptotic linearity type property (Assumption 5) with $\boldsymbol{\varphi}_0 = \sigma \mathbf{V}_0^{-1} \boldsymbol{\nabla} \mathbf{m}_0$, where $\mathbf{V}_0 = \int_{-\infty}^{\infty} \boldsymbol{\nabla} \mathbf{m}_0 \boldsymbol{\nabla} \mathbf{m}_0^T dF$ is a symmetric positive-definite matrix.

Theorem 8 *Let $0 < c < 1$ denote a small nonzero constant, and $C_1, C_2 \in \zeta$ so that $c < |C_1|\,|C_2|$. Under the parametric null hypothesis and the Assumptions 1 - 5, $\hat{\mathbb{H}}_n \xrightarrow{w} \mathbb{H}$, as $n \to \infty$, in the space $D[0,1]$, where $\mathbb{H}$ is a centered zero mean gaussian process with covariance function*

$$r(C_1, C_2) = \frac{1}{|C_1|\,|C_2|} \Big(G(C_1 \cap C_2) - \mathbf{g}_0^T(C_1)\mathbf{h}_0(C_2) - \mathbf{g}_0^T(C_2)\mathbf{h}_0(C_1) + \mathbf{g}_0^T(C_1)\boldsymbol{\Gamma}_0\mathbf{g}_0(C_2) \Big), \quad (8.2)$$

where

$$\begin{aligned} G(C) &= \int_C \sigma^2(u) dF(u), \\ \mathbf{g}_0(C) &= \int_C \boldsymbol{\nabla} \mathbf{m}_0(u) dF(u), \\ \mathbf{h}_0(C) &= \int_C \sigma(u) \boldsymbol{\varphi}_0(u) dF(u), \end{aligned}$$

with $\boldsymbol{\nabla} \mathbf{m}_0 = \boldsymbol{\nabla} \mathbf{m}_{\boldsymbol{\theta}|\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ the gradient with respect to $\boldsymbol{\theta}$ of $m(x, \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$.

Proof. The proof is immediate by applying the continuous mapping theorem and Theorem 1. □

In Theorem 9 we establish the limiting distribution of T_{RRUnij} under the parametric null hypothesis.

Theorem 9 *Let $\mathbb{Z}$ denote a standard Brownian Bridge on [0,1], and let $0 < c < 1$ denote a small nonzero constant, and $C_1, C_2 \in \zeta$ so that $c < |C_1|\,|C_2|$. Then, under the parametric null hypothesis and the Assumptions 1 - 5, the test statistic T_{RRUnij} converges in distribution to the supremum norm of the centered zero mean gaussian process $\mathbb{H}$, with covariance structure defined in Theorem 8.*
Proof. The proof is immediate by applying Theorem 8. □

To perform the regional residual tests, we need to estimate $\mathbf{V}_0$, $\mathbf{g}_0$, $\mathbf{h}_0$ and G. The strong consistency under H_0 of estimators of $\mathbf{V}_0$, $\mathbf{g}_0$, $\mathbf{h}_0$ and G is proved in Theorem 10.

Theorem 10 *Under the assumptions of Theorem 2 of Diebolt and Zuber (1999),*

$$\hat{\mathbf{V}}_n = \int_0^1 \boldsymbol{\nabla}\mathbf{m}_{\hat{\boldsymbol{\theta}}_n}(u)\boldsymbol{\nabla}\mathbf{m}^T_{\hat{\boldsymbol{\theta}}_n}(u)dF(u) \overset{a.s.}{\to} \int_0^1 \boldsymbol{\nabla}\mathbf{m}_0(u)\boldsymbol{\nabla}\mathbf{m}^T_0(u)dF(u) = \mathbf{V}_0, \quad n \to \infty,$$

$$\hat{\mathbf{g}}_n(C) = \int_C \boldsymbol{\nabla}\mathbf{m}_{\hat{\boldsymbol{\theta}}_n}(u)dF(u) \overset{a.s.}{\to} \int_C \boldsymbol{\nabla}\mathbf{m}_0(u)dF(u) = \mathbf{g}_0(C), \quad n \to \infty, \textit{ uniformly in s,t,}$$

$$\hat{G}_n(C) = n^{-1}\sum_{i=1}^n (Y_i - m(x_i, \hat{\boldsymbol{\theta}}_n))^2 I(s < x_i \leq t) \overset{a.s.}{\to} \int_C \sigma^2(u)dF(u) = G(C), \quad n \to \infty, \textit{ uniformly in s,t,}$$

$$\begin{aligned}\hat{\mathbf{h}}_n(C) &= \hat{\mathbf{V}}_n^{-1} n^{-1}\sum_{i=1}^n (Y_i - m(x_i, \hat{\boldsymbol{\theta}}_n))^2 \boldsymbol{\nabla}\mathbf{m}_{\hat{\boldsymbol{\theta}}_n}(x_i) I(s < x_i \leq t) \\ &\overset{a.s.}{\to} \mathbf{V}_0^{-1}\int_C \sigma^2 \boldsymbol{\nabla}\mathbf{m}_0(u)dF(u) = \mathbf{h}_0(C), \quad n \to \infty, \textit{ uniformly in s,t.}\end{aligned}$$

Proof. The proof is immediate by applying Theorem 2 of Diebolt and Zuber (1999) and the triangle inequality. □

8.2.2 RR test statistics based on standardized regional residuals

For the regional residual tests based on standardized regional residuals, we only provide some guidelines to obtain the asymptotic null distribution. As for all possible intervals $C \in \zeta$, standardized regional residuals correspond to the process

$$\frac{\hat{\mathbb{H}}_n(C)}{\sqrt{\hat{r}_n(C)}}, \tag{8.3}$$

where $\hat{r}_n(C)$ is the sample estimator of the variance function $r(C)$ in Theorem 8 and which depends on the design and on the parametric family of regression models under the null hypothesis. We believe that the limiting distribution of this standardized process can be written as

$$\frac{\hat{\mathbb{H}}_n(C)}{\sqrt{\hat{r}_n(C)}} \xrightarrow{w} \frac{\hat{\mathbb{H}}(C)}{\sqrt{\hat{r}(C)}}$$

where $\sqrt{\hat{r}_n(C)}$ is a bounded nonzero function. We do not provide a formal proof.

To illustrate the correspondence of the standardized process and the standardized regional residuals, we show in what follows the asymptotic equivalence of the variance structure of the process $\hat{\mathbb{H}}$ and the regional residuals calculated over intervals for both the no-effect and linear hypothesis. This is done by first simplifying the expression for the variance function of the process $\mathbb{H}$ for both the no-effect and the linear hypothesis. Secondly, the limit of the variance function for $\sqrt{n}R(A_{ij})$ is determined for $n \to \infty$.

No-effect hypothesis

The no-effect null hypothesis states

$$H_0 : m(x;\theta) = \theta_0, \quad x \in [0,1].$$

As $\nabla m_0(x) = 1$, the functions $g_0(x)$, V_0, $\varphi(x)$ and $h_0(x)$ become respectively $F(x)$, 1, $\sigma(x)$ and $\int_0^x \sigma^2(u)dF(u)$. In case of homoscedasticity, $C_1 = (s,t]$ and $C_2 = (v,w]$ both in ζ, and for F the cumulative distribution of a uniform random variable over the interval $[0,1]$, the covariance function of $\hat{\mathbb{H}}$ simplifies to

$$r((s,t],(v,w]) = \frac{\sigma^2}{|t-s|\,|w-v|}\left([|(s,t] \cap (v,w]| - |t-s|\,|w-v|\right).$$

Therefore, the variance of the process $\hat{\mathbb{H}}$ in interval C becomes

$$\text{var}(\hat{\mathbb{H}}(C)) = \frac{\sigma^2}{|\,C\,|}(1- |\,C\,|),$$

as was found in Section 8.1. For the regional residuals, we find

$$\text{var}(\sqrt{n}R(A_{ij})) = \sigma^2\frac{n}{n_{ij}} + \frac{1}{n_{ij}} - 1 - \frac{1}{n_{ij}^2} \overset{n\to\infty}{\to} \sigma^2\frac{1}{j-i}\left(1-(j-i)\right)$$

Linear effect hypothesis
Consider the linear effect hypothesis

$$H_0 : m(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x, \quad x \in [0,1].$$

If the variance function $\sigma^2(x) = \sigma^2$ for each $x \in [0,1]$, and F is the cumulative distribution function of a uniformly distributed random variable over [0,1], the vector functions $\boldsymbol{\nabla}\mathbf{m}_0$, $\mathbf{g}_0$ and $\mathbf{h}_0$, and the matrices $\mathbf{V}_0$ and $\boldsymbol{\Gamma}_0$, become

$$
\begin{aligned}
\boldsymbol{\nabla}\mathbf{m}_0^T(x_i) &= \left(\frac{\partial m}{\partial\theta_1}(x_i;\boldsymbol{\theta}), \frac{\partial m}{\partial\theta_0}(x_i;\boldsymbol{\theta})\right) = (x \quad 1\ ) \\
\mathbf{g}_0(x) &= \int_0^x \boldsymbol{\nabla}\mathbf{m}_0^T(y)dF(y) = \begin{pmatrix} \int_0^x y dF(y) \\ \int_0^x 1 dF(y) \end{pmatrix} = \begin{pmatrix} \frac{x^2}{2} \\ x \end{pmatrix} \\
\boldsymbol{\nabla}\mathbf{m}_0(x)\boldsymbol{\nabla}\mathbf{m}_0^T(x) &= \begin{pmatrix} x \\ 1 \end{pmatrix}(x \quad 1) = \begin{pmatrix} x^2 & x \\ x & 1 \end{pmatrix} \\
\mathbf{V}_0 &= \int_0^1 \boldsymbol{\nabla}\mathbf{m}_0(y)\boldsymbol{\nabla}\mathbf{m}_0^T(y)dF(y) \\
&= \begin{pmatrix} \frac{x^3}{3} & \frac{x^2}{2} \\ \frac{x^2}{2} & x \end{pmatrix}\Bigg|_0^1 = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \\
\mathbf{V}_0^{-1} &= \frac{\text{adj}(\mathbf{V}_0)}{\det(\mathbf{V}_0)} = 12\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 12 & -6 \\ -6 & 4 \end{pmatrix} \\
\mathbf{h}_0(x) &= \mathbf{V}_0^{-1}\int_0^x \sigma^2\boldsymbol{\nabla}\mathbf{m}_0^T(y)dF(y) = \mathbf{V}_0^{-1}\sigma^2\mathbf{g}_0(x) \\
&= \begin{pmatrix} 12 & -6 \\ -6 & 4 \end{pmatrix}\sigma^2\begin{pmatrix} \frac{x^2}{2} \\ x \end{pmatrix} = \sigma^2\begin{pmatrix} 6x^2 - 6x \\ -3x^2 + 4x \end{pmatrix} \\
G(x) &= \int_0^x \sigma^2 dF(y) = \sigma^2 F(x) = \sigma^2 x \\
\boldsymbol{\varphi}_0(x_i) &= \sigma\mathbf{V}_0^{-1}\boldsymbol{\nabla}\mathbf{m}_0(x) \\
\boldsymbol{\Gamma}_0 &= \int_0^1 \boldsymbol{\varphi}_0(x_i)\boldsymbol{\varphi}_0^T(x_i)dF(x) \\
&= \mathbf{V}_0^{-1}\sigma^2\left(\underbrace{\int_0^1 \boldsymbol{\nabla}\mathbf{m}_0(y)\boldsymbol{\nabla}\mathbf{m}_0^T(y)dF(y)}_{\mathbf{V}_0}\right)\mathbf{V}_0^{-1} = \sigma^2\mathbf{V}_0^{-1}
\end{aligned}
$$

The variance of the process $\hat{\mathbb{H}}$ in an interval $C = (s,t]$ is

$$
\begin{aligned}
r((s,t],(s,t]) &= \frac{1}{(t-s)^2}\Big(G((s,t]) - \mathbf{g}_0^T(s,t)\mathbf{h}_0(s,t) \\
&\quad -\mathbf{g}_0^T(s,t)\mathbf{h}_0(s,t) + \mathbf{g}_0^T(s,t)\boldsymbol{\Gamma}_0\mathbf{g}_0(s,t)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{(t-s)^2}\left(\sigma^2(F(t)-F(s)) - \mathbf{g}_0^T(s,t)\mathbf{h}_0(s,t)\right) \\
&= \frac{\sigma^2}{(t-s)^2}\Big((t-s) - (3(t^2-s^2)^2 - 6(t-s)(t^2-s^2) \\
&\quad +4(t-s)^2)\Big) \\
&= \frac{\sigma^2}{(t-s)}\left(1-(t-s)(3(t+s)^2-6(t+s)+4)\right).
\end{aligned}
$$

For the variance of the standardized regional residuals in an interval A_{ij}, we find after some straightforward, but lengthy calculations, a rather complex expression. Only the terms of asymptotic importance are shown in the expression below,

$$
\mathrm{var}(\sqrt{n}R(A_{ij})) \approx \sigma^2\left(\frac{1}{n_{ij}} - \frac{4n^2n_{ij}^2 - 6nn_{ij}(j^2-i^2) + 3(j^2-i^2)^2}{n^3n_{ij}^2}\right).
$$

For $n \to \infty$, we obtain

$$
\mathrm{var}(\sqrt{n}R(A_{ij})) \overset{n\to\infty}{\to} \frac{\sigma^2}{j-i}\left(1-(j-i)(3(j+i)^2-6(j+i)+4)\right).
$$

In the above calculations, we thus find that the asymptotic equivalence of the variance structure of the process $\mathbb{I}\hat{\mathrm{H}}$ and the regional residuals calculated over intervals, for both the no-effect and linear hypothesis. It also illustrates that the covariance structure can become very complicated for more complex regression models.

8.3 Consistency of the regional residual tests

The proof of the consistency of the regional residual tests presented here, is established along the same lines as the one presented in Su and Wei (1991) for tests based on the supremum of cumulative sums of residuals. For clarity of notation, recall that $m(\mathbf{x})$ denotes the true conditional mean of y given $\mathbf{x}$. The central null hypothesis states that m belongs to a given parametric family of functions,

$$
H_0 : m \in \mathcal{M} = \{m(\mathbf{x},\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\},
$$

where Θ is a p-dimensional proper parameter set in $\mathbb{R}^p$. The alternative hypothesis H_1 which we are interested in testing against, is that there does not exist a $p \times 1$ constant vector $\boldsymbol{\theta}$ such that $m(\mathbf{x}) = m(\mathbf{x},\boldsymbol{\theta})$, for all the $\mathbf{x}$ in the sample space.

Theorem 11 *Suppose that the null hypothesis is false in the sense that there does not exist a $p \times 1$ constant vector $\boldsymbol{\theta}$ such that $m(\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta})$, for all the $\mathbf{x}$ in the sample space, and suppose the assumptions 1 - 5 hold, then the regional residual tests have power tending to 1 as $n \to \infty$.*

Proof. Under H_1, as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n$ converges in probability to a constant vector $\boldsymbol{\theta}^*$. For each subset $C_{\alpha,\beta} \in \mathbb{R}^p$, such that at least one $\mathbf{x}_i \in C_{\alpha,\beta}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}[y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]I(\mathbf{x}_i \in C_{\alpha,\beta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{[y_i - m(\mathbf{x}_i)] + [m(\mathbf{x}_i) - m(\mathbf{x}_i, \boldsymbol{\theta}^*)] + [m(\mathbf{x}_i, \boldsymbol{\theta}^*) - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]\}I(\mathbf{x}_i \in C_{\alpha,\beta}).
\end{aligned} \tag{8.4}$$

Under H_1, there exists at least one subset $C_{\alpha_0,\beta_0} \in \mathbb{R}^p$ such that

$$n^{-1}\sum_{i=1}^{n}[m(\mathbf{x}_i) - m(\mathbf{x}_i, \boldsymbol{\theta}^*)]I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \xrightarrow{p} c, \tag{8.5}$$

where c is a nonzero constant. As a proof of this statement, it is sufficient to note that $c = 0$ for all subsets $C_{\alpha,\beta} \in \mathbb{R}^p$ implies that H_0 is true, whereas it is assumed here that H_0 is not true. For this particular subset C_{α_0,β_0}, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - m(\mathbf{x}_i)]I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \xrightarrow{p} 0. \tag{8.6}$$

Furthermore, since m has a bounded derivative, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}[m(\mathbf{x}_i, \boldsymbol{\theta}^*) - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \xrightarrow{p} 0. \tag{8.7}$$

The statements (8.4), (8.5), (8.6) and (8.7) imply that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \xrightarrow{p} c.$$

Since,

$$\sup_{C_{\alpha,\beta} \in \mathbb{R}^p}\left|\frac{n_{C_{\alpha,\beta}}}{n^{-3/2}}\mathbb{I}\hat{\mathrm{H}}_n(C_{\alpha,\beta})\right| \geq n^{-1}\left|\sum_{i=1}^{n}[y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)]I(\mathbf{x}_i \in C_{\alpha_0,\beta_0})\right|,$$

it follows that $n^{-1/2}T_{RRUnij}$ converges in probability to a nonzero positive constant as $n \to \infty$, and thus that $T_{RRUnij} \to \infty$ as $n \to \infty$. This establishes the

consistency of all unstandardized tests against H_1. Note that this is not limited to the RRUnij test, as multiple predictor variables are allowed.
For regional residual tests based on standardized regional residuals, the same idea could be used if we add the factor $\frac{1}{\sqrt{\hat{r}_n(C_{\alpha,\beta})}}$ to all statements, where $\sqrt{\hat{r}_n(C_{\alpha,\beta})}$ is a bounded, non zero function. In particular, the new statements imply that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)}{\sqrt{\hat{r}_n(C_{\alpha_0,\beta_0})}} I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \xrightarrow{p} \frac{c}{\sqrt{\hat{r}_n(C_{\alpha_0,\beta_0})}}.$$

Since,

$$\sup_{C_{\alpha,\beta}\in\mathbb{R}^p} \left| \frac{n_{C_{\alpha,\beta}}}{n^{-3/2}} \frac{\hat{\mathbb{H}}_n(C_{\alpha,\beta})}{\sqrt{\hat{r}_n(C_{\alpha,\beta})}} \right| \geq n^{-1} \left| \sum_{i=1}^{n} \frac{y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n)}{\sqrt{\hat{r}_n(C_{\alpha_0,\beta_0})}} I(\mathbf{x}_i \in C_{\alpha_0,\beta_0}) \right|,$$

it follows that $n^{-1/2}T_{RR.}$ converges in probability to a nonzero positive constant as $n \to \infty$, and thus that $T_{RR.} \to \infty$ as $n \to \infty$. This establishes the consistency of all regional residual tests against H_1. □

8.4 Conclusions

The asymptotic behaviour of interval based regional residuals is established in this chapter. The limiting distribution is a centered zero mean gaussian process with a complicated covariance structure for more complex models. The speed of convergence is rather slow. Therefore, we recommend the use of bootstrap null distributions in practice. Further, the consistency against the alternative H_1 of all supremum based tests in this dissertation was shown.

CHAPTER 9

Conclusions and further research

9.1 Conclusions

An important component of any modeling procedure is an assessment of the model fit, more specifically, an evaluation of how well model-based predicted outcomes coincide with the observed data. In this work a new type of Lack-of-Fit test and corresponding diagnostic plots are proposed and discussed for parametric regression models. The tests are based on so-called regional residuals, which are averages of classical residuals in subsets of the predictor space. Regional residuals are very suitable building blocks for constructing a lack-of-fit test. If deviations from the null model occur in a certain region of the predictor space, patterns of positive or negative residuals will show up in that neighbourhood, resulting in large absolute values of standardized regional residuals over these regions. Large absolute values thus suggest a possible lack-of-fit of the hypothesized model, located in the corresponding subset in the predictor space. To overcome the problem of multiplicity and to obtain a global measure of lack-of-fit, test statistics are defined as the supremum norm of standardized or unstandardized regional residuals over all subsets. The regional residual tests are omnibus in the sense that they are consistent against all fixed alternatives. In particular, simulation studies show that the tests are sensitive to *local* deviations from the hypothesized regression model, where local refers to a small subset of the predictor space over which the true and the hypothesized models do not agree.
We believe that important information is lost by summarizing all discrepancy measures into a single value. We therefore propose to complement the LOF test with a visualization of the individual regional residuals. The new plots *formally* identify regions in the predictor space where the model does not fit well and suggest in which area remedial measures may be necessary.
Smoothing based LOF tests are in general very powerful in detecting deviations from the null model, but their performance depends on the choice of the smoothing parameter. Regional residuals are calculated over *all* possible intervals, so as to avoid the choice of this smoothing parameter. In this way,

statistical tests based on regional residuals have power to detect both global and local deviations from the null model. As a consequence, our methodology is computationally intensive, but suggestions are made in subsection 9.2.1 to reduce the computational cost.
Finally, in contrast to some classical LOF tests, regional residual tests are applicable whether replicated design points are available or not. When no replicated design points are present, a regional residual calculated over a subset at one design point is simply the ordinary residual at each design point. However, for design points with multiple measurements, this regional residual is equal to the average of all the multiple classical residuals at that design point. This means that the availability or the presence of replicated design points is not an issue for regional residual based tests.

Single linear predictor

For a single linear predictor variable, the subsets are chosen to be intervals on the real line. The corresponding standardized regional residuals can be visualized in a heat map in the (i, j) plane, where the x-axis (y-axis) shows the starting point (end point) of the interval for which the standardized regional residual is calculated. The *formal regional residual plot* protects correctly for a family-wise error rate of α by only colouring the intervals for which the absolute value of the standardized regional residual exceeds the bootstrap α-level critical value of the test statistic. Coloured areas in this plot refer to particular regions where a statistically significant under- or overestimation of the data by the null model is detected at the α-level of significance. These regions can be very small, a few neighbouring observations or even a single outlying observation, or very large in case of global deviations from the null model.

The asymptotic null distribution of the test statistic is the supremum norm of a centered, zero mean Gaussian process with a complicated covariance function. However, since the convergence is slow, the asymptotic approximation may not be appropriate for small sample sizes. Therefore, we recommend a bootstrap procedure to obtain bootstrap p-values. The use of the wild bootstrap is recommended in practice, as it handles adequately heteroscedasticity of the error terms.

We standardize the regional residuals to make them comparable among one another. In practice, the residual variance is unknown, and, therefore, needs to be replaced by a variance estimator that is consistent under both the null

and the alternative hypotheses so as to obtain a powerful LOF test. The estimator based on the residual sum of squares (S_n^2), often overestimates under a lack-of-fit situation. The estimated standardized regional residuals appear then to be smaller than they really are, which might result some power loss. The use of variance estimators which are more robust against deviations from the null model may therefore be more appropriate.

As different variance estimators may influence the performance of the regional residual based tests, we may consider test statistics based on unstandardized regional residuals. Typically, the test statistic is then the supremum norm of weighted sums of classical residuals in all possible intervals. In case of global LOF, the weight factor $1/\sqrt{n}$, where n is the sample size, results in the best performance. The factor $1/\sqrt{n_{ij}}$, where n_{ij} denotes the number of observations in the interval, makes the test statistic more sensitive to local LOF, as the weighted sums over small intervals then become relatively more important. Regional residual tests based on standardized regional residuals seem to be a nice trade-off between the latter two statistics in case of both global and local LOF. For their ease of implementation in all parametric regression models, we further only consider standardized regional residual tests based on the estimator S_n^2 of the residual variance.

Single angular predictor

Our new methodology is extremely useful when the predictor variable is angular. Although ordinary least squares regression can be used to fit circular-linear regression models, classical LOF tests for linear-linear regression models often fail to detect deviations from the hypothesized model because their p-values strongly depend on the choice of the origin of the circular variate. Our regional residual test properly detects lack-of-fit on the circle, as it is origin independent. We have also illustrated that regional residuals, which are now calculated over all possible arcs on the circle, can be used to construct a regional residual plot. Combined with the testing procedure, this graphical diagnostic tool allows both global and local deviations to be detected and localized in the angular predictor space. We have also observed good powers for the smooth test of Fan and Huang (2001), which is also origin independent. This latter feature, however, has not been recognized before.

Multiple predictor variables

Our methodology is easily extended to multiple predictor variables. We dis-

cussed two possible extensions. The first one, based on marginal information for each predictor variable, is mainly useful to detect deviations from the null model in univariate directions. The second approach, on the other hand, takes the multivariate structure of the design space into account. For a d-dimensional covariate vector, d-dimensional spheres are constructed based on a distance measure in the predictor space. In this way, a more powerful test for local deviations in the higher dimensional predictor space is constructed and allows the detection of a broader class of alternatives. The advantage of this approach is that no order relation of the residuals has to be specified in advance, neither the choice of a smoothing parameter.

Generalized linear models

The lack-of-fit tests based on regional residuals and corresponding regional residual plots are further extended to the complete class of generalized linear models. Simulations in the logistic regression context suggest that the power of the proposed testing procedures is at least comparable to the power of popular classical methods. As before, our procedure is particularly sensitive to local LOF. Regional residual plots again formally locate the LOF in the predictor space.

9.2 Further research

9.2.1 Reduction of the computational cost

It would be most welcome to reduce the computational cost of our tests. In case of very large datasets that nowadays often occur, the proposed methodology would be too time consuming. Instead of calculating the regional residuals over all possible intervals with respect to a certain predictor variable, or over all possible spheres in the d-dimensional predictor space, only a selection of these regions could be studied. One could, for example, randomly select a number of these regions, as is illustrated in the POPS data example in Section 7.2.3. A lot of regional residuals contain the same information, so we believe that only including a randomly selected subset of all regions, provides reliable results. Of course, further investigation is necessary to obtain practical guidelines.

Another possible idea would be to follow the ideas of e.g. Landwehr et al. (1984) and Moons et al. (2004). They expect that if local deviations occur, it might be detected by considering close observations in the predictor space. Landwehr et al. (1984) construct clusters of similar observations in the predic-

tor space, while Moons et al. (2004) construct groups based on the recursive partitioning algorithm underlying classification trees. For both suggestions, however, many choices have to be made for the practical implementation. For example, the number of clusters, the choice of partitioning scheme and pruning process, including the number of final nodes and the number of observations in final nodes.

9.2.2 Categorical predictor variables

In this thesis we did not explicitly discuss categorical predictor variables. When the group levels are coded by real numbers, one could simply include them in the Euclidean distance measure and probably this procedure works fine. However, the choice of the subsets then depends on the codes assigned to the group levels. Therefore, properly handling categorical covariates would be an improvement. le Cessie and van Houwelingen (1995) provide some useful suggestions on handling categorical covariates. One could, for example, use a distance measure based on the number of categorical variables on which the observations differ. In particular, suppose the i^{th} covariate vector of observations consists of k categorical covariates, say $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$. Let c_j be the number of different categories of the j^{th} categorical variable. le Cessie and van Houwelingen (1995) define the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between observation i and j by

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\frac{c_1}{c_1 - 1} I(x_{i1} \neq x_{j1}) + \ldots + \frac{c_k}{c_k - 1} I(x_{ik} \neq x_{jk})},$$

where

$$I(\ldots) = \begin{cases} 1 & \text{if the proposition inside the brackets is true;} \\ 0 & \text{if it is false.} \end{cases}$$

The factors $\frac{c_k}{c_k - 1}$ are to adjust for the number of different categories of a variable. If one is faced with both categorical and continuous predictor variables, le Cessie and van Houwelingen (1995) use a modified distance measure, which is a combination of the Euclidean distance between the continuous and the distance measure for categorical variables as described above. The contributions for the continuous covariates in the distance measure are divided by two times the variance of the covariate. In this way the average of each term equals 1. We refer the reader to le Cessie and van Houwelingen (1995) for more details.

9.2.3 Spherical regional residuals in circular-linear regression

The spherical regional residuals of Section 6.2 may be extended to circular-linear regression in several ways. In Section 6.2 we have chosen to use the

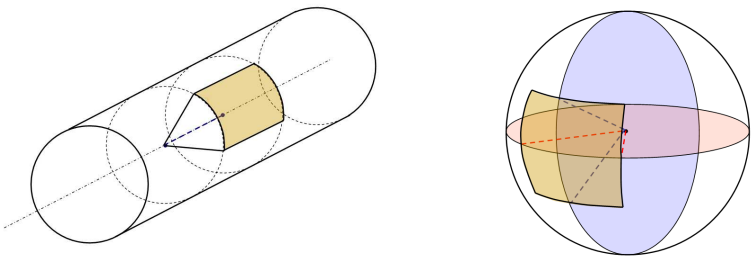

FIGURE 9.1: *Possible extensions of spherical regional residuals in case of a linear and a circular predictor variable (left panel) or two circular predictor variables (right panel).*

Euclidean distance measure, but of course other measures could be useful as well. The idea is to construct subsets of neighbouring points. Figure 9.1 shows a possible extension in case of a linear and a circular predictor variable (left panel) and in case of two circular predictor variables (right panel). Instead of rectangular areas, again spherical areas could be used as well. A frequently used circular distance measure for a single angular variate is to take the smaller of the two arcs between two angles ϕ_1 and ϕ_2, which can be expressed as

$$|\phi_1, \phi_2| = \min(|\phi_1 - \phi_2|, 2\pi - |\phi_1 - \phi_2|) = arc\cos[\cos(\phi_1 - \phi_2)].$$

An alternative, closely related definition of a circular distance is given by

$$|\phi_1, \phi_2| = 1 - \cos(\phi_1 - \phi_2).$$

For more details on circular distance measures, we refer the reader to the specific literature in this area, e.g. Jammalamadaka and SenGupta (2001), or Batschelet (1981).
These distance measures could then be combined with the Euclidean distance measure for predictor variables on the real line.

9.2.4 Smoothing based tests in circular-linear regression

Another solution to the LOF problem in circular-linear regression would be to adapt classical smoothing-based LOF tests by using a circular smoother (Giannitrapani, Bowman, and Scott (2005)). As smoothing based LOF tests have good power properties in linear-linear regression (Chapter 4 and Chapter 7), a good performance can also be expected here. Nevertheless, the major disadvantage remains the dependence on the choice of the smoother and the smoothing parameter.

9.2.5 Regional residual tests in generalized linear models

The simulation study in Section 7.1.5 indicates that the performance of the RR tests in generalized linear models depends on the type of residuals used to calculate the lack-of-fit test, but to the best of our knowledge no discussion in this context is available. A more extensive investigation is necessary to get better insight into the behaviour of these tests and to provide practical guidelines.

9.2.6 Limiting distributions for RR tests in multiple regression

The deduction of the asymptotic null distributions for RR tests in multiple regression is out of the scope of this thesis. Nevertheless, we believe that for spherical regional residuals, the asymptotic null distribution of the test statistic is again the supremum norm of a centered zero mean Gaussian process. The formal arguments are probably based on the higher-order results for empirical processes indexed by sets, as discussed e.g. in Chapter 26 in Shorack and Wellner (1986). Further investigations are certainly necessary.

Bibliography

Aerts, M., G. Claeskens, and J. Hart (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association 94*, 869–879.

Aerts, M., G. Claeskens, and J. Hart (2000). Testing lack of fit in multiple regression. *Biometrika 87, 2*, 405–424.

Alcalá, J., C. J.A., and W. González-Manteiga (1999). Goodness-of-fit test for linear models based on local polynomials. *Statistics & Probability Letters 42*, 39–46.

Azzalini, A. and A. Bowman (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society 55*(2), 549–557.

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

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society 139*(3), 318–355.

Barry, D. (1993). Testing for additivity of a regression function. *The Annals of Statistics 21*(1), 235–254.

Barry, D. and J. Hartigan (1990). An omnibus test for departures from constant mean. *Annals of Statistics 21*, 235–254.

Batschelet, E. (1981). *Circular Statistics in Biology*. London: Academic Press.

Bedrick, E. J. and J. R. Hill (1990). Outlier tests for logistic regression: A conditional approach. *Biometrika 77*(4), 815–827.

Boos, D. and J. Zhang (2000). Monte carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association 95*(450), 486–493.

Bowman, A. and S. Young (1996). Graphical comparison of nonparametric curves. *Applied Statistics 45*(1), 83–98.

Buckley, M. (1991). Detecting a smooth signal: Optimality of cusum based procedures. *Biometrika 78*(2), 253–62.

Chen, C.-F., J. D. Hart, and S. Wang (2001). Bootstrapping the order selection test. *Journal of Nonparametric Statistics 13*(6), 851–882.

Cheng, K. and J. Wu (1994). Testing goodness of fit for a parametric family of link functions. *Journal of the American Statistical Association 89*(426), 657–664.

Christensen, R. (1991). Lack-of-fit tests. *Journal of the American Statistical Association 86*(415), 752–756.

Cook, R. and S. Weisberg (1997). Graphics for assessing the adequacy of regression

models. *Journal of the American Statistical Association 92*(438), 490–499.

Copas, J. (1989). Unweighted sum of squares test for proportions. *Applied Statistics 38*(1), 71–80.

Davidson, R. and E. Flachaire (2001). The wild bootstrap, tamed at last. Working paper available on http://qed.econ.queensu.ca/pub/papers/abstracts/download/2001/1000.pdf.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

De Pypere, F. (2005). *Characterisation of Fluidised Bed Coating and Microcapsule Quality: A Generic Approach.* Ph. D. thesis, Ghent University, Ghent.

De Wiest, F. and H. Della Fiorentina (1975). Suggestions for a realistic definition of an air quality index relative to hydrocarbonaceous matter associated with airborne particles. *Atmospheric Environment 33*, 951–954.

DeBruyn, A. M. and J. J. Meeuwig (2001). Detecting lunar cycles in marine ecology: Periodic regression versus categorical ANOVA. *Marine Ecology Progress Series 214*, 307–310.

Dette, H. and A. Munk (1998). Testing heteroscedasticity in nonparametric regression. *J. R. Statist. Soc. 60*(4), 693–708.

Dette, H., A. Munk, and T. Wagner (1998). Estimating the variance in nonparametric regression - what is a reasonable choice? *Journal of the Royal Statistical Society 60*, 751–764.

Diebolt, J. (1995). A nonparametric test for the regression function: Asymptotic theory. *Journal of Statistical Planning and Inference 44*, 1–17.

Diebolt, J. and J. Zuber (1999). Goodness-of-fit tests for nonlinear heteroscedastic regression models. *Statistics & Probability Letters 42*, 53–60.

Draper, N. and H. Smith (1981). *Applied Regression Analysis* (Second ed.). USA: John Wiley & Sons.

Draper, N. R. and H. Smith (1998). *Applied Regression Analysis* (3rd ed.). Wiley series in probability and statistics. Texts and references section. New York: John Wiley & Sons.

Durbin, J. and G. Watson (1950). Testing for serial correlation in least squares regression. i. *Biometrika 37*(3/4), 409–428.

Eubank, R. and J. Hart (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics 20*(3), 1412–1425.

Eubank, R. and J. Hart (1993). Commonality of cusum, von neumann and smoothing-based goodness-of-fit tests. *Biometrika 80*(1), 89–98.

Eubank, R., J. Hart, D. Simpson, and L. Stefanski (1995). Testing for additivity in nonparametric regression. *The Annals of Statistics 23*(6), 1896–1920.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistiscal Modelling based on Generalized*

Linear Models. Springer Series in Statistics. New York, USA: Springer.

Fan, J. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association 91*(434), 674–688.

Fan, J. and L.-S. Huang (2001). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association 96*(454), 640–652.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika 34*(3/4), 320–334.

Fisher, N. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

Fisher, R. A. (1922). The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society 85*, 597–612.

Gasser, T., L. Sroka, and C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika 73*(3), 625–33.

Giannitrapani, M., A. W. Bowman, and M. E. Scott (2005). Additive models for correlated data with applications to air pollution monitoring. Technical report, The University Glasgow.

Gibbs, B., S. Kermasha, I. Alli, and C. Mulligan (1999). Encapsulation in the food industry: a review. *International Journal of Food Sciences and Nutrition 50*, 213–224.

Hall, P., J. Kay, and D. Titterington (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. Appl. Prob. 23*, 476–495.

Härdle, W. (1990). *Applied Nonparametric Regression*. New York, USA: Cambridge University Press.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*(4), 1926–1947.

Harper, W. (1967). The distribution of the mean half-square successive difference. *Biometrika 3/4*(54), 419–433.

Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.

Hosmer, D., T. Hosmer, S. le Cessie, and S. Lemeshow (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine 16*, 965–980.

Hosmer, D. and S. Lemeshow (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics A10*, 1043–1069.

Hosmer, D. and S. Lemeshow (2000). *Applied Logistic Regression* (2nd Edition ed.). Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Hosmer, D., S. Lemeshow, and J. Klar (1988). Goodness-of-fit testing for multiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal 30*(7), 1–14.

Hosmer, D. W. and N. L. Hjort (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine 21*, 2723–2738.

Hosmer, D. W., S. Taber, and S. Lemeshow (1991). The importance of assessing the fit of logistic regression models. *American Journal of Public Health 81*(12), 1630–1635.

Hurn, M., N. Barker, and T. Magath (1945). The determination of prothrombin time following the administration of dicumarol with specific reference to thromboplastin. *The Journal of laboratory and clinical medicine 30*, 432–447.

Jammalamadaka, S. R. and U. J. Lund (2005). The effect of wind direction on ozone levels - a case study. *to appear in Jour. Environmental and Ecological Statistics*.

Jammalamadaka, S. R. and A. SenGupta (2001). *Topics in Circular Statistics*. Singapore: World Scientific Press.

Joglekar, G., J. H. Schuenemeyer, and V. LaRiccia (1989). Lack of fit tests when replicates are not available. *The American Statistician 43*(3), 135–143.

Johnson, R. A. and T. E. Wehrly (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association 73*(363), 602–606.

Kuchibhatla, M. and J. Hart (1996). Smoothing-based lack-of-fit tests: Variations on a theme. *Nonparametric Statistics 7*, 1–22.

Kulasekera, K. and C. Gallagher (2002). Variance estimation in nonparametric multiple regression. *Communications in Statistics - Theory and Methods 31*(8), 1373–1383.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association 79*(385), 61–71.

Le, C. T., P. Liu, B. R. Lindgren, K. A. Daly, and G. S. Giebink (2003). Some statistical methods for investigating the date of birth as a disease indicator. *Statistics in medicine 22*, 2127–2135.

le Cessie, S. and H. C. van Houwelingen (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics 51*, 600–614.

le Cessie, S. and J. van Houwelingen (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics 47*, 1267–1282.

Ledwina, T. (1994, Sep.). Data-driven version of neyman's smooth test of fit. *Journal of the American Statistical Association 89*(427), 1000–1005.

Lee, G. and J. D. Hart (1998). An l_2 error test with order selection and thresholding. *Statistics & Probability Letters 39*, 61–72.

Lehmann, E. (1959). *Testing Statistical Hypotheses*. New York: John Wiley & Sons.

Lemeshow, S. and D. Hosmer (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology 115*, 92–106.

Li, C.-S. (2005). Using local linear kernel smoothers to test the lack of fit of nonlinear regression models. *Statistical Methodology 2*, 267–284.

Lin, D., L. Wei, and Z. Ying (2002). Model-checking techniques based on cumulative

residuals. *Biometrics 58*, 1–12.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics 21*, 255–285.

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

Montgomery, D. and E. Peck (1982). *Introduction to Linear Regression Analysis*. New York: John Wiley.

Moons, E., M. Aerts, and G. Wets (2004). A tree based lack-of-fit test for multiple logistic regression. *Statistics in Medicine 23*, 1425–1438.

Munk, A., N. Bissantz, T. Wagner, and G. Freitag (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society, Series B, Statistical Methodoloy 67*, 19–41.

Neill, J. W. and D. E. Johnson (1984). Testing for lack-of-fit in regression - a review. *Communications in Statistics - Theory & Methods 13*(4), 485–511.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models* (Fourth Edition ed.). Mc Graw Hill.

Opsomer, J. and M. Francisco-Fernández (2006). Finding local departures from a parametric model using nonparametric regression. *submitted to Statistical Papers*.

Pan, Z. and D. Y. Lin (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics 61*, 1000–1009.

Pardoe, I. (2001). A bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics 10*(4), 617–627.

Peixoto, J. (1990). A property of well-formulated polynomial regression models. *American Statistician 44*, 26–30.

Pregibon, D. (1981). Logisitic regression diagnostics. *The Annals of Statistics 9*(4), 705–724.

Pulkstenis, E. and T. J. Robinson (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine 21*, 79–93.

Rayner, J. and D. Best (1989). *Smooth Tests of Goodness-of-Fit*. New York, USA: Oxford University Press.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics 12*(4), 1215–1230.

Ryan, B., E. Wishart, and D. Shaw (1976). The growth rates and densities of ice crystals between $-3^{\circ}c$ and $-21^{\circ}c$. *Journal of the Atmospheric Sciences 33*, 842–850.

Shorack, G. R. and J. A. Wellner (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons, Inc.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–643.

Stute, W., W. González Manteiga, and M. Presedo Quindimil (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association 93*(441), 141–149.

Stute, W., S. Thies, and L.-X. Zhu (1998). Model checks for regression: An innovation process approach. *The Annals of Statistics 26*(5), 1916–1934.

Su, J. Q. and L. Wei (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association 86*(414), 420–426.

Su, Z. and S.-S. Yang (2006). A note on lack-of-fit tests for linear models without replication. *Journal of the American Statistical Association 101*(473), 205–210.

Tong, T. and Y. Wang (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika 92*(4), 821–830.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika 67*(1), 250–251.

Van Der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK.: Cambridge University Press.

Verloove, S. and R. Verwey (1988). Project on preterm and small-for-gestational age infants in the netherlands, 1983. University Microfilms International. No. 8807276 Ann Arbor, MI.

von Neumann, J. (1941). Distribution of the ratio of the mean square succesive difference to the variance. *The Annals of Mathematical Statistics 12*(4), 367–395.

von Neumann, J., R. Kent, H. Bellinson, and B. Hart (1941). The mean square successive difference. *Annals of Mathematical Statistics 12*, 153–162.

Watson, G. (1961). Goodness-of-fit tests on a circle. *Biometrika 48*, 109–114.

Williams, D. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics 36*(2), 181–191.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*(4), 1261–1295.

Yanagimoto, T. and M. Yanagimoto (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.

Zuber, J. (1996). Quelques tests non paramétriques d'absence d'effet en régression. Technical report, Ecole Polytechnique Fédérale de Lausanne.

Zuber, J. (1999). *Un Test Chi-Carré d'adéquation de modèles paramétriques en régression*. Phd thesis, École Polytechnique Fédérale de Lausanne.

Samenvatting

Een belangrijk onderdeel in elke regressie analyse is het controleren van de modelkwaliteit, onder meer door het vergelijken van de model gebaseerde voorspellingen met de observaties. In dit werk wordt een nieuw type van *lack-of-fit* (LOF) toetsen voorgesteld, in combinatie met bijhorende diagnostische grafieken. De toetsen zijn gebaseerd op zogenaamde *regionale residuen*, gedefineerd als het gemiddelde van klassieke residuen in deelgebieden van de ruimte van de onafhankelijke variabelen, de predictorruimte. Regionale residuen zijn zeer geschikte bouwstenen voor het construeren van een LOF toets. Indien er afwijkingen ten opzichte van het beschouwde model aanwezig zijn in een bepaald gebied in de predictorruimte, dan verwachten we in deze regio groepen van positieve of negatieve residuen. Deze zullen resulteren in grote absolute waarden van gestandardizeerde regionale residuen. Grote absolute waarden suggeren dus een mogelijke afwijking van het model, gesitueerd in het overeenkomstige deelgebied van de predictorruimte. De voorgestelde teststatistiek, de supremum norm van alle gestandardiseerde of ongestandardiseerde residuen, is een globale maat voor afwijkingen van het beschouwde model, en controleert voor een globaal significantie niveau α. Toetsen gebaseerd op regionale residuen, verder afgekort als RR toesten, zijn omnibus, in die zin dat ze consistent zijn tegen alle vaste alternatieve modellen. De RR toetsen zijn in het bijzonder gevoelig voor *lokale* afwijkingen van het model onder de nulhypothese, en zijn bovendien in staat om de afwijkingen te lokaliseren binnen de predictorruimte. Met "lokaal" wordt verwezen naar kleine gebieden in de predictorruimte waarvoor het werkelijke en het beschouwde nul model niet overeenkomen.
Belangrijke informatie gaat echter verloren door de afwijkingen in één globale maat samen te vatten. Daarom stellen we voor om de toets te gebruiken in combinatie van een grafische visualisatie van de individuele regionale residuen. Deze nieuwe grafieken identificeren op *formele* wijze gebieden waar het model geen goede voorspellingen oplevert.
Gladde toetsen uit de literatuur zijn over het algemeen krachtig in het detecteren van modelafwijkingen, maar hangen af van de al dan niet subjectieve keuze van een gladheidsparameter. Regionale residuen daarentegen worden over *alle* mogelijke deelgebieden berekend, zodat het onnodig is om een gladheidsparameter te kiezen. Op die manier is de RR toets in staat om

zowel globale als lokale afwijkingen van het model onder de nul hypothese te detecteren. Dit leidt wel tot een computationeel intensieve methode, hoewel enkele suggesties voor het reduceren van de computationele kost werden voorgesteld.
In tegenstelling tot verschillende klassieke LOF toetsen, zijn RR toetsen zowel toepasbaar in experimenten met al dan niet herhaalde waarnemingen. Indien er geen herhalingen voorhanden zijn is het regionaal residu, berekend over een deelgebied met één enkel punt uit de predictorruimte, gelijk aan het klassieke residu in dit punt. Wanneer echter herhaalde waarnemingen voorhanden zijn, dan is het regionale residu gelijk aan het gemiddelde van de klassieke residuen van de herhaalde waarnemingen op het ene punt.

De asymptotische nuldistributie van de RR toetsingsgrootheid is in het meest eenvoudige geval een gecentreerd Gaussiaans proces met gemiddelde nul en een ingewikkelde covariantiestructuur. Aangezien de convergentie traag is, is de asymptotische benadering niet geschikt voor kleine steekproefgroottes. We raden dan ook aan om de nuldistributie te schatten met behulp van een gepaste bootstrap methode.

De methode is in het bijzonder ook toepasbaar voor circulaire predictoren. Ondanks het feit dat de kleinste kwadraten methode kan gebruikt worden voor het schatten van een regressiemodel met een circulaire predictor, zijn de meeste klassieke LOF testen niet geschikt om systematische afwijkingen van dit model te detecteren. De p-waarden van de klassieke toetsen hangen immers vaak af van de keuze van de oorsprong van de circulaire variabele. Testen die gebaseerd zijn op regionale residuen, berekend over alle mogelijke bogen in de predictor, zijn echter oorsprong-invariant, en dus erg geschikt om *LOF op de cirkel* te detecteren. Wanneer er afwijkingen van het model met de RR toets gedetecteerd worden, worden de regio's met systematische afwijkingen van het model opnieuw op een formele manier in de predictorruimte gevisualiseerd.

De methode is eenvoudig uitbreidbaar naar situaties met meerdere voorspellingsvariabelen. De voorgestelde methode is gebaseerd op een afstandsmaat in de predictorruimte, waardoor er, in tegenstelling tot verschillende klassieke toetsen, geen specifieke ordening van de residuen hoeft gekozen te worden. Tenslotte worden in dit werk ook uitbreidingen naar de volledige klasse van veralgemeende lineaire modellen geïllustreerd en bediscussieerd.