Jaar 2005

Faculteit Wetenschappen

Non- and Semi-parametric Techniques for Handling Missing Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde,
te verdedigen door

Niel HENS

Promotor: Prof. dr. M. Aerts
Prof. dr. G. Molenberghs

If knowledge can create problems, it is not through ignorance that we can solve them.

Isaac Asimov (1920-1992)

Acknowledgment

Dit doctoraatswerk zou niet tot stand gekomen zijn zonder de medewerking en steun van verschillende personen.

Vooreerst wil ik mijn promotoren, Prof. dr. Marc Aerts en Prof. dr. Geert Molenberghs bedanken. Marc, bedankt voor de uitstekende begeleiding, het luisterend oor en de goede raad waarop ik steeds kon terugvallen. Geert, bedankt voor de fijne samenwerking en uitstekende bijdragen.

This work could not have been completed without the pleasant and fruitful collaboration I had with my co-authors, in alphabetical order: Frank Boelaert, Liesbeth Bruckers, Gerda Claeskens, Christel Faes, Ivy Jansen, Hans Laevens, Koen Mintiens, Ziv Shkedy, Niko Speybroeck, Herbert Thijs and Geert Verbeke.

Ook mijn andere collega's van het Centrum voor Statistiek wil ik hartelijk bedanken voor hun bijdrage, hoe klein of hoe groot deze ook was. Thank you all, for the fantastic atmosphere which you all create at the Center for Statistics.

Verder wil ik mijn vrienden bedanken die me de nodige ontspanning konden brengen. Een speciaal bedankje aan iedereen van SBBK en HBBC voor hun interesse en begrip op momenten dat het niet mogelijk was om hobby en beroep te combineren.

Een bijzonder woordje van dank voor mijn ouders en familie. Mama, papa zonder jullie onvoorwaardelijke steun en liefde was dit werk niet mogelijk geweest. Kelly, een laatste woordje voor jou. Jij, als geen ander, bent het zonnetje in huis met je liefde, vertrouwen, vriendschap en steun. Bedankt voor alles.

Niel Hens

Diepenbeek, 10 juni 2005

Contents

List of Abbreviations

AC	Available Cases
AIC	Akaike Information Criterion
AIC^{cor}	Corrected Akaike Information Criterion
AIC$_I$	Imputation-based Akaike Information Criterion
AIC$_S$	Smooth Akaike Information Criterion
AIC$_W$	Weighted Akaike Information Criterion
ANOVA	Analysis Of Variance
BoHV-1	Bovine Herpesvirus-1
BIC	Bayesian Information Criterion
BIC$_W$	Weighted Bayesian Information Criterion
CC	Complete Cases
CD	Case Deletion
CFWGEE	Constrained, Flexible, Weighted Generalized Estimating Equations
Cov	Covariance
Corr	Correlation
CP	Cost Complexity
Cp	Mallow's Cp
Cp$_W$	Weighted Mallow's Cp
CV	Cross-validation
EM	Expectation Maximization

FLIC	Functional Living Index: Cancer
FOI	Force Of Infection
GAM	Generalized Additive Model
GCV	Generalized Cross-validation
GEE	Generalized Estimating Equations
GI	Global Influence
HH	Household
HIS	Health Interview Survey
IPW	Inverse Probability Weighting
KL	Kullback-Leibler
KWGI	Kernel Weighted Global Influence
KWLI	Kernel Weighted Local Influence
LD	Likelihood Displacement
LI	Local Influence
LR	Local Resampling
LRT	Likelihood Ratio Test
LSR	Local Semi-parametric Resampling
MAR	Missing At Random
MASE	Mean Averaged Squared Error
MCAR	Missing Completely At Random
MI	Multiple Imputation
ML	Maximum Likelihood
MNAR	Missing Not At Random
MSE	Mean Squared Error
MT	Missing Together
NPSI	Non-parametric Single Imputation
NW	Nadaraya-Watson
OD	Original Data
OLS	Ordinary Least Squares

PMI	Parametric Multiple Imputation
PSI	Parametric Single Imputation
PSU	Primary Sampling Unit
RF	Random Forest
SE	Standard Error
SIR	Susceptible-Infected-Recovered
SSU	Secondary Sampling Unit
TSU	Tertiary Sampling Unit
UBRE	Unbiased Risk Estimation
var	Variance
WCC	Weighted Complete Cases
WEE	Weighted Estimating Equations
WGEE	Weighted Generalized Estimating Equations
xstd	Cross-validation Relative Error Standard Deviation
xerror	Cross-validation Relative Error

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Why is there a need for using Semi-parametric Techniques in Handling Missing Data?

Missing data arise in various settings, including surveys, clinical trials and epidemiological studies. With or without missing data, the goal of a statistical analysis is to make valid and efficient inferences about a population of interest. The issue of missing values complicates this process. Early on, modelling incomplete data relied on the use of parametric models (see e.g. Afifi and Elsahoff, 1966; Ibrahim, 1990). Recently there is a general trend towards non- and semi-parametric approaches to relax assumptions on which parametric models typically rely.

Two different reasonings towards the use of non- and semi-parametric modelling techniques exist. A first point of view is well described by Silverman (1985): "An initial non-parametric estimate may well suggest a suitable parametric model (such as linear regression), but nevertheless will give the data more of a chance to speak for themselves in choosing the model to be fitted."An alternative point of view arises from a statement by Box (1980): "Known facts (data) suggest a tentative model, implicit or explicit, which in turn suggests a particular examination and analysis of data and/or the need to acquire further data; analysis may then suggest a modified model that may require further practical illumination and so on."This results in an iterative procedure. Non-parametric techniques offer an ideal tool to obtain such a suitable parametric model (see Hastie and Tibshirani, 1987; Simonoff, 1996; Hart, 1997).

Non- and semi-parametric procedures in general will not be as efficient as model-based techniques when there is a posited model, and the model is appropriate.

However, if the assumed model is not the correct one, inferences can be worse than useless, leading to misleading interpretations of the data.

In this thesis, non- and semi-parametric techniques will be used to relax assumptions when analyzing incomplete data. We will first provide an overview of existing approaches for handling missing data.

1.2 Approaches to Missing Data

In real datasets, like, e.g., surveys and clinical trials, it is quite common to have observations with missing values for one or more input features. The first issue in dealing with the problem is determining whether the missing data mechanism has distorted the observed data.

Little and Rubin (1987) and Rubin (1987) distinguish between basically three missing data mechanisms. Data are said to be missing at random (MAR) if the mechanism resulting in its omission is independent of its (unobserved) value. If its omission is also independent of the observed values, than the missingness process is said to be missing completely at random (MCAR). In any other case the process is missing not at random (MNAR), i.e., the missingness process depends on the unobserved values. Some more detail of terminology will be provided in the following sections.

1.2.1 Standard Methodology

The literature presents various methods to handle missing data. They can roughly be classified into four groups.

Complete Case Analysis

When some variables are not observed for some of the units, one can omit these units from the analysis. These so-called “complete cases”are then analyzed as they are. This method is easy but can lead to serious biases and inefficiency (Little and Rubin, 1987).

Imputation-based Methods

Multiple imputation was formally introduced by Rubin (1978). Several other sources, such as Rubin and Schenker (1986), Little and Rubin (1987), Tanner and Wong (1987) and Schafer (1997)’s book give excellent and easy-to-read descriptions of the technique. The concept of multiple imputation refers to replacing each missing value with more than one imputed value. The goal is to combine the simplicity of imputation strategies, with unbiasedness in both point estimates and measures of precision. A problem of simple imputation procedures is that these may yield inconsistent point estimates as soon as the missingness mechanism surpasses MCAR. Another problem is that the variability of the estimators is underestimated, since imputed values are treated as observed values. By imputing several values for a single missing component, this uncertainty is explicitly acknowledged. There are several ways to impute missing values by multiple imputation. One of them is to draw from the posterior distribution based on the complete cases. Another flexible technique is to impute the missing values using classification trees as described by Hastie *et al.* (2001).

Weighting Methods

A third approach is based on the complete cases but now weighting them with the inverse of the probability that a case is observed as introduced by Flanders and Greenland (1991) and Zhao and Lipsitz (1992). In this way cases with a low probability to be observed gain more influence in the analysis and thus represent the probable missing values in the neighbourhood. One can look at this approach as an implicit imputation of missing values.

If this probability is unknown, which in general is the case, it can be estimated for instance using a non- or semi-parametric technique, e.g., kernel-based density estimation, splines or classification trees.

Recently Carpenter and Kenward (2005) compared the weighting and imputation procedure. They discuss the merits and demerits of both methods.

Fully Model-based Procedures

The number of publications on missing data modelling procedures increases exponentially and so does the use of model-based procedures. Such procedures rely on modelling the partially missing data using estimation methods such as, for example, maximum likelihood. They are based on untestable model assumptions and therefore sensitivity analyses are indispensable and should be part of the analysis. Recent literature (see e.g. Scharfstein *et al.*, 1999; Wang *et al.*, 2004) uses semi-parametric techniques to relax upon assumptions made, but still a sensitivity analysis ought to be conducted. Let us give a brief introduction to some popular parametric and semi-parametric model-based procedures for longitudinal data in the next section.

1.2.2 Modelling Longitudinal Data with Missing Values

In a longitudinal study, each unit is measured on several occasions. It is not unusual for some sequences of measurements to terminate early for reasons outside the control of the investigator, any unit so affected is often called a dropout. We will restrict our attention to longitudinal data with dropout and fully observed covariates. We refer to Roy and Lin (2002) for a discussion on modelling dropout when covariate values are missing. Denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$, $i = 1, \cdots, N$; the full data response vector. Define $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})$, a vector of missingness indicators for which the elements are given by $j = 1, \cdots, n_i$; $i = 1, \cdots, N$;

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Denote X to be the matrix of covariates.

Let us denote $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ where $\boldsymbol{Y}_i^o$ are those measurements Y_{ij} for which $R_{ij} = 1$ (observed) and $\boldsymbol{Y}_i^m$ those for which $R_{ij} = 0$ (missing). Let the data that would be measured when there were no missing data be referred to as the "Original Data". The "Full Data"is defined as the original data together with the missingness indicator, $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$. The "Available Cases"are those cases which are observed and the "Complete Cases"refer to the units for which none of the measurements are missing. Although of potential interest, auxiliary variables are not discussed here and therefore omitted from notation.

With this terminology a more refined distinction between the different missing data mechanisms can be made (Little and Rubin, 1987; Rubin, 1987). Missing values of $\boldsymbol{Y}$ are said to be *missing completely at random* (MCAR) if the probability of nonresponse does not depend on covariates, $\boldsymbol{X}$, nor on $\boldsymbol{Y}$. When the probability of nonresponse depends only on the covariates $\boldsymbol{X}$ and conditionally on $\boldsymbol{X}$ does not depend on $\boldsymbol{Y}$, the missingness mechanism is said to be *covariate-dependent MCAR.* If the dependency of the probability of nonresponse is allowed to depend only on the observed components of $\boldsymbol{Y}$ and possibly on $\boldsymbol{X}$, the missingness mechanism is *missing at random* (MAR), while if it is allowed to depend only on observed measurements, measured at the current or previous occasions is said to be *sequential MAR.* In this way the probability to be missing does not depend on missing data nor on future observed data. The latter could however be true for MAR data even if the missingness pattern is monotone, i.e., there are no intermittent missing values. Data are said to be "missing not at random"(MNAR) if the probability for the data to be missing depends on the unobserved data and possibly on the observed data.

Denote the distribution of missing data as $f_{\boldsymbol{\psi}}(\boldsymbol{R}|\boldsymbol{Y}, \boldsymbol{X})$. Further denote the measurement model $f_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X})$ and the joint model $f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{Y},\boldsymbol{R}|\boldsymbol{X})$. When dealing with MCAR or MAR data, likelihood based inference for the full data parameter $\boldsymbol{\theta}$ can be based on the likelihood of the observed data, assuming that the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are separable (meaning that their parameter spaces are non-overlapping). This is the ignorability condition (Little and Rubin, 1987). It follows from this assumption that

$$f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{Y}^o,\boldsymbol{R}|\boldsymbol{X}) = f_{\boldsymbol{\theta}}(\boldsymbol{Y}^o|\boldsymbol{X})f_{\boldsymbol{\psi}}(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y}^o). \tag{1.2}$$

In some situations the observed-data likelihood is difficult to use and methods like the EM-algorithm are required to optimize.

One can also use generalized estimating equations that use only observed response data in a semi-parametric way

$$\sum_{i=1}^{n} w_i \Psi(\boldsymbol{Y}_i;\boldsymbol{\theta}) = 0, \tag{1.3}$$

where w_i is the inverse of the marginal probability for an observation to drop out and Ψ the derivative of the log(quasi)likelihood. Under certain regularity conditions these equations provide a consistent estimate of $\boldsymbol{\theta}$. For more details we refer to Robins *et al.* (1994, 1995).

The semi-parametric approach relaxes upon the assumptions made but requires the proper estimation of the marginal probability for an observation to drop out which is not necessary for likelihood-based methods.

Turning to the situation of MNAR, likelihood based methods can be classified into three main frameworks: (1) selection models, (2) pattern mixture models and (3) shared parameter models. These three methods correspond to three different factorizations of the joint density

$$\begin{array}{ll}
(1) & f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{Y},\boldsymbol{R}|\boldsymbol{X}) = f_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X})f_{\boldsymbol{\psi}}(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{X}),\\
(2) & f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{Y},\boldsymbol{R}|\boldsymbol{X}) = f_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{R},\boldsymbol{X})f_{\boldsymbol{\psi}}(\boldsymbol{R}|\boldsymbol{X}),\\
(3) & f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\boldsymbol{Y},\boldsymbol{R}|\boldsymbol{X}) = \displaystyle\int f_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{\eta},\boldsymbol{X})f_{\boldsymbol{\psi}}(\boldsymbol{R}|\boldsymbol{\eta},\boldsymbol{X})dF(\boldsymbol{\eta}|\boldsymbol{X}),
\end{array}$$

where in the last expression $\boldsymbol{\eta}$ denotes a shared parameter.

Selection models consist of two parts. A measurement part and a missingness part. It has the appealing property that it expresses the taxonomy of Little and Rubin (1987) in a straightforward way by including or excluding different parameters in the dropout model. A pattern mixture model, uses a different model for each missing data pattern. In this way it focuses on the measurement model for a given missingness pattern and not on the global measurement model which is addressed by selection models. A shared parameter model or frailty model uses a random effect to induce dependency between the responses and the missing data process.

Let us express the different missingness processes when using the selection model factorization. First of all, Missing Completely At Random (MCAR):

$$f(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{\psi}) = f(\boldsymbol{R}|\boldsymbol{\psi}).$$

Secondly Missing At Random (MAR):

$$f(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{\psi}) = f(\boldsymbol{R}|\boldsymbol{Y}^o,\boldsymbol{\psi}).$$

Finally Missing Not At Random (MNAR) where the missingness process depends on the missing values:

$$f(\boldsymbol{R}|\boldsymbol{Y},\boldsymbol{\psi}) = f(\boldsymbol{R}|\boldsymbol{Y}^m,\boldsymbol{Y}^o,\boldsymbol{\psi}).$$

All of these methods are based on untestable assumptions and therefore a great deal of literature has focused on sensitivity analyses in this context.

Recently, attention has been devoted to semi-parametric methods to model MNAR dropout in Scharfstein *et al.* (1999), Rotnitzky *et al.* (1998), Fitzmaurice and Laird (2000) and Lin and Ying (2003). We refer to Hogan *et al.* (2004) and Molenberghs *et al.* (2004) for a further discussion on models handling dropout in longitudinal studies.

Let us now turn to an overview of some non- and semi-parametric techniques, which will be used and referred to throughout this thesis.

1.3 Non- and Semi-parametric Techniques

In a first part, we will describe some flexible modelling techniques, while in a second part, the bootstrap and jackknife are briefly introduced.

1.3.1 Modelling Techniques

In this section, non- and semi-parametric modelling techniques will be briefly introduced. While parametric techniques rely on several assumptions, non- and semi-parametric techniques relax on these assumptions. As with non-parametric procedures in general, non-parametric modelling techniques will not be as efficient as model-based techniques when the assumed is close to the true model.

In a parametric model the relationship, between a response variable and several explanatory variables, can be expressed in different ways, subject to different assumptions. Fractional polynomials were first introduced by Royston and Altman (1994). They provide flexibility while attaining the advantages of a parametric model. Other well known flexible parametric modelling techniques are power models as introduced by Cox and Hinkley (1974) and Davidian and Giltinan (1995). We will however restrict ourselves to fractional polynomials.

Fractional Polynomials

For a given degree m and a variable $x > 0$, fractional polynomials are defined as

$$\eta_m(x; \boldsymbol{\beta}, \mathbf{p}) = \sum_{i=0}^{m} \beta_i H_i(x), \tag{1.4}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ is the vector of regression parameters, $\mathbf{p} = (p_1, \ldots, p_m)$ a vector of powers $p_1 \leq \ldots \leq p_m$, which are positive or negative integers or fractions and $H_i(x)$ is a transformation given by

$$H_i(x) = \begin{cases} x^{p_i} & \text{if } p_i \neq p_{i-1} \\ H_{i-1}(x) \times \log x & \text{if } p_i = p_{i-1} \end{cases} \tag{1.5}$$

with $p_0 \equiv 0$ and $H_0 \equiv 1$. Royston and Altman (1994) argue that polynomials with degree higher than $m = 2$ are rarely required in practice. The powers themselves are proposed to be taken from $\{-2, -1, -0.5, 0, 0.5, 2, \ldots, \max(3, m)\}$ but other powers can be chosen too.

Non- and Semi-Parametric Modelling Techniques

Although fractional polynomials already offer a great deal of flexibility, parametric modelling is sometimes too confined. In this section some semi- and non-parametric techniques will be summarized. In what follows we will focus on smoothing techniques in a regression setting, since they will provide flexible tools to describe the underlying relationship between a response variable and several explanatory variables.

Smooth Regression

Suppose we have n observations $(x_1, y_1), \ldots, (x_n, y_n)$ and interest goes to $\boldsymbol{\mu} = (\mu(x_1), \ldots, \mu(x_n))$ with $\mu(x) = E(Y|x)$. In what follows we consider linear estimators, i.e., estimators of the form

$$\hat{\mu}_\lambda(x) = \sum_{i=1}^{n} w(x, x_i; \lambda) y_i, \tag{1.6}$$

where $w(x, x_i; \lambda)$ is a collection of weight functions that depend on one or more parameters λ. If, e.g., $w(x, x_i; \lambda) = n^{-1}, \quad i = 1, \ldots, n$, we obtain the sample mean of $\boldsymbol{y} = (y_1, \ldots, y_n)$.

A conceptually simple approach is to let the weight sequence $\{w(x, x_i; \lambda)\}_{i=1}^n$ be defined in the following way

$$w(x, x_i; \lambda) = \frac{K_\lambda(x_i - x)}{n^{-1}\sum_{i=1}^n K_\lambda(x_i - x)}, \tag{1.7}$$

where $K_\lambda(\cdot) = K(\cdot/\lambda)$ and K is a function satisfying $\int K(u)du = 1$. K is called a kernel function, while λ is often referred to as the bandwidth. Usually K is chosen to be a unimodal probability density function that is symmetric around zero like, e.g., the standard normal density or Epanechnikov kernel which enjoys some optimality properties (Härdle, 1990). The choice of the bandwidth λ corresponds with the window of values where the averaging of the y's in a neighbourhood of x is done over. Too small a window produces an undersmooth estimate, while too large a window will produce an oversmoothed estimate. Techniques to find an optimal bandwidth rely on minimizing the mean (integrated) squared error which in simple settings can be shown to depend on the second derivative of the unknown true regression function. Since the true regression function is not known, alternative techniques such as cross-validation were developed.

Cross-validation as described by Green and Silverman (1994) has the objective to provide an optimal bandwidth by minimizing the estimated mean integrated squared error

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^n \int [\hat{\mu}_{\lambda,[-i]}(x) - \hat{\mu}_\lambda(x)]^2 dx, \tag{1.8}$$

where $\hat{\mu}_{\lambda,[-i]}$ is the regression estimator (1.6) after deleting the i-th observation, $i = 1, \ldots, n$.

The estimators as defined by (1.7) are called *kernel smoothers*. The form (1.7) has been proposed by Watson (1964) and Nadaraya (1964) and is therefore known as the Nadaraya-Watson estimator.

A very popular approach in smooth regression is *local polynomial regression* which can be seen as an extension of kernel smoothing. In this local polynomial regression problem, the objective is to minimize

$$\sum_{i=1}^n \{Y_i - \beta_0 - \beta_1(x_i - x) \cdots - \beta_p(x_i - x)^p\}^2 K_\lambda(x_i - x), \tag{1.9}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. $\hat{\beta}_0$ estimates $\mu(x)$ while $\hat{\beta}_1, \ldots, \hat{\beta}_p$ estimate higher order derivatives of $\mu(x)$. An important part of local polynomial regression is to determine the appropriate power p and bandwidth λ. If $p = 0$, we obtain the Nadaraya-Watson estimate based on (1.7). We refer to Fan and Gijbels (1996) and Aerts and Claeskens (1997) for a more thorough discussion and a generalization towards multi-parameter models.

Recently *spline smoothing* has gained a lot of attention. Splines are generally defined as piecewise polynomials (Eubank, 1988) in which curve (or line) segments are constructed individually and then pieced together. There are different types of splines which can be roughly divided into smoothing splines, regression splines and penalized regression splines. Let us first give a basic definition of a cubic smoothing spline.

Consider

$$\sum_{i=1}^{n}(\mu(x_i) - y_i)^2 + \lambda \int [\mu''(x)]^2 dx, \tag{1.10}$$

where λ is a positive constant. For a given function μ, the term $\sum_{i=1}^{n}(\mu(x_i) - y_i)^2$ provides a measure of how well μ fits the data whereas $\lambda \int [\mu''(x)]^2 dx$ measures the smoothness of μ. The constant λ is the smoothing or penalty parameter that controls the trade-off between closely matching the data and having a smooth model. As $\lambda \to \infty$ the penalty increases and the smoother converges to an ordinary least squares (OLS) fitted cubic polynomial. When $\lambda \to 0$ the penalty decreases and the smoother converges to an OLS fitted regression spline through the data. The minimizer, $\hat{\mu}$ of (1.10) is a spline with all distinct values $x_1, \ldots, x_n$ as knots, which is a cubic polynomial on each interval $[x_{i-1}, x_i]$, $i = 2, \ldots, n$ and has two continuous derivatives.

In general, the placement of the knots and the determination of the penalty are very important for a spline. The difference between smoothing splines (#knots $= n$), regression splines (#knots $< n$) and penalized regression splines (regression splines with penalization for the number of knots) lies in the number of knots chosen. Basic references in this field are Eilers and Marx (1996), Ruppert and Carroll (2000) and Ruppert *et al.* (2003).

The choice of the smoothing parameter is crucial in the practical use of splines. While several methods as Akaike's information criterion (AIC), unbiased risk estimation (UBRE), and generalized maximum likelihood (see Wahba, 1990; Hurvich *et al.*, 1998) have been introduced, generalized cross-validation (GCV) is one method of smoothing parameter selection that has proven effective and has good theoretical properties.

Generalized cross-validation introduced by Craven and Wahba (1979) is based on minimizing

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}_{\lambda,[-i]}(y) - \hat{\mu}_{\lambda}(y)]^2 \left\{ \frac{1 - [S_\lambda]_{ii}}{n^{-1}\text{trace}(I - S_\lambda)} \right\}, \tag{1.11}$$

where S_λ is the smoothing matrix so that $\hat{\mu}_\lambda(y) = S_\lambda y$.

For multidimensional problems non-parametric smoothers face the problem of the 'curse of dimensionality'. In general, the "curse of dimensionality"(Bellman, 1961) refers to the exponential growth of hypervolume as a function of dimensionality. For non-parametric regression, this translates into sparseness of data which causes the variances of the estimates to be unacceptably large. Generalized additive models (GAM), as introduced by Hastie and Tibshirani (1987) can be used to cope with this problem. While $g(\mu_i) = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ji}$ represents a strict parametric generalized linear model (McCullagh and Nelder, 1989), where g is a known monotonic differentiable 'link function' and β_i are the parameters to be estimated, a generalized additive model has the form $g(\mu_i) = \beta_0 + \sum_{j=1}^{n} s_j(x_{ji})$, where the s_i are estimated using linear smoothers and backfitting.

While the work of Hastie and Tibshirani (1987) is considered to be the foundation of generalized additive models, Wood (2001, 2005) and Wood and Augustin (2002) have done a great deal of work on the application of the technique using penalized regression splines (Wahba, 1980, 1990; Marx and Eilers, 1998; Wood, 2000) instead of linear smoothers. Further methodology and theory was developed by, e.g., Aerts *et al.* (2002b). Model selection and inference when backfitting with linear smoothers (Hastie and Tibshirani, 1987) presents difficulties, while the mathematically elegant work of Wahba (1990) on generalized spline smoothing provides a rigorous framework for model selection and inference with generalized additive models. A 'middle way' between these approaches was the use of penalized regression splines to construct GAMs. The availability of the R package '*mgcv*' has made the use of GAMs very popular.

While GAMs with penalized regression splines offer a flexible modelling technique, it is still a hard task to apply them on data with many explanatory variables (like in a data-mining setting). Tree-based methods such as *classification and regression trees* (Breiman *et al.*, 1984) can be used for this purpose. Tree-based methods partition the covariate-space into rectangles and fit a simple model in each one of them. They are conceptually simple yet powerful. In what follows we restrict attention to classification trees.

Classification Trees

The classification tree methodology is a classification method where, following specific splitting rules, disjoint subsets of the data are constructed. These subsets are called nodes. Further splitting is repeated several times within these nodes. We focus on binary classification trees, where splitting occurs into exactly two child nodes.

This partitioning process results in a saturated tree. A tree is saturated in the sense that the offspring nodes subject to further division cannot be split. The saturated binary tree is then pruned to an optimal sized tree. This is the so-called pruning process. The final step is the selection process, which determines the final tree. The point is to find the subtree of the saturated tree that is most 'predictive' of the outcome and least vulnerable to noise in the data. Selection of the 'right-sized' tree is based on the cost complexity measure. This function is defined as the cost for the tree plus a complexity parameter times the tree size. In many typical applications, costs simply correspond to the proportion of misclassified observations, but other modifications are possible too (Zhang and Singer, 1999). V-fold cross-validation is useful when no test sample is available and the learning sample, i.e., the sample which was used to construct the tree, is too small to have the test sample taken from it. The classification tree of the specified size is computed V times, each time leaving out one of the subsamples from the computations, and using that subsample as a test sample for cross-validation. The CV costs computed for each of the V test samples are then averaged to give the V-fold estimate of the CV costs. While there is nothing wrong with choosing the tree with the minimum cost as the 'right-sized' tree, often there will be several trees with cross-validation (CV) costs close to the minimum. Breiman *et al.*(1984) make the reasonable suggestion that one should choose as the 'right-sized' tree the smallest-sized (least complex) tree whose costs do not differ appreciably from the minimum costs. They proposed a '1 SE rule' for making this selection, i.e., chose the 'right-sized' tree to be the smallest-sized tree whose costs do not exceed the minimum cost plus 1 times the standard error of the cost at the minimum. For more details we refer to Breiman *et al.* (1984), Zhang and Singer (1999) and Hastie *et al.* (2001).

The applications of tree-based methods in statistics nowadays are vast (e.g. Hens *et al.*, 2002; Speybroeck *et al.*, 2004). A general overview of the applicability was given by Segal (1995). Among more recent developments we find several areas of applications as, e.g., longitudinal data analysis and survival analysis (Segal, 1992; Shannon and Banks, 1999; Feldesman, 2002; Moons *et al.*, 2004).

One of the disadvantages of tree-based methods is that it gives a discrete, non-smooth model and that it is very unstable, which results in high variability. Bagging and boosting have been developed to deal with these issues. Boosting is a method which constructs a set of successive trees via iteratively reweighting. Observations which are misclassified, gain more weight and a new tree is built using these weights. The resulting classifier is a weighted average of the successive trees. The adaptive boosting of Freund and Shapire (1997) is very popular. Boosting works best with small trees.

Bagging or bootstrap aggregating (Breiman, 1996) was introduced to reduce the variance of a predictor. In a tree-based methodology, trees are grown on bootstrap samples of the learning data and then combined to obtain a more accurate prediction. Bagging works best with large samples. A theoretical foundation for bagging was offered by Bühlmann and Yu (2002).

A lot of variants of boosting and bagging have been developed. One of the most important ones is the *random forest methodology* (Breiman, 2001). For a random forest (RF), trees are grown on bootstrap samples of the learning data, as is done for bagging, but now for each bootstrap sample m out of p variables are randomly chosen at each node. These m variables are searched through for the best split and the largest tree possible is grown but not pruned. The default value for m is $\sqrt{p}$, but the technique appears to be relatively insensitive to m. The resulting trees are then combined to obtain a more accurate prediction (majority vote for binary classification trees).

The strength of a random forest is that it converges, unlike boosting, and so overfitting is not a problem. RFs are able to provide variable importance measures, which could be of interest in reducing the number of variables taken in a statistical analysis. RFs provide proximity measures between observations resulting in outlier detection and clustering.

RFs as well as single classification trees can be used to impute missing data in a straightforward way. They can be seen as hot-deck imputation methods.

Let us now turn to a brief introduction on the jackknife and the bootstrap.

1.3.2 The Jackknife and the Bootstrap

In practical situations interest often goes out to unknown parameters which have to be estimated from the empirical data at hand. Some disadvantages arise when estimating a parameter using traditional approaches. Among them are the need of a large sample size, a correctly postulated model and the ease to derive the theoretical formula or its approximation for each problem under consideration. Moreover one should be able to estimate accuracy measures. A first alternative is the jackknife, as originally described by Quenouille (1956), where the bias of an estimator was determined by deleting one observation each time from the original dataset and recalculating the estimator based on the rest of the data. The jackknife has become a more valuable tool since Tukey (1958) showed that the jackknife can also be used to construct variance estimators. A second alternative is the bootstrap which can be seen as an extension to the jackknife procedure. While the jackknife only utilizes n of $2^n - 1$ non empty subsets of a dataset of size n, the bootstrap can use more than n or even all $2^n - 1$ subsets to construct estimators (Efron, 1979).

There are two bootstrap situations one can consider, a parametric and a non-parametric bootstrap. The parametric bootstrap uses a particular mathematical model to regenerate data while the non-parametric bootstrap does not rely on such a model. Even if there is a plausible parametric model, a non-parametric analysis can still be useful to assess the robustness of conclusions drawn from a parametric analysis. For a more thorough discussion on bootstrap methods we refer to Davison and Hinkley (1997).

1.4 Objectives of this Thesis

In this thesis a variety of non- and semi-parametric techniques as introduced in Section 1.3 are used to handle missing data problems. The material presented here clearly shows the benefits of relaxing assumptions. Handling incomplete data problems by means of non- and semi-parametric techniques requires the availability of powerful computing resources. It will be clear that this still is a limitation to some of the methods presented here. Let us now give an overview of the different topics in this thesis.

Many authors as, e.g., Lipsitz *et al.* (1998), Rubin and Schenker (1986) and Heitjan and Little (1991) have addressed the use of semi- and non-parametric techniques relaxing on strong assumptions made by parametric techniques to impute missing values. Aerts *et al.* (2002a) propose to use local multiple imputation in a regression setting with nonresponse. Chapter 2 describes this kernel based imputation procedure which makes use of a non-parametric regression relationship between a partially observed response and fully observed covariate. The approach is related to the approximate Bayesian bootstrap method (see Efron, 1979; Little and Rubin, 1987) and can be seen as an extension of the local single imputation of Cheng (1994) to a proper local multiple imputation approach. An essential ingredient of the algorithm is the local generation of responses. Throughout the chapter, interest goes out to a marginal parameter of the response distribution.

In a regression analysis, selecting an appropriate model from a candidate set of models is based on, e.g., the Akaike Information Criterion (AIC). If however observations are incomplete, the use of complete cases can lead to wrong model choices. In Chapters 3 and 4, two modifications of the AIC-criterion are proposed. In Chapter 3, inverse probability weights are used, in analogy to the missing data method described in Section 1.2.1, to improve upon model selection. The method is applicable to both incomplete data and design-based samples (Hens *et al.*, 2005a). If the weights are unknown, they are estimated using generalized additive models with penalized regression splines as introduced in Section 1.3.1. Whenever only a few complete cases are available by deleting every observation with at least one missing value, weighting is not adequate anymore and imputation can provide a solution. Therefore, Chapter 4 focuses on an imputation-based AIC-criterion where imputation is non-parametric in nature by using generalized additive models with penalized regression splines. The simulations in the latter chapter reveal potential benefits of model selection after smoothing for fully observed regression data. In Chapter 5 we will illustrate these two modified AIC versions in a case study and contrast them with tree-based methods who deal with both missing values and design.

From the overview of existing material to deal with dropout in longitudinal studies in Section 1.2.2, it is clear that a sensitivity analysis should be part of any statistical analysis. Next to providing an overview of existing sensitivity tools, Chapter 6 describes a non-parametric sensitivity tool called 'kernel weighted influence' as derived by Hens *et al.* (2005b). It uses a 'kernel based neighbourhood' concept to explore the global and local influence towards non-random missingness for types of observations instead of observations itself in a selection model framework. These sensitivity tools pick up a lot of different anomalies in the data not only deviations from the MAR-assumption. A method to oppose missing at random versus missing not at random in a selection model framework is the likelihood ratio test. In Chapter 7, the bootstrap will be used in an attempt to generate the null distribution of the likelihood ratio test statistic opposing missing not at random versus missing at random in a selection model framework.

In Chapter 8, generalized estimating equations are used to determine the force of infection for binary clustered data. The impact of missing data on the analysis is illustrated and an inverse probability weighted estimating equation is proposed. The weights are estimated non-parametrically by a generalized additive model with penalized regression splines. Several other complications in the dataset are dealt with, including the constraint for the age-specific seroprevalence to be monotone increasing. Deriving confidence intervals under these constraints is done using the bootstrap. The application of these techniques in the context of veterinary epidemiology is new and therefore considered to be a motivation for interdisciplinary collaboration between statisticians and veterinary epidemiologists working in this field.

1.5 Key Examples

In this section, the datasets which will be used throughout this work are introduced. The Vorozole data (Section 1.5.1) and the Mastitis data (Section 1.5.1) are examples of longitudinal studies with dropout. The Cervix Cancer Screening data (Section 1.5.2) have been collected as a part of the Belgian health survey held in 1997 and has a considerable amount of incomplete measurements. The Bovine Herpes Virus-1 Data (Section 1.5.4) relate to the field of veterinary epidemiology and include features such as clustering and missingness.

1.5.1 Vorozole Data

This study was an open-label, multicenter, parallel group design conducted at 67 North American centers. Patients were randomized to either vorozole (225 patients, 2.5 mg taken once daily) or megestrole acetate (227 patients, 40 mg four times daily). The patient population consisted of postmenopausal patients with histologically confirmed estrogen-receptor positive metastatic breast carcinoma. All 452 randomized patients were followed until disease progression or death. The main objective was to compare the treatment group with respect to response rate while secondary objectives included a comparison relative to duration of response, time to progression, survival, safety, pain relief, performance status and quality of life. Full details of the study are reported in Goss *et al.* (1999). Here we focus on overall quality of life, measured by the total Functional Living Index: Cancer (FLIC, Schipper *et al.*, 1984). Precisely, a higher FLIC score is the more desirable outcome.

Patients underwent screening and for those deemed eligible a detailed examination at baseline took place. Further measurements were taken at month 1, then from month 2 at bi-monthly intervals until month 44. The average total FLIC score was 116.3 (s.e. 1.3) for the vorozole group, and 117.1 (s.e. 1.3) for megestrole acetate group. These total FLIC scores were calculated based on 199 resp. 213 patients. Goss *et al.* (1999) analyzed the FLIC score using a two-way ANOVA model with effects for treatment, disease status, as well as their interaction and found no significant difference.

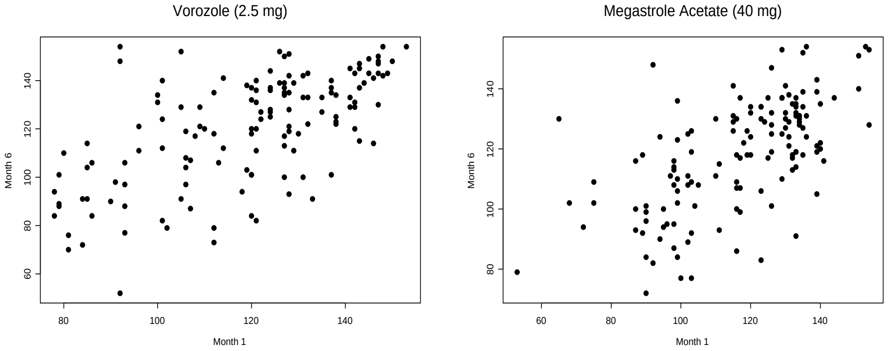


Figure 1.1: Vorozole data: Scatterplot of the FLIC scores at month 6 versus month 1 for the two treatment groups separately.

In Figure 1.1, focus is on the FLIC scores at month 6 versus FLIC scores at month 1 for the two treatment groups separately when both measurements are available. The mean FLIC score at month 6 for the vorozole group was 123.32 (s.e. 3.69), and for the megestrole acetate group 119.15 (s.e. 4.19). In the megestrole acetate group 102 patients (48%) have dropped out at month 6 while in the vorozole group 112 patients (53%) have dropped out at month 6 resulting in a considerable amount of missing data.

In Chapter 2, the mean FLIC score for patients at month 6 for both treatment arms separately is estimated based on local multiple imputation, which uses a regression relation between the FLIC scores at month 6 with those at month 1 to impute data.

1.5.2 Cervix Cancer Screening Data

According to the *Nationaal Kankerregister* (Haelterman, 1999), cervix cancer is the fifth most common cancer among women in Belgium in the period of 1993-1995. Therefore, it is not surprising that for health policy goals cervix cancer is an important point of attention. Interest goes to differences between the group of women, aged 25-64, not having a smear and those that did have a smear taken in the past three years. Data were available as a subset of the first Belgian Health Interview Survey (HIS) which took place in 1997. An brief introduction on the HIS is given here.

Table 1.1: The Cervix Cancer Screening data: explanatory variables.

Variable	Measurements	Variable	Measurements
Civil Status	Married	Educational Level	No Diploma
	Divorced		Primary Education
	Widow(er)		Lower Secondary Education
	Single		Higher Secondary Education
Age	'0-14'		Higher Education
	'15-24'	Financial Status	Difficult to Pay Health Costs
	'25-44'		No Problems to Pay Health Costs
	'45-64'	Drug Consumption	Number of Drugs Taken
	'65+'		

In the HIS, a total sample of 10,000 interviews (0.1% of the Belgian population) was planned, equally spread over the year 1997. For the three regions of Belgium (Flemish region, Walloon region and Brussels region) the number of individuals to be successfully interviewed was preset at 3500, 3500 and 3000, respectively. An oversampling was planned for the German Community of Belgium (in the district Eupen-Malmédy), with 300 successful interviews. A detailed description of the sampling scheme used in the HIS was published elsewhere (Quataert *et al.*, 1998). The most important features are summarized in what follows. Sampling was based on a combination of stratification, multistage sampling, and clustering (Kish, 1995).

There were two stratification levels. First, stratification was done at the regional level, to ensure that the preset regional level could be reached. Secondly, stratification was conducted at the level of provinces, proportional to their size. Next, the individuals' sample is selected in three stages within each stratum. The first stage, yielding primary sampling units (PSU), consists of municipalities and sampling is carried out proportionally to (population) size via systematic sampling. Whenever a municipality is selected (and it can be more than once), a group of 50 persons is to be interviewed within this municipality. The next stage of random selection operates on households (HHs, secondary sampling units or SSU) according to a clustered systematic sampling procedure upon ordering of the HHs by statistical sector, size and age of the reference person. At this level, matching HHs are provided in case a HH refuses to participate. Finally, individuals or tertiary sampling units (TSU) are selected within HHs in such a way that 4 persons at most are interviewed in each HH and the reference person and his/her partner are automatically selected.

To investigate the cervix cancer screening, only women aged between 25 and 64 were selected from the HIS. The explanatory variables of interest for these 2893 women are shown in Table 1.1[1] (Women without uterus are excluded from the analysis).

While the response variable was fully observed, 56% of the women had one or missing values for the explanatory variables. While missing data issues in the health interview survey have been addressed by Burzykowski *et al.* (1999), Quataert *et al.* (1998) focused on the design of the survey. In Chapter 5 the data are analyzed while dealing with design issues and missing data. A weighted and imputation-based AIC-criterion, as introduced in Chapter 3 and 4, will be used to select an appropriate logistic regression model, while a classification tree is used to provide a non-parametric alternative to logistic regression accounting for the design and missing values.

1.5.3 Mastitis Data

In this dataset the occurrence of the infectious disease of the udder, called mastitis, in dairy cows was studied. The milk yields in thousands of liters of 107 cows from a single herd in two consecutive years were available. In the first year all cows were supposedly free of mastitis and in the second year 27 cows became infected. Mastitis typically leads to a reduction in milk yield. There is a view among dairy scientists, widely held, that mastitis is more likely to occur in high yielding cows. It is however difficult to examine such a relationship due to the effects of mastitis. Figure 1.2 shows a profile plot of the Mastitis data.

In Figure 1.3, a scatterplot of the original data is given together with a plot of the increments, i.e., the difference between the second and the first measurement against the first measurement.

The Mastitis data have been analyzed by Kenward (1998) for an informal sensitivity analysis and further by Molenberghs *et al.* (2001) with the local influence methodology while using the selection model proposed by Diggle and Kenward (1994). In Chapter 6, weighted influence measures are applied to these data as a part of a sensitivity analysis.

[1] The questionnaire can be consulted at www.iph.fgov.be/epidemio/epien/crospen/hisen/table.htm.

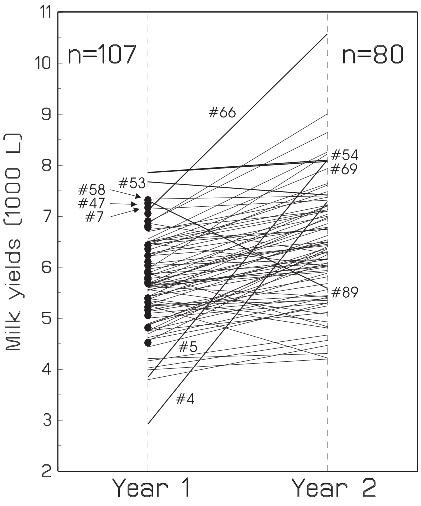


Figure 1.2: Mastitis data: Profile plot.

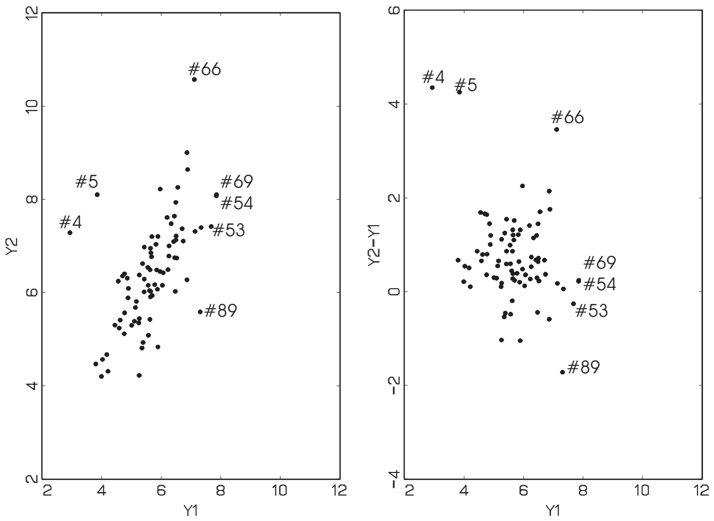


Figure 1.3: Mastitis data: Scatter plot. In the left panel the milk yield for year 2 was plotted versus the milk yield at year 1. In the right panel the increase in milk yield from year 1 to year 2 was plotted versus the milk yield at year 1.

1.5.4 Bovine Herpes Virus-1 Data

The bovine herpesvirus 1 (BoHV-1) is a transmissible disease in cattle, which is of economic importance and significance to international trade. It is spread worldwide. To facilitate the free trade of cattle, several European countries implemented eradication programs for BoHV-1. BoHV-1 causes infectious bovine rhinotracheitis, an enzootic disease. The BoHV-1 seroprevalence (apparent prevalence) in the Belgian cattle population was determined by a large serological survey, conducted from December 1997 to March 1998 (Boelaert *et al.*, 2000; Speybroeck *et al.*, 2003). The sample taken was stratified for province. Within each province, 1% of the total number of herds was sampled. The blood samples, which were taken from all animals in the selected herds, were tested for antibodies against BoHV-1 by using an ELISA-test, specific for the BoHV-1 glycoprotein B (gB). Additional characteristics as gender, type of the herd (dairy, mixed or beef), purchased or homebred and size of the herd were recorded. In total 11284 animals were investigated. In Table 1.2, a complete overview of the variables is given.

Table 1.2: Overview of the different variables in the BoHV-1 dataset.

Variable	Description
gB	ELISA-test positive for glycoprotein B, or not
herd	number of the herd
animal	number of the animal
province	province (nine, Brabant Walloon and Flemish Brabant together)
herdtype	dairy, mixed or beef
herd size	size of the herd
densanim	density of animals in the municipalities (number of cattle/km^2)
densherd	density of herds in the municipalities (number of herds/km^2)
age	age of the animal (in months)
sex	gender of the animal
purchase	purchased or homebred

From these variables 'age', 'sex' and 'gB' had a small amount of missing values, 0.23%, 0.12%, 0.32%, respectively. The 'purchase' variable, indicating whether an animal was homebred or purchased had 2091 missing values (19.00 %).

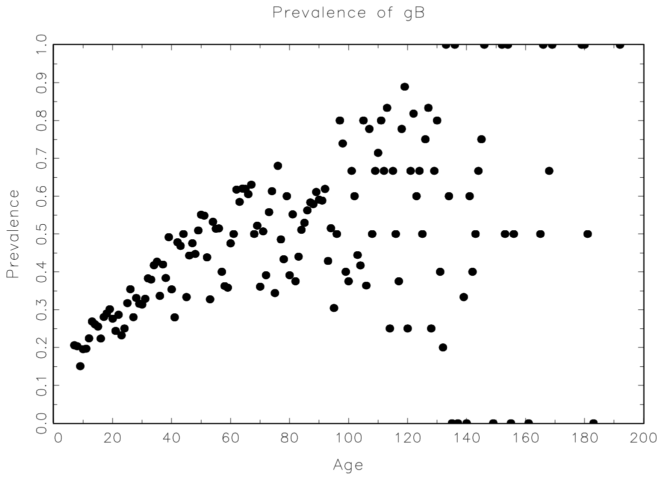


Figure 1.4: BoHV-1 study. Seroprevalence plot as a function of age (in months).

It is often of interest to look at the seroprevalence as a function of age. Since animals younger than 6 months typically have high seroprevalence of gB-antibodies because of acquired maternal antibodies and not necessarily due to infection with the BoHV-1, we will restrict ourselves to the animals older than 6 months. In Figure 1.4, the age-specific prevalence of gB-antibodies is displayed. There is a clear increase of seroprevalence with age. In Chapter 8, we will present a flexible population-averaged model to relate the seroprevalence with the recorded variables and derive the force of infection thereof.

Chapter 2

Local Multiple Imputation

2.1 Introduction

Datasets with missing values arise frequently in statistical practice. Population surveys inevitably face the problem of incomplete data, missing data create difficulties in quality of life studies, in cancer clinical trials, etc. There exist many ways to deal with missing data problems, ranging from the most naive one focusing on the complete cases only to well-defined parametric, semi-parametric and non-parametric approaches.

In this chapter we will introduce non-parametric — smoothing — methods to obtain multiple imputation estimators in a non-Bayesian framework. The proposed approach is novel in several aspects. Unlike most of the literature which deals with missing covariate values, this method allows for missing response data.

The onset to the use of kernel methods for imputation of missing values is given by Titterington and Sedransk (1989), who used kernel density estimation in combination with a non-parametric bootstrap for imputing values. In their method, relationships between variables are not directly accounted for. For single imputation in a nonresponse setting, Cheng (1994) and Chu and Cheng (1995) used kernel estimators in a regression model. To overcome the curse of dimensionality when using high dimensional smoothing operations, Wang *et al.* (2004) and Little and An (2004) propose the use of propensity scores to obtain semi-parametric imputation methods.

For missing covariate data, smoothing methods have been applied by Wang *et al.* (1998) to estimate selection probabilities. Other semi-parametric approaches, in the sense of not having to specify a fully parametric model, although not directly in a

Table 2.1: Example: Cholesterol levels for heart-attack patients measured 2, 4 and 14 days after attack. Source: Ryan and Joiner (1994)

Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
270	218	156	236	234	–	210	214	242	142	116	–
280	200	–	272	276	256	160	146	142	220	182	216
226	238	248	242	288	–	186	190	168	266	236	236
206	244	–	318	258	200	294	240	264	282	294	–
234	220	264	224	200	–	276	220	188	282	186	182
360	352	294	310	202	214	280	218	–	278	248	198
288	278	–	288	248	256	244	270	280	236	242	204

smoothing context, are constructed for drop-out models in Scharfstein *et al.* (1999).

Many authors, e.g., Lipsitz *et al.* (1998), Rubin and Schenker (1986), Heitjan and Little (1991), have addressed the use of semi- and non-parametric techniques to impute missing values in a wide variety of other settings.

We will make use of a non-parametric regression relationship between a partially observed response variable and a fully observed covariate to augment the data.

As an illustrative example, consider data on the serum-cholesterol levels of heart-attack patients in Table 2.1 (data from Ryan and Joiner 1994, analyzed by Schafer 1997). For all patients, treated for heart attacks, cholesterol levels were measured at 2 time points after the attack (Y_1 and Y_2). For only a part of the patients, an additional measurement Y_3 was taken at a third occasion. Schafer (1997) demonstrates the use of multiple imputation to estimate the parameters of greatest interest which in this case appear to be functions of the means, such as comparisons or contrasts among μ_1, μ_2 and μ_3. Examples include μ_3, the mean cholesterol level at the last occasion, $\mu_1 - \mu_3$, the average decrease in cholesterol level and $100(\mu_1 - \mu_3)/\mu_1$, the percentage decrease. In parametric multiple imputation, the original data are regarded as a random sample from a trivariate normal distribution, an assumption which implies that the marginal distribution of $Y_i, i = 1, 2, 3$ is normal and that the conditional distribution of, e.g., Y_3 given Y_1 is also normal with a conditional mean function which is linear in Y_1. Two kinds of assumptions can be distinguished: distributional assumptions and assumptions concerning parameter functions. The need for methods relaxing both types of assumptions has been recognized by many authors.

Our approach is based on local imputation methods. Whereas multiple imputation is mainly regarded as a Bayesian technique, the proposed methods are essentially bootstrap based (see also Efron, 1994). In the next section, we introduce two local bootstrap methods, the *local resampling* method which is fully non-parametric and hence relaxes both types of the above mentioned assumptions, and the *local semi-parametric resampling* method which still assumes that the conditional distributions are, e.g., locally normal but which allows non-linear conditional mean structures.

We focus on settings such as in the above example where some of the variables are fully observed and some involve missing measurements. The parameter of interest is essentially a marginal parameter of an incompletely observed variable. The regression relationship with a completely observed variable is exploited to impute values for the missing items. Throughout, we assume an ignorable nonresponse mechanism.

In the next section we introduce some basic notation and explain the imputation algorithm. Section 2.3 focuses on the local bootstrap method, asymptotic results are presented in Section 2.4 and the selection of the different smoothing parameters involved in the procedure, is considered in Section 2.5. Section 2.6 summarizes the results of a simulation study. An application to the Vorozole data and a discussion are provided in Sections 2.7 and 2.8, respectively.

2.2 Local Imputation Scheme

Consider one completely observed continuous variable X and one incompletely observed continuous variable Y. The parameter of interest is a function $\theta(\mu_X, \mu_Y)$ of the two means. Since there is no missing X value, the problem reduces to consistent estimation of $\mu = \mu_Y$; estimators of other moments of Y and functions thereof can be obtained in a straightforward manner. Extensions to more general settings and other parameters are discussed in Section 2.8.

The main idea is to exploit the assumed regression relationship between X and Y to yield better estimators for μ. Let $Z_i = (X_i, Y_i, \delta_i), i = 1, \dots, n$, be independent observations, where $\delta_i = 0$ if Y_i is missing and $\delta_i = 1$ otherwise. Under the strongly ignorable missing at random assumption (Rosenbaum and Rubin, 1983)

$$\pi(X) := E(\delta|X) = E(\delta|X, Y). \tag{2.1}$$

In other words, Y and δ are conditionally independent given X. Note that this assumption is weaker than missingness completely at random since dependence on the observed variable X is allowed. Little and Rubin (1987, p. 15), term data of this type missing at random but not observed at random.

Most of the parametric imputation schemes assume an underlying parametric function, such that $Y_i = \mu(X_i) + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n$ are independent mean zero random variables and $\mu(\cdot)$ is known up to a finite dimensional parameter, see e.g. Schenker and Welsh (1988).

In case this model assumption is not correct, biased results will be obtained. As an example, consider a linear single imputation model in X, where the true $\mu(\cdot)$ is non-linear. A linear imputation model calculates $\hat{Y}_j$ from a linear regression model based on the complete data vectors (X, Y). By straightforward computation one finds that the expected value of the usual estimator for the global mean of Y, which is constructed by taking the average of those Y-values which are observed, and the estimated values $\hat{Y}_j$, if Y_j is missing, differs from $E\{\mu(X)\}$ because $E(\hat{Y}_j|X)$ differs from $\mu(X)$. That is, if our assumed regression model is not correct, there will be a bias.

One of the key ideas in this chapter is to use non-parametric regression techniques to impute the missing values, this to avoid making such model assumptions. As in parametric methods, it is implicitly assumed that there is a statistical relationship between X and Y. Cheng (1994) uses a non-parametric kernel estimator to impute single missing Y observations. Such a single imputation can be considered as a non-parametric version of the so-called ‘poor man’s data augmentation’, which is known to underestimate variability, especially in cases with substantial missingness. Little and Rubin (1987) call such a method an *improper* imputation method.

Our approach extends this local single imputation of Cheng (1994) to a non- or semi-parametric version of a ‘proper’ imputation method and is related to the approximate Bayesian bootstrap method as described in equation (3.7) of Efron (1994); see also Little and Rubin (1987, Section 12.4). An essential ingredient of the algorithm is the local generation of Y observations. Let x be a specific value of X at which a Y value is to be generated and let $w_j(x), j = 1, \ldots, n$, denote positive weights with $\sum_{j=1}^n w_j(x) = 1$. The local resampling method generates a Y value from the distribution $\mathcal{L}(x)$ with cumulative distribution function

$$\sum_{j=1}^{n} w_j(x) I\{Y_j \leq y\}. \tag{2.2}$$

Detailed treatment of the choice of weights is postponed to Section 2.3. First we describe the steps of the local m-fold multiple imputation algorithm, where as an example, attention is restricted to a normal likelihood in Step 2.

Step 1: Resampling step

Fix ℓ between 1 and m. For each observation $i = 1, \ldots, n$, if $\delta_i = 1$, generate $Y_i^*(\ell)$

from the distribution $\mathcal{L}(X_i)$. This is a non-parametric resampling of the observed data vectors.

Step 2: Imputation step

Fix ℓ between 1 and m. Given the data from Step 1, we create imputations for the missing Y values. This can be done in several ways, using local resampling or local semi-parametric resampling. More explicitly, conditional on the resampled data $(X_i, Y_i^*(\ell), \delta_i), i = 1, \ldots, n$, we construct a distribution $\mathcal{L}_\ell^*(X_i)$, for local resampling, or local estimators $\hat{\mu}_\ell^*(X_i)$, $\hat{\sigma}_\ell^{*2}(X_i)$, for local semi-parametric resampling. If Y_i is missing, that is if $\delta_i = 0$, we generate $Y_i^+(\ell)$ from $\mathcal{L}_\ell^*(X_i)$, for local resampling, or, for local semi-parametric resampling, we generate $Y_i^+(\ell)$ from $N\{\hat{\mu}_\ell^*(X_i), \hat{\sigma}_\ell^{*2}(X_i)\}$. It is clear that local semi-parametric resampling is more efficient if normality holds. In both Step 1 and Step 2, data are generated independently for $i = 1, \ldots, n$, $\ell = 1, \ldots, m$.

Step 3: Construction of the final estimators

For $\tilde{Y}_i(\ell) = \delta_i Y_i + (1-\delta_i) Y_i^+(\ell)$, $\hat{\mu}(\ell) = n^{-1} \sum_{i=1}^n \tilde{Y}_i(\ell)$ is the estimator of the mean based on the ℓ-th augmented dataset, and the final multiple-imputation estimator for μ is

$$\hat{\mu} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\mu}(\ell). \tag{2.3}$$

The algorithm has the same structure as its parametric counterpart. Since an imputed observation Y^+ is subject to extra variability, Step 1 is needed for obtaining a proper imputation method (Efron, 1994). This extra randomness can be introduced in different ways. The triplets $(X_i, Y_i, \delta_i), i = 1, \ldots, n$, could be resampled with replacement; this is case resampling. We opted for an alternative approach, where Y values are generated, conditional on X and δ, incorporating the regression relationship between X and Y in a non-parametric way. This approach is legitimate because of assumption (2.1), which states that the missingness mechanism is noninformative for the parameter μ of interest. The advantage of the method is that it generates samples with exactly the same range of X values as in the original sample, avoiding samples which might only poorly reflect the regression structure, the latter which is essential in Step 2.

The non-parametric imputation method is applicable in a wide variety of statistical models, and can be used for discrete response data. The small adaptation needed for semi-parametric resampling is the specification of the appropriate distribution function in Step 2 of the algorithm. For examples of local likelihood estimators in multi-parameter families, see Aerts and Claeskens (1997).

A local bootstrap method both avoids parametric assumptions and allows much more flexibility in the regression design than does for example hot-deck imputation (Rao and Shao, 1992), where one requires a covariate to take on only a few different values, with replication.

2.3 Local Bootstrap Methods

The choice of the weights in the resampling scheme is crucial. Global uniform weights $\delta_j / \sum_j \delta_j$ would simply result in mean imputation of the Y-values, ignoring the regression structure completely. More useful are kernel weights of the type

$$w_j(x) = \frac{\delta_j K_h\left(x - X_j\right)}{\sum_{k=1}^{n} \delta_k K_h\left(x - X_k\right)}, \tag{2.4}$$

where the kernel $K(\cdot)$ is a symmetric unimodal probability density function, $K_h(u) = K(u/h)/h$, and $h = h_n$ is a bandwidth parameter converging to zero as the sample size increases. It is not necessary to use the same set of weights in the resampling and imputation steps. In particular, since the smoothing weights in Step 2 use a resampled set of data, it is advisable to use a second bandwidth $g = g_n$ in a possibly different kernel L for the construction of the weights in the imputation step. In case of possible confusion, choice of bandwidth will be included in the notation.

Local weights (2.4) are defined such that observed Y_j values, of which the corresponding X_j is closer to the specific value x, and which are in an area with larger chance of having missing observations, get larger weights. The latter is readily understood by rewriting the weights (2.4) as $w_j(x) = \delta_j \tilde{w}_j(x)/\hat{\pi}(x)$ where the classical Nadaraya-Watson weights $\tilde{w}_j(x) = K_h\left(x - X_j\right) / \sum_{k=1}^{n} K_h\left(x - X_k\right)$ and where $\hat{\pi}(x) = \sum_{j=1}^{n} K_h\left(x - X_j\right) \delta_j / \sum_{j=1}^{n} K_h\left(x - X_j\right)$ is the kernel estimator for $\pi(x)$. Thus we do not have to make any parametric assumptions about the missingness probability distribution since this is automatically taken care of by the kernel weights. The effect of $\hat{\pi}(x)$ on the weights stresses the importance of the few available but highly informative Y observations in a ‘sparse’ area with a lot of missingness.

In the complete data case, Aerts *et al.* (1994) have shown that distribution (2.2) is consistent and asymptotically normal for estimating the conditional distribution of Y given $X = x$. They also showed that a resampling scheme based on this distribution leads to a consistent bootstrap procedure. In an analogous way it can be shown that, if Y values are missing, (2.2) is a consistent estimator for the distribution function $P(Y \leq y | X = x, \delta = 1)$. Its mean equals the well-known Nadaraya-Watson estimator at x from the complete cases (Nadaraya, 1964; Watson, 1964), which can be rewritten as

$$\hat{\mu}(x) = \sum_{j=1}^{n} \tilde{w}_j(x)\ \delta_j Y_j / \hat{\pi}(x), \tag{2.5}$$

where the numerator is the kernel estimator of $E(\delta Y | X = x)$ and the denominator estimates $E(\delta | X = x)$ non-parametrically. It immediately follows from assumption (2.1), that $\hat{\mu}(x)$ is an estimator of $E(Y|X = x)$. The variance of (2.2),

$$\hat{\sigma}^2(x) = [\sum_{j=1}^{n} \delta_j \{Y_j - \hat{\mu}(x)\}^2 K_h(x - X_j)] / \sum_{j=1}^{n} \delta_j K_h(x - X_j),$$

is a consistent non-parametric variance estimator. These provide alternatives to the local likelihood estimators in local semi-parametric resampling.

Given the known limitations of Nadaraya-Watson weights, alternative sets of local weights are worth considering, such as biased bootstrap weights (Hall and Presnell, 1999), constrained to make the adjusted estimator unbiased for linear functions. Here we define

$$\begin{aligned} \breve{w}_j(x) \quad &= \quad \delta_j K_h(x - X_j) \left\{1 + c(x - X_j) K_h(x - X_j)\right\}^{-1} \\ &\quad \times \left[\sum_{k=1}^{n} \delta_k K_h(x - X_k) \left\{1 + c(x - X_k) \times K_h(x - X_k)\right\}^{-1} \right]^{-1}, \end{aligned} \tag{2.6}$$

where c is the solution to the equation

$$\sum_{j=1}^{n} \delta_j (x - X_j) K_h(x - X_j) \left\{1 + c(x - X_j) K_h(x - X_j)\right\}^{-1} = 0. \tag{2.7}$$

These weights are asymptotically equivalent to local linear weights. Hence they automatically correct for boundary bias, while remaining positive. Alternative or additional constraints on the resampling distribution can be imposed in a similar way.

If the proportion of missingness would be known, missing data could be dealt with as in a weighted-distributions regression setting, for which Ahmad (1995), see also Jones (1991), derives a kernel estimator analogue to the direct sampling case. The corresponding weights, with π estimated by the kernel estimator $\hat{\pi}$, are defined as

$$\breve{w}_j(x) = \delta_j \tilde{w}_j(x) \{\hat{\pi}(X_j) \sum_{k=1}^{n} \delta_k \tilde{w}_k(x) / \hat{\pi}(X_k)\}^{-1}. \tag{2.8}$$

The important difference from the weights (2.4) is the evaluation of $\hat{\pi}$ at the covariates X_j.

The performance of the above weights will be numerically illustrated in Section 2.6.1, where it turns out that the precise choice of weights has little effect on the final estimator.

If more than one variable is completely observed, local methods could take all of them into account. However, in high dimensions kernel-based methods might lose some of their attractiveness because of the curse of dimensionality.

2.4 Asymptotic Expressions of Bias and Variance

The final estimator $\hat{\mu}$ is consistent, under conditions similar to those in Cheng (1994). Smoothness conditions require $\mu(x)$, the conditional mean of Y given $X = x$, the density function, $f_X(x)$, and the function $\pi(x)$, to possess at least two bounded derivatives, bandwidth sequences to tend to zero at a rate faster than $n^{-1/3}$, and kernel functions K and L in both steps to be bounded and symmetric probability density functions with finite second moments. Although the proofs of the following theorems are provided for the local resampling algorithm; the proofs for the local semi-parametric resampling are very similar. We also assume that Y has a finite second moment, and that all required expected values are finite.

A first result shows that the final estimator $\hat{\mu}$ is asymptotically unbiased and that the bias depends on both bandwidth sequences in a typical non-parametric way.

Property 1. *For some constants c_1 and c_2,*

$$E(\hat{\mu}) = \mu + c_1 h^2 + c_2 g^2 + o(h^2 + g^2) \ , \qquad \text{as } n \to \infty. \tag{2.9}$$

The asymptotic variance of $\hat{\mu}$, as $n \to \infty$, with additional constants $c_3, \ldots, c_6$ and with $\sigma^2(X) = \mathrm{var}(Y|X)$, is given by

$$\begin{aligned}\mathrm{var}(\hat{\mu}) \quad = \quad & (mn)^{-1} E[\sigma^2(X)\{1 - \pi(X)\}/\pi(X)] + n^{-1}[E\{\sigma^2(X)/\pi(X)\} + \mathrm{var}\{\mu(X)\}] \\ & + n^{-2}(c_3 h^{-1} + c_4 g^{-1}) + n^{-1}(c_5 h^2 + c_6 g^2) + o\{(h^2 + g^2)n^{-1}\},\end{aligned} \tag{2.10}$$

showing that $\hat{\mu}$ is root-n consistent as an estimator for μ.

Proof. Let us denote by $E(\cdot|O)$ the expectation conditional on $Z_1, \ldots, Z_n$ and by $E(\cdot|O, R)$ the expectation conditional on $Z_1, \ldots, Z_n, Z_1^*, \ldots, Z_n^*$, where $Z_i^* = (X_i, Y_i^*, \delta_i)$.

Suppose that the weights on the observed data, with g a twice continuously differentiable function fulfil $\sum_{j=1}^n E\{w_j(X; h) g(X_j)\} \to E\{g(X)\} + O(h^2)$. This condition holds for the weights studied in Section 2.3. Since $\hat{\mu}$ is defined as

$$\hat{\mu} = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_j(\ell),$$

where $\tilde{Y}_j(\ell) = \delta_j Y_j + (1-\delta_j)Y_j^+(\ell)$, the first term contributes to

$$E(\delta_j Y_j) = E\{E(\delta_j Y_j|X_j)\} = E\{\pi(X)\mu(X)\}.$$

Next, we look at the expectation of the second term,

$$E\{(1-\delta_j)Y_j^+(\ell)\} = E[(1-\delta_j)E\{Y_j^+(\ell)|O,R\}].$$

By definition of $Y_j^+(\ell)$,

$$E\{(1-\delta_j)Y_j^+(\ell)\} = E[(1-\delta_j)E\{Y_j^+(\ell)|O,R\}] = E\{(1-\delta_j)\hat{\mu}_\ell^*(X_j;g)\}.$$

Using the explicit formula $\hat{\mu}_\ell^*(X_j;g) = \sum_{i=1}^n w_i(X_j;g)Y_i^*(\ell)$, conditioning on the observed data and using a Taylor expansion, hereby making use of the symmetry of the kernel functions K and L, we obtain that

$$E\{(1-\delta_j)\tilde{Y}_j(\ell)\} = E\{\mu(X)(1-\pi(X))\} + O(h^2+g^2).$$

Together with the result for the first term, $E(\delta_j Y_j)$, this concludes the first part of the proof.

Conditioning on observed and first-stage resampled data, we obtain $\text{var}(\hat{\mu}) = E\{\text{var}(\hat{\mu}|O,R)\} + \text{var}\{E(\hat{\mu}|O,R)\}$. By definition of the multiple imputation estimator $\hat{\mu}$, we have

$$\begin{aligned} E\{\text{var}(\hat{\mu}|O,R)\} &= \frac{1}{mn}E[(1-\delta_1)\,\text{var}\{Y_1^+(1)|O,R\}] \\ &= \frac{1}{mn}E[(1-\delta_1)E\{\hat{\sigma}_1^{*2}(X_1;g)|O\}], \qquad (2.11)\end{aligned}$$

where

$$\hat{\sigma}_1^{*2}(X_1;g) = \sum_{j=1}^{n}\{Y_j^*(1) - \hat{\mu}_1^*(X_1;g)\}^2 w_j(X_1;g).$$

The inner expectation in (2.11), which is conditional on the observed data, is most easily obtained by explicitly rewriting $\{Y_j^* - \hat{\mu}^*(X_i;g)\}^2$ as $(Y_j^*)^2 - 2Y_j^*\hat{\mu}^*(X_i;g) + \{\hat{\mu}^*(X_i;g)\}^2$, and by calculating the conditional expectation of each term separately, using computations similar to those in the proof of (2.9). Proceeding this way, we obtain that

$$E\{\text{var}(\hat{\mu}|O,R)\} = \frac{1}{mn}E[\{1-\pi(X)\}\sigma^2(X)] + O\{(h^2+g^2)n^{-1}\}.$$

Next, we turn to

$$\begin{aligned}\operatorname{var}\{E(\hat{\mu}|O,R)\} &= \operatorname{var}[E\{E(\hat{\mu}|O,R)|O\}] + E[\operatorname{var}\{E(\hat{\mu}|O,R)|O\}] \\ &= \tfrac{1}{n}\operatorname{var}[\delta_1 Y_1 + (1-\delta_1)E\{\hat{\mu}_1^*(X_1;g)|O\}] \\ &\qquad + \tfrac{1}{mn}E[(1-\delta_1)\operatorname{var}\{\hat{\mu}_1^*(X_1;g)|O\}]. \qquad (2.12)\end{aligned}$$

Similar calculations as before yield, for the first term in (2.12),

$$\tfrac{1}{n}[E\{\sigma^2(X)/\pi(X)\} + \operatorname{var}\{\mu(X)\} + O\{(h^{-1}+g^{-1})n^{-1}\} + O\{(h^2+g^2)n^{-1}\}],$$

and, for the second term,

$$\tfrac{1}{mn}\left(E[\{1-\pi(X)\}^2\sigma^2(X)/\pi(X)] + O(h^2+g^2)\right),$$

from which the result follows. □

The second term on the right-hand side of $\operatorname{var}(\hat{\mu})$ in equation 2.10 represents the variance of a single mean imputation, as shown in Cheng (1994). The first term stems from the multiple imputation approach with additional randomness generated in Step 1. In the case of no missingness, the leading term in (2.10) reduces to $\operatorname{var}(Y)/n$, as expected. The constants c_i depend on the second derivatives of $\mu(x)$, $f_X(x)$ and $\pi(x)$, with respect to x, as well as on second moments of the kernel functions.

The following central limit result holds

Theorem 1.

$$\sqrt{n}\operatorname{var}(\hat{\mu})^{-1/2}\{\hat{\mu} - E(\hat{\mu})\} \to N(0,1), \qquad (2.13)$$

in distribution, with mean and variance as given by (2.9) and (2.10).

Proof. Define

$$V_{1n} = \tfrac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\hat{\mu}_1^*(X_i;g) - E\{\mu(X)\} + \tfrac{1}{n}\sum_{i=1}^{n}\delta_i Y_i,$$

$$V_{2n} = E(V_{1n}|O).$$

Conditional on $Z_1^*,\ldots,Z_n^*$, $\sqrt{n}(\hat{\mu} - \mu - V_{1n}) \to N_1$, in distribution and, conditional on $Z_1,\ldots,Z_n$, $\sqrt{n}(V_{1n} - V_{2n}) \to N_2$, in distribution. Unconditionally, $\sqrt{n}V_{2n} \to N_3$, in distribution, where the N_i have a normal distribution. Since V_{2n} features only observed data, normality is readily obtained. Distributions of N_1 and

N_2 can be obtained by separating the randomness induced by the bootstrap resampling as in the following triangular arrays:

$$\begin{aligned}\sqrt{n}(V_{1n}-V_{2n}) &= \sum_{j=1}^{n}\Bigg[n^{-1/2}(1-\delta_j)\sum_{i=1}^{n}w_i(X_j;g)\sum_{\ell=1}^{n}\Big\{Y_\ell-\sum_{k=1}^{n}w_k(X_i;h)Y_k\Big\}\\ &\qquad\times I\{w_{(\ell-1)}(X_j;h)<U_{1j}\le w_{(\ell)}(X_j;h)\}\Bigg],\\ \sqrt{n}(\hat{\mu}-\mu-V_{1n}) &= \sum_{i=1}^{n}\Bigg[n^{-1/2}(1-\delta_i)\sum_{\ell=1}^{n}\{Y_\ell^*-\sum_{k=1}^{n}w_k(X_i;g)Y_k^*\}\\ &\qquad\times I\{w_{(\ell-1)}(X_i;g)<U_{2\ell}\le w_{(\ell)}(X_i;g)\}\Bigg],\end{aligned}$$

where $w_{(k)}(X_i;h)=\sum_{j=1}^{k}w_j(X_i;h)$ and $w_{(0)}=0$. The independent random variables U_{1j} (respectively U_{2j}) follow a uniform distribution on (0, 1), and are independent of $Z_1,\dots,Z_n$ (respectively $Z_1^*,\dots,Z_n^*$). A central limit theorem result for the triangular arrays above is obtained via Theorem 2.1 of Janssen and Mikosch (1997).

Application twice of Lemma 1 of Schenker and Welsh (1988) yields the desired result that $\sqrt{n}(\hat{\mu}-\mu)-N\to 0$ in distribution, where N is as the convolution of the three distributions above, namely a normal random variable with mean and variance as already calculated in (2.9) and (2.10). □

The latter results lead to the following asymptotic expression for the mean squared error of $\hat{\mu}$,

$$\begin{aligned}\mathrm{MSE}(\hat{\mu}) &= c_0n^{-1}+(c_1h^2+c_2g^2)^2+(c_3h^{-1}+c_4g^{-1})n^{-2}\\ &\quad+(c_5h^2+c_6g^2)n^{-1}+o\{(h^2+g^2)n^{-1}\},\end{aligned}\qquad(2.14)$$

where $c_0=m^{-1}E[\sigma^2(X)\{1-\pi(X)\}/\pi(X)]+E\{\sigma^2(X)/\pi(X)\}+\mathrm{var}\{\mu(X)\}$.

In the remainder of this section we examine the behaviour of a particular estimator of $\mathrm{var}(\hat{\mu})$ by showing how it relates to expression (2.10).

In parametric multiple imputation estimation, the variance of $\hat{\mu}$ is typically estimated by $S^2(\hat{\mu})=\hat{W}+(1+m^{-1})\hat{B}$, where $\hat{W}$ is the average within-imputation variance estimator, i.e., $\hat{W}=m^{-1}\sum_{\ell=1}^{m}S_\ell^2$, where nS_ℓ^2 is the unbiased sample variance within the ℓ-th augmented dataset, and $\hat{B}$ is the between-imputation variance, i.e., $\hat{B}=(m-1)^{-1}\sum_{\ell=1}^{m}\{\hat{\mu}(\ell)-m^{-1}\sum_{k=1}^{m}\hat{\mu}(k)\}^2$. The following proposition, which proves the asymptotic unbiasedness of $S^2(\hat{\mu})$ as an estimator of $\mathrm{var}(\hat{\mu})$, holds.

Property 2.

$$E(\hat{W}) = \frac{1}{n}\big[\mathrm{var}\{\mu(X)\}+E\{\sigma^2(X)\}\big]+O\{(h^2+g^2)n^{-1}\}\qquad(2.15)$$

$$E(\hat{B}) = \frac{1}{n}E\left\{\frac{1-\pi(X)}{\pi(X)}\sigma^2(X)\right\}+O\{(h^2+g^2)n^{-1}\}.\qquad(2.16)$$

Proof. By straightforward calculation we get that, with $\mu_2(X) = E(Y^2|X)$,

$$E(\hat{W}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} E\left[\{\delta_i Y_i + (1-\delta_i)Y_i^+(1)\}^2 - \frac{1}{n}\left\{ \sum_{i=1}^{n} \delta_i Y_i + (1-\delta_i)Y_i^+(1) \right\}^2 \right]$$

$$\begin{aligned} &= \tfrac{1}{n}E\{\mu_2(X)\} - \tfrac{1}{n}\left[E\{\pi(X)\mu(X)\}\right]^2 - \tfrac{2}{n}E\{\pi(X)\mu(X)\} \cdot E[\{1-\pi(X)\}\mu(X)] \\ &\quad -\tfrac{1}{n}(E[\{1-\pi(X)\}\mu(X)])^2 + O\{(h^2+g^2)n^{-1}\} \\ &= \tfrac{1}{n}E\{\mu_2(X)\} - \tfrac{1}{n}[E\{\mu(X)\}]^2 + O\{(h^2+g^2)n^{-1}\} \\ &= \tfrac{1}{n}[\text{var}\{\mu(X)\} + E\{\sigma^2(X)\}] + O\{(h^2+g^2)n^{-1}\}, \end{aligned}$$

which is result (2.15).

For $\ell = 1, \ldots, m$, define the random variables

$$D_n(\ell) = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)Y_i^+(\ell).$$

Being a sample variance, the estimator $\hat{B}$ is an unbiased estimator of the variance of $D_n(1)$, conditional on the observed data. Hence,

$$E(\hat{B}) = E[\text{var}\{D_n(1)|O,R\}] + E(\text{var}[E\{D_n(1)|O,R\}|O]). \tag{2.17}$$

By definition of $D_n(1)$ and $Y_i^+(1)$,

$$\text{var}\{D_n(1)|O,R\} = \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\,\text{var}\{Y_i^+(1)|O,R\} = \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\hat{\sigma}_1^{*2}(X_i;g).$$

Since this depends on both the observed and the first-stage resampled data, we calculate the first term of (2.17) via $E(E[\text{var}\{D_n(1)|O,R\}|O])$. As in the proof of (2.10), the expectation of the resulting random variable is given by

$$\tfrac{1}{n}E[\{1-\pi(X)\}\sigma^2(X)] + O\{(h^2+g^2)n^{-1}\}.$$

The second term in (2.17) can be shown to equal

$$\tfrac{1}{n}E[\{1-\pi(X)\}^2\sigma^2(X)/\pi(X)] + O\{(h^2+g^2)n^{-1}\},$$

from which (2.16) follows. □

The construction of the variance estimator $S^2(\hat{\mu})$ is simple and is exactly the same as in parametric multiple imputation methods. This is an advantage over other estimators of this variance, such as the non-parametric estimator of Cheng (1994), where an additional smoothing parameter needs to be selected.

2.5 Optimal Bandwidths

As with any other non-parametric method, also here we cannot go without the choice of smoothing parameters. Asymptotic optimal bandwidths are obtained in Section 2.5.1, followed by jackknife bandwidth selection in Section 2.5.2.

2.5.1 Asymptotically Optimal Bandwidths

The asymptotically optimal bandwidths minimize the dominant terms in (2.14). Terms of order $O(h/n)$ are negligible compared to order $O\{(nh)^{-2}\}$ terms, as long as $h = O(n^{-\alpha})$, with $\alpha > 1/3$. The same holds for g, where the order of g is not restricted to be the same as the order of h.

By differentiating (2.14) and omitting all negligible terms, we find that both bandwidths are $O(n^{-2/5})$, yet with different constants, depending on c_1, c_2, c_4 and c_5.

Since the constants in front of the $n^{-2/5}$ are functions of higher derivatives of $\mu(x)$, these cannot be computed exactly for any dataset. Data-driven bandwidth selection is to be advised, although in practice any 'reasonable' bandwidth choice will give satisfactory results.

Interesting to observe is that the order of the optimal bandwidths is not the $O(n^{-1/5})$ typically obtained in non-parametric regression estimation. It compares to the rates obtained in non-parametric density estimation. Although we use the regression relationship between the random variables X and Y, the non-parametric kernel weights are mainly used to construct a probability distribution from which 'new' response values are to be generated.

2.5.2 Jackknife Bandwidth Selection

Using the asymptotic optimal order derived in the previous section, jackknife ideas can be utilized to estimate the MSE of $\hat{\mu}$ for different choices of the bandwidths h and g. A data driven selection of both smoothing parameters can then be based on the minimization of the MSE. Because of the double resampling estimation procedure, there are different ways to implement a sensible jackknife method.

As a first possibility one could simply apply Quenouille's original jackknife method (Quenouille, 1956) to each of the augmented datasets, $\tilde{Y}_i(\ell), i = 1, \ldots, n$; $\ell = 1, \ldots, m$. This method is not successful in our case because it cannot estimate the bias, which here results from the non-parametric estimation used to obtain the variables $\tilde{Y}_i(\ell)$. Moreover some simulations showed that it results in unreasonable variance estimates.

A second possibility is to implement the jackknife at the start of the estimation procedure, on the original data. This approach requires some modifications of the classical jackknife. Indeed, as can be seen from expression (2.9), the bias is not of a "parametric order" $1/n$ but determined by the leading term $c_1h^2+c_2g^2$. With $\alpha > 0$ some exponent to be specified later, define the following jackknife pseudo-values

$$\hat{\mu}_{n,i} = \frac{n^\alpha \hat{\mu} - (n-1)^\alpha \hat{\mu}^{(-i)}}{n^\alpha - (n-1)^\alpha} \tag{2.18}$$

where $\hat{\mu}^{(-i)}$ is defined exactly as $\hat{\mu}$ but based on all but the i-th observation (X_i, Y_i, δ_i). Using (2.9), the average of the pseudo-values

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{n,i} \tag{2.19}$$

has approximately the following expectation

$$E(\bar{\mu}) \approx \mu + c_1 \frac{n^\alpha h_n^2 - (n-1)^\alpha h_{n-1}^2}{n^\alpha - (n-1)^\alpha} + c_2 \frac{n^\alpha g_n^2 - (n-1)^\alpha g_{n-1}^2}{n^\alpha - (n-1)^\alpha}. \tag{2.20}$$

Taking $h = C_h n^{-2/5}$ and $g = C_g n^{-2/5}$, expression (2.20) can be rewritten as

$$E(\bar{\mu}) \approx \mu + c_1 C_h \frac{n^{\alpha-4/5} - (n-1)^{\alpha-4/5}}{n^\alpha - (n-1)^\alpha} + c_2 C_g \frac{n^{\alpha-4/5} - (n-1)^{\alpha-4/5}}{n^\alpha - (n-1)^\alpha}. \tag{2.21}$$

By the choice $\alpha = 4/5$, the leading bias term of $\hat{\mu}$ cancels out, leading to a bias-corrected estimator $\bar{\mu}$ and called the generalized jackknife statistic (Gray and Schucany, 1972). Since the bias of $\hat{\mu}$ is not of the order $1/n$, the choice $\alpha = 1$ corresponding to Quenouille's original jackknife pseudovalues is not appropriate here.

The difference $\widehat{\text{bias}}(\hat{\mu}) = \hat{\mu} - \bar{\mu}$ is known as the jackknife bias estimator and the jackknife variance estimator for $\hat{\mu}$ is given by (see e.g. Efron and Tibshirani, 1993)

$$\widehat{\text{var}}(\hat{\mu}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\mu}_{n,i} - \bar{\mu})^2.$$

Both $\widehat{\text{bias}}(\hat{\mu})$ and $\widehat{\text{var}}(\hat{\mu})$ depend on the values of the unknown constants C_h and C_g. Optimal choices can then be derived by minimizing the estimated mean squared error

$$\widehat{\text{mse}}(\hat{\mu})(C_h, C_g) = \widehat{\text{bias}}^2(\hat{\mu})(C_h, C_g) + \widehat{\text{var}}(\hat{\mu})(C_h, C_g).$$

This method has been implemented but led to highly variable estimates for bias and variance. The reason for this failure is the generation of new responses in step 1 and step 2. Within each jackknife run, deleting the i-th observation, complete new data are generated, causing far too high variability in pseudo-values.

A third procedure retains all data generated in steps 1 and 2, but modifies, for each $i = 1, \ldots, n$, the imputed data $Y_j^+(\ell)$ to $Y_j^{+(-i)}(\ell)$ by shifting them to a new mean reflecting the deletion of the i-th observation while using $h_{n-1} = C_h(n-1)^{-2/5}$ and $g_{n-1} = C_g(n-1)^{-2/5}$. This idea was inspired by the adjusted jackknife as proposed by Rao and Shao (1992). When using all data $(X_1, Y_1, \delta_1), \ldots, (X_n, Y_n, \delta_n)$, the conditional mean of $Y_i^+(\ell)$ is given by

$$\hat{\mu}_n^+(X_i; K, L, h_n, g_n) = \frac{\displaystyle\sum_{k=1}^{n} \delta_k L\left(\frac{X_i - X_k}{g_n}\right) \hat{\mu}(X_k; K, h_n)}{\displaystyle\sum_{k=1}^{n} \delta_k L\left(\frac{X_i - X_k}{g_n}\right)} \tag{2.22}$$

where $\hat{\mu}(X_k; K, h_n)$ is defined by

$$\hat{\mu}(x; K, h) = \frac{\sum_{j=1}^{n} \delta_j K\left(\frac{x - X_j}{h}\right) Y_j}{\sum_{j=1}^{n} \delta_j K\left(\frac{x - X_j}{h}\right)}. \tag{2.23}$$

Within each jackknife run i (referring to deletion of the i-th observation, $i = 1, \ldots, n$), the imputed observation $Y_j^+(\ell)$ is replaced by the adjusted imputed value

$$Y_j^{+(-i)}(\ell) = Y_j^+(\ell) + \{\hat{\mu}_{n-1}^{+(-i)}(X_j; K, L, h_{n-1}, g_{n-1}) - \hat{\mu}_n^+(X_j; K, L, h_n, g_n)\}, \tag{2.24}$$

where $\hat{\mu}_{n-1}^{+(-i)}(X_j; K, L, h_{n-1}, g_{n-1})$ is defined as $\hat{\mu}_n^+(X_i; K, L, h_n, g_n)$ in (2.22) but with the i-th observation excluded and using bandwidths h_{n-1} and g_{n-1}.

The values $\hat{\mu}_{(-i)}$ are defined in the same way as before but now without the i-th observation and based on the jackknife imputed values. The pseudo-values are again defined as in (2.18). Using similar arguments as those in (2.19) to (2.21), it can be shown that the leading bias term disappears for the estimator $\bar{\mu}$.

As illustrated in the next section, this jackknife procedure succeeds in selecting a proper choice of C_h and C_g. An in-depth study of the theoretical properties and the finite sample behaviour of this jackknife bandwidth selector is beyond the scope of this chapter.

2.6 Simulation Results

In this section we apply the above developed methods to simulated data, and perform a comparison with other approaches dealing with this type of missingness.

2.6.1 A Simulation Study

The following methods for multiple imputation are included in this simulation study. The first, naive approach uses the complete cases only. Among the parametric methods we consider single imputation (Buck, 1960) and multiple imputation, according to Rubin (1978, 1987) and Efron (1994). These methods all assume a parametric regression relationship between Y and X. Rubin's multiple imputation assumes joint normality of (X, Y). In Efron's bootstrap approach, the complete cases are resampled and used to fit a linear regression model of Y on X in order to impute Y-values from a normal distribution with estimated linear conditional mean function and estimated constant variance.

Three non-parametric approaches are also included. The first is a single imputation method, in which a local linear estimator of the conditional mean is used to impute for missing Y values (Cheng, 1994). The other two methods are those studied in this chapter, namely multiple imputation by local resampling or local semi-parametric resampling, employing different sets of local weights (1) w_j, (2) $\breve{w}_j$, (3) $\breve{w}_j$; see Section 2.3.

In a first scenario, Y observations are generated from a normal distribution with conditional mean $\mu(x) = -3{+}x{+}7x^2$ and conditional variance $\sigma^2(x) = \exp(3{+}0.2x)$. The completely observed X variable follows a uniform distribution on the interval [0,10]. Values are missing with conditional probability $1 - \pi(x) = \{1 + \exp(0.5 - 0.1(x{-}5)^2)\}^{-1}$, which is largest at the ends of the interval. With these specifications, the true value of the parameter of interest is $\mu = E\{\mu(X)\} = 235.33$ and the total percentage of missingness is $E\{\pi(X)\} = 0.57$. In this and all other scenarios we took the number of multiple imputations to be $m = 3$. Other values, $m = 5$ and $m = 10$, gave very comparable results and are not shown.

We generated 1000 samples $\{(X_i, Y_i, \delta_i), i = 1, \ldots, n\}$. Table 2.2 summarizes the main results for n=200. An arbitrary sample from this setting is shown in the left upper panel of Figure 2.1. The 75 solid dots are observed, the other 125 y-values are missing. As shown in the right lower panel, more response values are missing at both ends of the $[0, 1]$ interval. The quadratic mean and variance function are also shown.

In all non-parametric imputation methods, the standard normal kernel function was used and all bandwidths were kept fixed. For the non-parametric single imputation only one bandwidth is needed and was taken as 1.5. For the local semi-parametric resampling the bandwidth in Step 1 was $h = 0.25$ and in Step 2 we chose $g = 1.5$. The local resampling method used the same bandwidth $h = g = 0.25$ in both steps. These choices are based on some initial experiments.

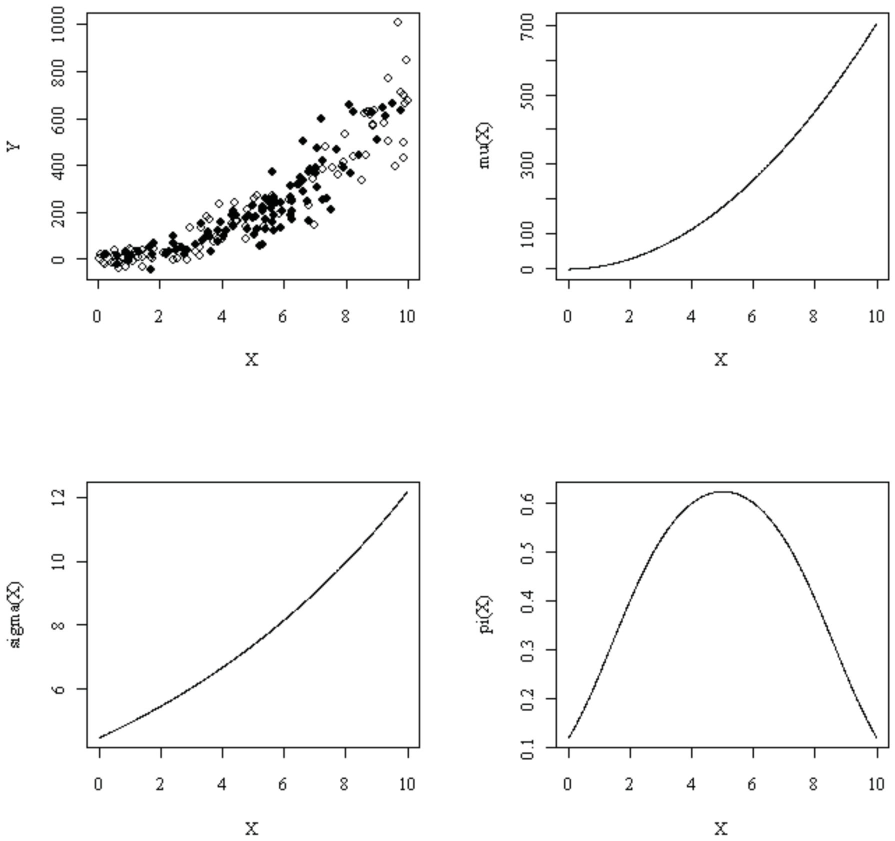


Figure 2.1: Scenario 1, simulation setting 1: an arbitrary sample (left upper panel), the mean function $\mu(x)$ (right upper panel), the function $\sigma(x)$ (left lower panel) and the probability $\pi(x)$ (right lower panel)

For each imputation method and each run we computed the multiple imputation estimate $\hat{\mu}$, its estimated standard error $se(\hat{\mu})$ and a 95% confidence interval $\hat{\mu} \pm 1.96se(\hat{\mu})$. Averages of the point estimates are shown in columns 1 and 2 of Table 2.2. Column 3 shows the simulated standard error of $\hat{\mu}$. Columns 4 and 5 show the average length of the 1000 confidence intervals and the simulated coverage probability. Rubin and Schenker (1986) suggested adjusting additionally for the multiple imputation by using critical points based on a t-distribution with $(m-1)\{1+(m/(m+1))(\hat{W}/\hat{B})\}^2$ degrees of freedom. The average lengths of these adjusted confidence intervals and simulated coverage probabilities are shown in columns 6 and 7, only for the multiple imputation methods.

Table 2.2: Simulation results for the first scenario. For each method: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). True value is $\mu = 235.33$. PSI: parametric single imputation method, Rubin PMI: Rubin's parametric multiple imputation method, Efron PMI: Efron's parametric multiple imputation method, NPSI: non-parametric single imputation method, LSR: local semi-parametric resampling method, LR: local resampling method. For the latter two, local weights (1) w_j, (2) $\breve{w}_j$, (3) $\check{w}_j$ are used.

method	ave($\hat{\mu}$)	ave($S(\hat{\mu})$)	sse($\hat{\mu}$)	ave.CI	sim.cov	adj. ave CI	adj.cov
All data	235.41	15.83	15.77	62.07	0.954	-	-
CC	214.71	19.86	19.82	77.87	0.788	-	-
PSI	215.33	14.93	16.93	58.54	0.682	-	-
Rubin PMI	215.23	17.26	17.22	67.67	0.759	70.03	0.778
Efron PMI	215.12	17.08	17.38	66.94	0.739	69.07	0.755
NPSI	236.57	15.22	18.11	59.78	0.889	-	-
LSR(1)	235.86	17.58	18.13	68.96	0.925	72.39	0.925
LSR(2)	237.09	17.30	18.62	67.83	0.917	69.85	0.920
LSR(3)	236.55	17.64	18.42	69.16	0.925	72.31	0.932
LR(1)	233.53	17.38	18.71	68.13	0.919	71.97	0.924
LR(2)	234.45	17.20	18.66	67.43	0.919	69.85	0.921
LR(3)	234.20	17.52	18.87	68.69	0.916	71.97	0.920

Next to linearity of $\mu(x)$, all parametric multiple imputation methods assume a constant variance $\sigma^2(x)$. Moreover Rubin's parametric multiple imputation assumes X to be normally distributed. The local resampling and local semi-parametric resampling approaches do not violate any model specifications.

As expected, the complete-cases method and the parametric imputation methods clearly underestimate the true mean μ while the non-parametric approaches perform much better. A comparison of the averages of the estimated standard errors and the simulated standard errors confirms the need for multiple imputation.

Note that the average lengths of the confidence intervals and the associated coverage probabilities are equal or larger for the construction based on a t random

Table 2.3: Simulation results for the second scenario. For each method: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). True value is $\mu = 17.33$. LSR: local semi-parametric resampling method, LR: local resampling method. Local weights (1) w_j, (2) $\breve{w}_j$, (3) $\check{w}_j$ are used.

method	ave.($\hat{\mu}$)	ave.($S(\hat{\mu})$)	sse($\hat{\mu}$)	ave.CI	sim.cov.	adj. ave. CI	adj.cov.
LSR(1)	17.75	1.74	1.79	6.83	0.938	7.53	0.948
LSR(2)	18.24	1.72	1.94	6.75	0.906	7.30	0.918
LSR(3)	17.66	1.72	1.76	6.73	0.936	7.36	0.946
LR(1)	18.00	1.72	1.82	6.76	0.927	7.35	0.933
LR(2)	18.48	1.77	2.01	6.94	0.898	7.59	0.918
LR(3)	17.53	1.70	1.76	6.66	0.936	7.22	0.945

variable (Rubin and Schenker, 1986) for all multiple imputation methods. This approach reduces to the normal confidence intervals for single imputation.

For this scenario there is not much difference between the local semi-parametric resampling and local resampling methods; both improve significantly upon the parametric methods. Also, there are almost no differences between the different local weighting schemes.

In a second scenario, response data follow a 6:4 mixture of $N\{\mu(x), \sigma^2(x)\}$ and $\text{Exp}\{1/\mu(x)\}$, where $\mu(x) = 6 + (x-2)(x-4) + 5\cos(\pi x)$, $\sigma(x) = \exp(0.02x)$, and $\text{logit}\{\pi(x)\}{=}2 - 0.4x$, resulting in $\mu = 17.33$. Since there is more misspecification, differences between parametric and non-parametric methods are more pronounced. Table 2.3 gives the simulation results for the local methods using bandwidths $h = 1$ and $g = 1.5$ for $n = 200$.

In this scenario, the semi-parametric methods, using a normal local likelihood in Step 2, turn out to be quite robust against the model misspecification. The local linearized weights $\breve{w}_j$ result in somewhat lower coverage probabilities, caused by a slight overestimation of μ. Better results might be obtained if the bandwidth were be optimized in each simulation run. The precise choice of local weights turns out to be of less importance.

For the simulated dataset shown in Figure 2.1, densities of the augmented response values are shown in Figure 2.2. The panel on the left shows densities of all response values (observed and missing), of the observed only and of the parametri-

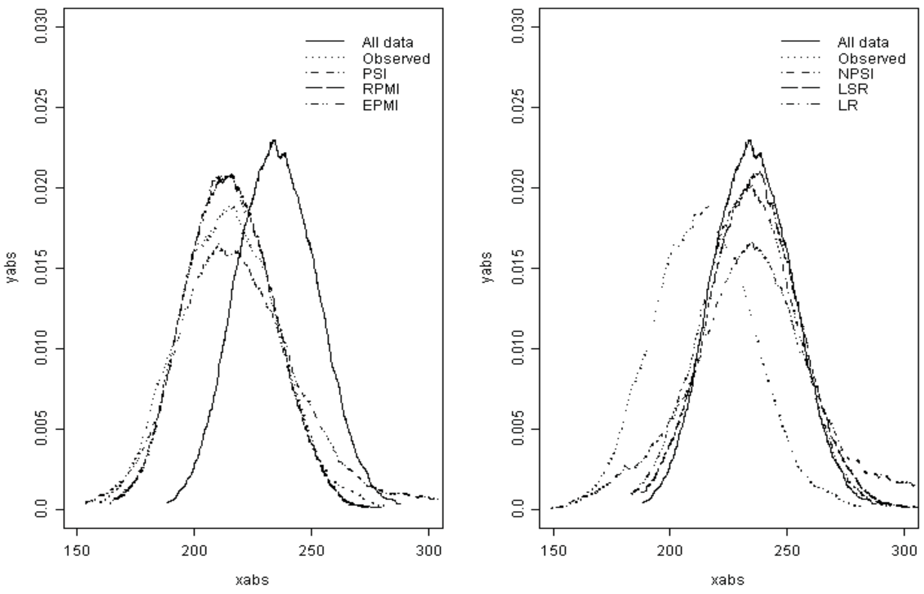


Figure 2.2: Scenario 1, simulation setting 1: Densities of response values (all and observed only) and augmented values for the different parametric imputation methods (left panel) and different non-parametric methods (right panel), for a sample within simulation setting 1.

cally augmented values. The panel on the right shows similar densities based on our non- and semi-parametric imputation method. For illustrative purposes only the weights w_j were used. It illustrates that the parametric methods impute values at the wrong location and that variability is wrongly incorporated by single imputation methods.

Several parameters and underlying functions may influence the behaviour of the different imputation methods. We experimented with some other simulation settings (componentwise modifications of setting 1). Results are shown in Table 2.4. In setting 2, the sample size was reduced from 200 to 100. The results are similar. The coverage of the single imputation based confidence intervals are unacceptable. The difference between a parametric and non-parametric approach is now less pronounced and, for local imputation, the estimator $S(\hat{\mu})$ seems to suffer from underestimation when the sample size is getting smaller. This is no surprise because local or non-parametric methods typically need more data to be really successful.

Setting 3 differs from setting 1 by another choice of $\pi(x) = \{1+\exp(-1+0.1(x-5)^2)\}^{-1}$ leading to a lower total percentage of missingness (45.90%). All methods are performing somewhat better now but the overall conclusion is the same.

Turning to a setting where more assumptions of the parametric MI methods are fulfilled, we first considered setting 4, similar to setting 1 but now with a constant conditional variance $\sigma^2(X) = 51$. This doesn't seem to lead to large changes or improvements for the parametric methods. The single non-parametric imputation method benefits most from this.

In a last setting 5 we chose a linear regression relation $\mu(x) = -3+70*x$ together with another probability $\pi(x) = \{1+\exp(-3+0.5*x)\}^{-1}$ leading to a true value of $\mu = 347$ and a total percentage of 41.57 % missingness. As can be seen from Table 2.5, the LR and LSR do not outperform the parametric approaches. The latter ones now use a correctly specified regression model. But the loss in efficiency by using unnecessarily a local imputation method remains very reasonable.

We experimented with some other variations of setting 1, all leading to essentially the same conclusions. The local imputation method improves upon the classical methods when one or more of the parametric assumptions are violated. When all assumptions underlying the parametric multiple imputation methods are fulfilled, local resampling and local semi-parametric resampling do not outperform the parametric approaches, although the loss in efficiency incurred by using unnecessarily a local imputation method remains small.

2.6.2 Jackknife Data Driven Bandwidth Selection

As an illustration, we applied the jackknife method of Section 2.5.2 to a randomly chosen sample obtained from the first scenario in Section 2.6.1, using the weights w_j, defined in (2.4). For the local resampling imputation, the grid 0.2, 0.25, 0.3, 0.5, 1, 2.5, 5, 20, 30, 40 was used for both constants C_h and C_g. In this way 100 estimates of $\hat{\mu}$ and the corresponding mean squared error of $\hat{\mu}$ were calculated. This resulted in a surface as shown in Figure 2.3(a). Figure 2.3(b) shows the estimated mean squared error as a function of $\hat{\mu}$ using a loess fit. This shows that lower values of the mean squared error correspond to estimates in the neighbourhood of the true value $\mu = 235.33$. The minimum is attained at $\hat{\mu} = 233.15$, with bandwidths $h = 0.601$ and $g = 0.024$. This latter plot also shows that different choices for h and g can lead to a wide range of $\hat{\mu}$-values, from about 225 to 260, indicating that a precise bandwidth choice is not unimportant.

Jackknifing with local semi-parametric imputation was also examined for the same sample, using a C_h, C_g-grid based on 1, 1.5, 2.5, 5, 7.5, 10, 15, 20, 30, 40.

Table 2.4: Scenario 1, simulation results for settings 2, 3 and 4: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). Local weights (1) w_j, (2) $\breve{w}_j$, (3) $\check{w}_j$ are used. True value $\mu = 235.33$.

Setting 2	average($\hat{\mu}$)	average($S(\hat{\mu})$)	simulated	average	sim. cov.	adj. average	adj. cov.
			se($\hat{\mu}$)	length of CI	prob.	length of CI	prob.
All data	235.52	22.38	22.67	87.72	0.946	-	-
CC	215.76	28.16	28.72	110.37	0.867	-	-
PSI	214.86	21.10	24.15	82.71	0.770	-	-
Rubin PMI	214.97	24.45	24.57	95.86	0.822	99.33	0.828
Efron PMI	213.94	24.28	24.50	95.17	0.815	98.62	0.827
NPSI	235.67	21.46	26.49	84.11	0.883	-	-
LSR(1)	235.03	24.15	27.05	94.66	0.913	97.97	0.925
LSR(2)	235.81	23.71	27.03	92.93	0.904	94.72	0.907
LSR(3)	235.23	24.13	27.20	94.60	0.914	97.74	0.921
LR(1)	232.20	23.45	27.33	91.92	0.892	94.75	0.898
LR(2)	232.89	23.09	27.23	90.53	0.886	92.07	0.889
LR(3)	232.56	23.47	27.74	92.01	0.886	94.80	0.890
Setting 3	average($\hat{\mu}$)	average($S(\hat{\mu})$)	simulated	average	sim. cov.	adj. average	adj. cov.
			se($\hat{\mu}$)	length of CI	prob.	length of CI	prob.
All data	235.41	15.83	15.77	62.07	0.953	-	-
CC	217.84	18.38	19.29	72.06	0.798	-	-
PSI	218.39	15.06	16.43	59.04	0.744	-	-
Rubin PMI	218.63	16.77	16.83	65.75	0.798	67.02	0.805
Efron PMI	218.31	16.72	16.85	65.53	0.794	66.73	0.802
NPSI	236.41	15.33	17.21	60.08	0.916	-	-
LSR(1)	235.83	17.15	17.41	67.23	0.942	69.33	0.945
LSR(2)	236.83	16.94	17.65	66.40	0.940	67.70	0.940
LSR(3)	236.09	17.20	17.44	67.42	0.938	69.56	0.941
LR(1)	233.70	16.96	17.28	66.47	0.930	68.27	0.932
LR(2)	234.73	17.01	17.42	66.69	0.943	68.31	0.946
LR(3)	234.47	17.11	17.44	67.06	0.933	69.01	0.938
Setting 4	average($\hat{\mu}$)	average($S(\hat{\mu})$)	simulated	average	sim. cov.	adj. average	adj. cov.
			se($\hat{\mu}$)	length of CI	prob.	length of CI	prob.
All data	235.38	15.38	15.38	60.30	0.953	-	-
CC	251.03	22.65	22.52	88.77	0.895	-	-
PSI	250.19	15.02	16.08	58.89	0.823	-	-
Rubin PMI	250.26	16.30	16.28	63.88	0.852	64.60	0.856
Efron PMI	250.10	16.31	16.28	63.92	0.854	64.65	0.855
NPSI	241.16	15.13	15.61	59.31	0.936	-	-
LSR(1)	241.43	15.88	15.78	62.24	0.941	62.54	0.942
LSR(2)	241.55	15.88	15.73	62.23	0.941	62.51	0.942
LSR(3)	241.29	15.87	15.77	62.19	0.943	62.48	0.945
LR(1)	235.72	15.98	15.89	62.65	0.954	62.97	0.954
LR(2)	235.49	15.99	15.67	62.68	0.952	63.02	0.955
LR(3)	235.45	15.96	15.84	62.55	0.953	62.88	0.953

Table 2.5: Scenario 1, simulation results for setting 5: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). Local weights (1) w_j, (2) $\breve{w}_j$, (3) $\check{w}_j$ are used. True value $\mu = 347$.

Setting 5	average($\hat{\mu}$)	average($S(\hat{\mu})$)	simulated	average	sim. cov.	adj. average	adj. cov.
			se($\hat{\mu}$)	length of CI	prob.	length of CI	prob.
All data	347.07	14.72	14.79	57.72	0.956	-	-
CC	252.36	16.69	16.84	65.43	0.000	-	-
PSI	347.12	14.53	15.23	56.95	0.942	-	-
Rubin PMI	347.33	15.50	15.42	60.76	0.956	61.37	0.959
Efron PMI	347.26	15.45	15.31	60.55	0.951	61.10	0.953
NPSI	347.16	14.54	15.43	56.99	0.935	-	-
LSR(1)	346.73	15.51	15.51	60.81	0.950	61.54	0.95
LSR(2)	347.59	15.39	15.54	60.31	0.949	60.75	0.95
LSR(3)	347.01	15.51	15.47	60.79	0.951	61.47	0.951
LR(1)	345.17	15.33	15.67	60.08	0.934	60.72	0.936
LR(2)	346.06	15.25	15.72	59.80	0.934	60.22	0.937
LR(3)	345.88	15.34	15.70	60.13	0.933	60.74	0.934

Larger values were needed, which seems plausible for a partly parametric approach. A plot of the estimated mean squared error versus $\hat{\mu}$ is shown in Figure 2.3(c).

The loess curve indicates a steeper descent towards the minimum, but on the other hand there is more variability. Estimates in the range of 224 to 231 have more or less the same associated mean squared error. For this sample, however, the local resampling method seems to do better. A similar experiment was done with sample size equal to 100 instead of 200. The result is shown in Figure 2.3(d). The curve seems to flatten out at its minimum of $\hat{\mu} = 236.39$, corresponding to $h = 0.396$ and $g = 3.170$.

Our conclusion is that the jackknife method gives promising results, but further research is needed.

2.7 The Vorozole Data

In this section we will illustrate both the local resampling and the local semi-parametric resampling on the Vorozole data as introduced in Section 1.5.1.

Goss *et al.* (1999) analyzed the Functional Living Index: Cancer (FLIC) using a two-way ANOVA model with effects for treatment, disease status, as well as their interaction. No significant difference was found. These data were further analyzed by Michiels *et al.* (1999).

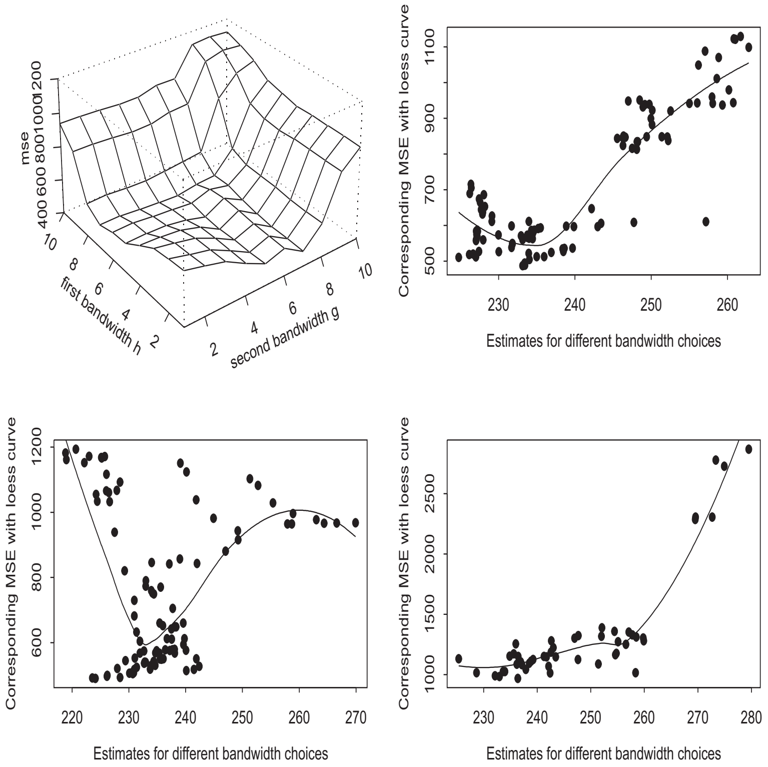


Figure 2.3: Scenario 1: (a) estimated mean squared error response surface plot for local resampling method with $n = 200$, (b) estimated mean squared error plotted against $\hat{\mu}$ for local resampling method with $n = 200$, (c) estimated mean squared error plotted against $\hat{\mu}$ for local semi-parametric resampling method with $n = 200$, (d) estimated mean squared error plotted against $\hat{\mu}$ for local resampling method with $n = 100$.

Here we estimate the mean FLIC score for patients at month 6 for both treatment arms separately, using a regression relation between the FLIC scores at month 1 (X) with those at month 6 (Y). Patients with no FLIC score at month 1 where excluded. About 50 % of all patients dropped out. A lower FLIC score at month 1 corresponds with a higher drop-out probability at month 6.

Table 2.6 summarizes the results. As expected the mean of the available, complete cases clearly overestimates the true mean FLIC score at month 6, for both treatments. The single imputation method PSI seems to correct for this but heavily underestimates the standard error. For both non-parametric multiple imputation methods LSR and LR using weights w_i, Table 2.7 and Figure 2.4 show the jackknife selected optimal bandwidths (gridsearch). There is almost no difference between the LSR and LR optimal bandwidths whereas the choices are like reversed for both treatments. Results for LSR and LR using weights $\breve{w}_j$ and $\breve{w}_j$ give similar findings.

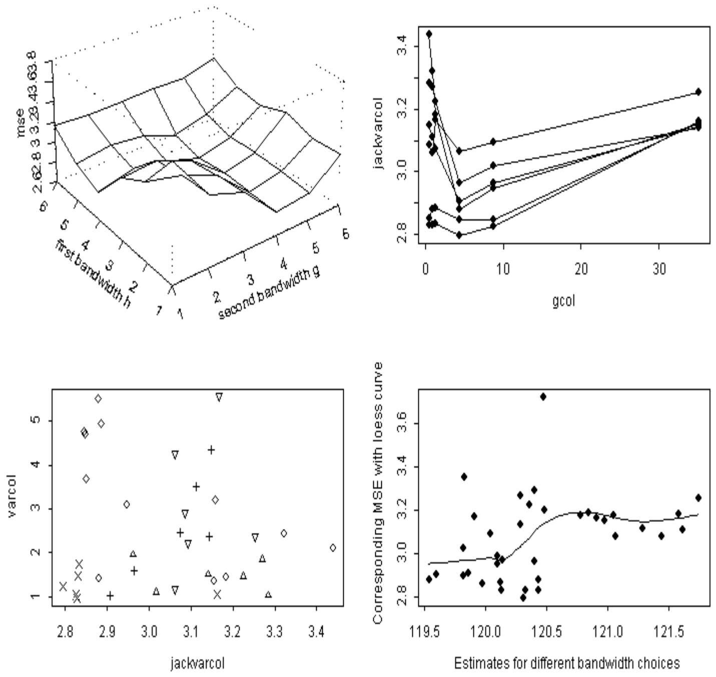


Figure 2.4: Vorozole data: (a) estimated mean squared error response surface plot for local resampling method, (b) estimated mean squared error plotted against $\hat{\mu}$ for local resampling, (c) estimated mean squared error plotted against $\hat{\mu}$ for local semi-parametric resampling method, (d) estimated mean squared error plotted against $\hat{\mu}$ for local resampling method.

Within a treatment arm, there is a striking similarity between all estimates based on multiple imputation. Whereas the simulations indicate that even a slightly misspecified parametric MI method can have serious problems, no such conclusions can be formulated here.

All point estimates take higher values for the vorozole arm but the confidence intervals indicate that the differences are not significant, which is in line with the results of Goss *et al.* (1999).

Table 2.6: Results for the Vorozole data: Estimates for the mean FLIC score $\hat{\mu}$ at month 6 with estimated standard error $s(\hat{\mu})$ and corresponding 95% confidence interval, for the vorozole (rows 1–2) and for the megestrole acetate (rows 3–4) treatment.

Method	CC only	PSI	Rubin PMI	Efron PMI	LSR(1)	LR(1)
$\hat{\mu}$	123.32	120.55	119.32	120.66	120.44	121.11
$s(\hat{\mu})$	3.69	1.35	2.80	5.62	3.24	4.00
CI	(119.5,127.1)	(118.3,122.8)	(116.0,122.6)	(116.0,125.3)	(116.9,124.0)	(117.2,125.0)
adj CI	-	-	(115.9,122.8)	(114.3,127.1)	(116.4,124.5)	(116.3,126.0)
$\hat{\mu}$	119.15	116.68	116.02	116.93	116.86	116.38
$s(\hat{\mu})$	4.19	1.79	2.96	2.68	4.98	2.60
CI	(115.1,123.2)	(114.1,119.3)	(112.7,119.4)	(113.7,120.1)	(112.5,121.2)	(113.2,119.5)
adj CI	-	-	(112.6,119.5)	(113.7,120.2)	(111.4,122.4)	(113.1,119.6)

Table 2.7: Results for the Vorozole data: Jackknife selected bandwidth for the LR and LRS methods.

Method	LSR(1)		LR(1)	
Bandwidth	h	g	h	g
vorozole	2.95	0.95	2.95	0.59
megestrole acetate	1.18	2.94	1.76	2.94

2.8 Discussion

Dealing with missing data via parametric multiple imputation methods usually implies stating several strong assumptions about both the distribution of the data and about underlying regression relationships. If such parametric assumptions do not hold, the multiply imputed data are not appropriate and might produce inconsistent estimates and thus misleading results. In this chapter, a fully non-parametric and a semi-parametric imputation method were introduced. Focus was on missing response data and in particular on the overall mean of the response variable Y. Estimators of other moments of Y and functions thereof can be obtained in a straightforward manner. For example, the average k-th sample moment $\hat{\mu}_k = \sum_{\ell=1}^{m} \hat{\mu}_k(\ell)/m$ where $\hat{\mu}(\ell) = \sum_{i=1}^{n} \{\tilde{Y}_i(\ell)\}^k/n$, can be shown to be a consistent estimator of the k-th moment of Y.

The non-parametric imputation method is applicable in a wide variety of statistical models, and can in the same way be used for discrete response data. The small adaptation needed for semi-parametric resampling is the specification of the appropriate distribution function in step 2 of the algorithm.

If there is more than one parameter completely observed, local methods could take all of them into account. However, in high dimensions kernel based methods might loose some of their attractiveness because of the ever present curse of dimensionality. The semi-parametric imputation methods proposed by Wang *et al.* (2004) and Little and An (2004) overcome this deficiency in the single imputation setting.

In some specific situations, imputation of missing values when missingness is non-ignorable, has been addressed by several authors, e.g., Greenlees *et al.* (1982). In general, however this is not straightforward.

Chapter 3

Weighted Model Selection for Incomplete and Design-based Samples

3.1 Introduction

In a regression analysis, starting from a rich enough family of models and based on the data at hand, one or a few good models can be selected, e.g., using the Akaike Information Criterion (AIC). In case of missing data, simple deletion of the subsample of incomplete observations and treating the resulting subsample of so-called *complete cases* as a simple random sample has been shown to possibly lead to biased estimates, even when using a correct model (see e.g. Little, 1992; Zhao *et al.*, 1996). A similar problem occurs when the observations come from a complex survey design, i.e., when sampling from a finite population with unequal selection probabilities. Indeed, the probability that an observation is incomplete can also be considered as a selection probability for that observation to be included in the sample or not. Analyzing such design-based data as a simple random sample can also introduce bias (Horvitz and Thompson, 1952).

There is a vast literature on parametric and non-parametric models in case of incomplete or design-based samples, but most of it concerns estimation (assuming a correct model) rather than model selection. The naive use of model selection criteria however turn outs to be unreliable in case of the aforementioned complications in the data. Indeed, treating the complete cases or the design-based sample as just a simple random sample can invoke some effects to appear or disappear and thus suggest another (incorrect) model to be more adequate for the data at hand.

In the context of incomplete data, selection methods like the predictive divergence for incomplete observations (PDIO, Shimodaira, 1994) and the complete data AIC (AICcd, Cavanaugh and Shumway, 1998) have been proposed. These methods rely on modelling the complete data likelihood, which introduces an additional model selection problem, namely the selection of an appropriate model for the missingness mechanism (if not missing completely at random). In this chapter we focus on selecting appropriate models for the measurement part, while treating the missingness mechanism as a nuisance. We propose a modification of the AIC-criterion for regression models, based on reweighting the complete cases by their inverse selection probabilities. The latter selection probabilities, if unknown, are preferably estimated non-parametrically (using ,e.g., splines), in this way avoiding the selection of a parametric model with its assumptions for the missingness process. This weighting of completely observed cases can be seen as an implicit imputation of missing observations and is valid when the probability to be missing depends upon the observed values but not on the unobserved values (MAR in the terminology of Little and Rubin 1987, Section 1.2).

For the closely related situation of design-based samples, model selection has not been really investigated. In the next section, the motivating study illustrates both complications of missingness and design-based sampling. In Section 3.3, the weighted AIC-criterion is introduced and discussed, mainly for parametric models, but its applicability is also extended to non-parametric models. Indeed, analogous to the selection of an optimal model from a set of parametric candidate models, one can choose the optimal smoothing parameter in non-parametric regression based on the corrected AIC-criterion, as shown by Hurvich *et al.* (1998). We will modify this criterion to handle incomplete and design-based samples. In Section 3.4, a simulation study shows the improved performance of the weighted AIC-criterion. Section 3.5 and Section 3.6 discuss some other weighted model selection criteria and possible avenues of other model selection techniques.

3.2 Cervix Cancer Screening

The Cervix Cancer Screening data, as a part of the Belgian Health Interview Survey of 1997, were introduced in Section 1.5.2. In this particular dataset, two complications arise. Firstly, sampling in the HIS was based on a combination of stratification, multistage sampling and clustering (Kish, 1995). Secondly, about 30% of the 2893 women had one or more missing covariates for the variables of interest. These design issues, together with the likely occurrence of data to be missing, are inherent to surveys and should be taken into account when selecting an optimal model from a candidate set of models.

In Table 5.2 and 5.3, an overview of twelve different models, based on the variables in Table 5.1, is given together with the original AIC-criterion, three weighted versions and two imputation-based versions.

The first modification, 'AIC_{W_1}', corrects for the survey design, the second and fourth version, 'AIC_{W_2}' and 'AIC_I', correct for incomplete data and the combination of both can be found in versions, 'AIC_{W_1,W_2}' and 'AIC_{I,W_1}'. Table 5.3 shows that different models are chosen by the different versions of the AIC-criterion; so it indicates that ignoring missingness or ignoring the sampling design can possibly lead to inappropriate model choices. We refer to Chapter 5 for a more thorough discussion.

Based on a theoretical justification, the weighted AICs are defined in the next section.

3.3 Weighted Akaike Information Criterion

Based on observations $(\boldsymbol{x_i}, y_i), i = 1, ..., n$, consider the regression model

$$\boldsymbol{y} \sim f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta}), \tag{3.1}$$

where

$$\boldsymbol{y} = (y_1, \dots, y_n)^T, \ \ \boldsymbol{\theta} = (\theta(\boldsymbol{x}_1), \dots, \theta(\boldsymbol{x}_n))^T, \ \ \boldsymbol{\eta} = (\eta(\boldsymbol{x}_1), \dots, \eta(\boldsymbol{x}_n))^T.$$

Here f denotes the joint density of $\boldsymbol{y}$ (given $\boldsymbol{x}$), $\boldsymbol{\theta}$ the parameter of interest and $\boldsymbol{\eta}$ a nuisance parameter. The aim is to select an optimal or a few good models amongst a set of candidate models. Several model selection criteria have been developed, in different settings and with different types of complexities in data and models (see e.g. Akaike, 1973; Takeuchi, 1976; Schwarz, 1978; Spiegelhalter *et al.*, 2002).

Assume we start from a collection of models, in particular we consider models of the form (3.1) . The well-known AIC criterion (Akaike, 1973)

$$\text{AIC} = -2L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) + 2K, \tag{3.2}$$

with $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ denoting the loglikelihood of the model and $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$ the maximum likelihood (ML) estimator of $(\boldsymbol{\theta}, \boldsymbol{\eta})$, originates from information theory. Here K stands for the total number of estimated parameters, nuisance parameters included. The second term in the AIC formula is often interpreted as a penalization for complexity. The AIC was designed to be an approximately unbiased estimator of the expected *Kullback-Leibler Information* (KL). In general, the KL information between model f_0 (denoting the 'true' model) and model f (the approximating model (3.1)) is defined as (ignoring an 'historical' factor 2)

$$I(f_0, f) = E\{\ \log(\frac{f_0(\boldsymbol{y})}{f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta})})\}, \tag{3.3}$$

(expectation with respect to the true model) and can be interpreted as the information loss using f to approximate f_0, or as the distance from f_0 to f. This KL distance is not a metric, but it has the property that $I(f_0, f) \geq 0$ with equality only if $f \equiv f_0$.

3.3.1 Missing Data

In case of missing data, the naive use of only complete cases in the definition of $I(f_0, f)$ can lead to serious deficiencies in its applicability to measure the distance between models (and consequently also in the use of its empirical version, the AIC-criterion). For simplicity, let us consider classical regression and suppose data are generated by a true model

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \sigma_0^2 I_n), \tag{3.4}$$

where $\boldsymbol{\mu}_0 = (\mu_0(1), \ldots, \mu_0(n))^T$, $\mathcal{N}_n$ denotes an n-variate normal distribution and I_n the $n \times n$ identity matrix . Consider the approximating, or candidate, family of models

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 I_n), \tag{3.5}$$

where $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1; \boldsymbol{\theta}), , \ldots, \mu(\boldsymbol{x}_n; \boldsymbol{\theta}))^T$.

For this setting, $E\{\log f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta})\}$ can be written as (f now denoting the univariate normal density)

$$\begin{aligned} E\{\sum_{i=1}^{n} \log f(y_i; \mu(\boldsymbol{x_i}), \sigma^2)\} &= -\frac{n}{2}\log(2\pi\sigma^2) \\ &\quad -E\left[\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right]/(2\sigma^2). \end{aligned} \tag{3.6}$$

Using an analogous expression for $E\{\log f_0(\boldsymbol{y})\}$, it is easy to verify that

$$I(f_0, f) = \frac{n}{2}\log(\sigma^2/\sigma_0^2) + \frac{n}{2}\{\frac{\sigma_0^2}{\sigma^2} - 1\} + \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2). \tag{3.7}$$

It follows that this measure is minimized as a function of σ^2 and $\boldsymbol{\mu}(\boldsymbol{\theta})$ (and equals 0) by taking $\sigma^2 = \sigma_0^2$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}_0$.

Now, let us introduce the missingness process. For $i = 1, \ldots, n$, define the indicator $\delta_i = 1$ if $(\boldsymbol{x_i}, y_i)$ is fully observed and 0 otherwise. In general it is possible that $\pi_i = P(\delta_i = 1) = \pi(\boldsymbol{x_i}, y_i, z_i)$, so the probability that the i-th observation is not fully observed is allowed to depend on $\boldsymbol{x_i}$, y_i or even on the value z_i of an other, completely ignored, variable. In this chapter we restrict attention to the MAR setting, implying that π_i does not depend on z_i, that it additionally does not depend on $\boldsymbol{x_i}$ (resp. y_i) in case $\boldsymbol{x_i}$ (resp. y_i) might be missing.

The use of complete cases (CC) only (those for which $\delta_i = 1$) (and hence ignoring the missing data mechanism) is translated in a replacement of (3.6) by

$$\begin{aligned} E\{\sum_{i=1}^{n} \delta_i \log f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)\} &= -\frac{E\{\text{trace}(D)\}}{2} \log(2\pi\sigma^2) \\ &\quad -E\left[\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T D \{y - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right] / (2\sigma^2), \end{aligned} \tag{3.8}$$

where $D = \text{diag}(\delta_1, \ldots, \delta_n)$. As a function of σ^2 and $\boldsymbol{\mu}(\boldsymbol{\theta})$, and using a saturated model $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ for the mean function, this expression (3.8) is maximized and the corresponding CC version of the KL distance

$$\begin{aligned} I_{CC}(f_0, f) &= E\{\sum_{i=1}^{n} \delta_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} \\ &= \frac{E\{\text{trace}(D)\}}{2} \log(\frac{\sigma^2}{\sigma_0^2}) + E\left[\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T D \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right] / (2\sigma^2) \\ &\quad + E\{\boldsymbol{z}^T D \boldsymbol{z}\} \frac{1}{2} \left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + E\{\boldsymbol{z}^T D\} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})) \left(\frac{\sigma_0}{\sigma^2}\right), \end{aligned} \tag{3.9}$$

(with $\boldsymbol{z} = (\boldsymbol{y} - \boldsymbol{\mu}_0)/\sigma_0$) is minimized at

$$\tilde{\theta}_i = \frac{E\{y_i \pi_i\}}{E\{\pi_i\}} = \mu_0(i) + \frac{\text{Cov}(y_i, \pi_i)}{E\{\pi_i\}}, \tag{3.10}$$

and

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} E[\pi_i \{y_i - \tilde{\theta}_i\}^2]}{\sum_{i=1}^{n} E\{\pi_i\}}. \tag{3.11}$$

In the above expressions and in what follows, moment related operators like the expectation E or the covariance (Cov) act on the random variables y_i and δ_i and treat $\boldsymbol{x_i}$ as nonrandom.

First of all, under a MCAR (missing completely at random) mechanism, $\pi_i = \pi$ and the above solutions simplify and are equal to the 'true' values, $\mu_0(i)$ and σ_0^2, respectively. The same holds in the MAR case that y_i is missing with probability $\pi_i = \pi(\boldsymbol{x_i})$, only depending on $\boldsymbol{x_i}$. If however π_i does depend on y_i in a way that $\mathrm{Cov}(y_i, \pi_i) \neq 0$, $I_{CC}(f_0, f)$ reaches a different minimum at (3.10) and (3.11). In fact, since by definition $I_{CC}(f_0, f_0) = 0$, this minimal value is negative (which is undesirable for a distance measure). If, e.g., y_i and π_i are positively correlated, then $\tilde{\mu}_i > \mu_0(i)$. This is to be expected since observations with smaller values of y_i are discarded with higher probability. Also for nonsaturated models for $\boldsymbol{\mu}(\boldsymbol{\theta})$, such kind of anomalies can be shown.

The AIC-criterion (3.2) based on the complete cases is given by

$$\mathrm{AIC}_{CC} = -2\sum_{i=1}^{n} \delta_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_{CC}), \hat{\sigma}^2_{CC})] + 2K, \tag{3.12}$$

where $\hat{\boldsymbol{\theta}}_{CC}$ and $\hat{\sigma}^2_{CC}$ are the ML estimators, maximizing the CC-loglikelihood (as described by the first term in (3.12)). For classical regression and ignoring constants, this can be simplified to

$$\mathrm{AIC}_{CC} = \left(\sum_{i=1}^{n} \delta_i\right) \log(\hat{\sigma}^2_{CC}) + 2K. \tag{3.13}$$

In case of MCAR, criterion (3.12) (or 3.13) is an approximately unbiased estimate of $I_{CC}(f_0, f)$ and is expected to behave appropriately (the missingness just results in an implicit sample size reduction). But for the MAR setting with missingness probabilities depending on the response, nothing guarantees that the above AIC criteria will serve any longer as useful model selection criteria.

The shortcomings of a CC approach, as described above, can be circumvented by a simple modification of the KL distance $I_{CC}(f_0, f)$ and corresponding AIC_{CC}-criterion. This modification is inspired by the technique of weighted estimation. Assuming a correct model is used, Flanders and Greenland (1991) and Zhao and Lipsitz (1992) showed that the use of weighted estimators, solving the weighted estimating equations (WEE)

$$\sum_{i=1}^{n} w_i \Psi(y_i; \boldsymbol{\theta}, \boldsymbol{\eta}) = 0, \tag{3.14}$$

with Ψ the derivative of the log(quasi)likelihood and with weights w_i inversely proportional to the missingness probabilities, are consistent and asymptotically unbiased. The idea of WEE was inspired by the Horvitz-Thompson estimator in the closely related setting of design-based samples with unequal selection probabilities (see Horvitz and Thompson, 1952). In Section 3.3.2, we further exploit this setting and its similarity with missing data for model selection.

Analogous to (3.14), a weighted KL distance can be defined as

$$I(f_0, f; w) = E\{\sum_{i=1}^{n} w_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\}. \tag{3.15}$$

Taking the weights

$$w_i = \delta_i/\pi_i, \tag{3.16}$$

the deficient distance $I_{CC}(f_0, f)$ is rectified and turned into the original data KL distance ('original' referring to the 'full' data, before introducing missingness). Indeed,

$$E\{\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} \quad = \quad \sum_{i=1}^{n} E\{\log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\}.$$

In a similar way, the weighted AIC-criterion

$$\text{AIC}_W = -2\sum_{i=1}^{n} w_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_W), \hat{\sigma}_W^2)] + 2K, \tag{3.17}$$

with w_i as in (3.16) and with $\hat{\boldsymbol{\theta}}_W$ and $\hat{\sigma}_W^2$ the weighted ML estimators (maximizing the weighted maximum likelihood), is expected to behave appropriately, i.e., to correct for the missing data. Indeed, denote $\hat{\boldsymbol{\theta}}_{\text{O}}$ and $\hat{\sigma}_{\text{O}}^2$ the ML estimators based on the original data, and consider the Taylor expansion (linear terms cancelling out)

$$\begin{aligned} & -2\sum_{i=1}^{n} w_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_{\text{O}}), \hat{\sigma}_{\text{O}}^2)] \qquad (3.18) \\ \approx \quad & \text{AIC}_W - 2\left((\hat{\boldsymbol{\theta}}_{\text{O}} - \hat{\boldsymbol{\theta}}_W)\ (\hat{\sigma}_{\text{O}}^2 - \hat{\sigma}_W^2)\right) \mathcal{I}_n(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}_W^2) \left((\hat{\boldsymbol{\theta}}_{\text{O}} - \hat{\boldsymbol{\theta}}_W)\ (\hat{\sigma}_{\text{O}}^2 - \hat{\sigma}_W^2)\right)^T, \end{aligned}$$

where the matrix $\mathcal{I}_n$ is the matrix of second derivatives of the weighted log-likelihood, evaluated at $(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}_W^2)$. The expected value of the left-hand side equals the expected value of the AIC-criterion based on the original data. Since both estimates, the 'original' $(\hat{\boldsymbol{\theta}}_{\text{O}}, \hat{\sigma}_{\text{O}}^2)$ and the 'weighted' $(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}_W^2)$, are estimating the same parameter (being the true value $(\boldsymbol{\theta}_0, \sigma_0^2)$ in case the model under consideration is a correct model), the second term in the right hand side is negligible, at least in a first order approximation.

For a normal regression model with $\mu(\boldsymbol{x_i}, \boldsymbol{\theta}) = \boldsymbol{x_i\theta}$, $i = 1, \ldots, n$, where $\boldsymbol{x_i} = (1\ x_{i1} \ldots n_{ip})$ and $\boldsymbol{\theta} = (\theta_0\ \theta_1 \ldots \theta_p)^T$, the weighted AIC-criterion can be rewritten in terms of squared residuals

$$\text{AIC}_W = (\sum_{i=1}^{n} w_i) \log\left(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i}\right) + 2(p+2), \tag{3.19}$$

where e_i are the residuals from the fitted model, using weighted ML. In the context of robust model selection procedures, Agostinelli (2002) introduced a robust modification of the AIC-criterion, based on the weighted likelihood methodology. He proposed a similar weighted AIC_W-criterion, but with weights downplaying the contribution of highly influential outliers.

Of course, typically the missing probabilities are unknown and have to be estimated, introducing essentially two further complications: i) finding appropriate estimates $\hat{\pi}_i$ which is again a model selection problem and ii) the effect on the characteristics of AIC_W when using weights

$$\hat{w}_i = \delta_i / \hat{\pi}_i. \tag{3.20}$$

Regarding the first complication, we suggest the use of a non-parametric or flexible semi-parametric estimator (generalized additive models (gam) or, e.g., classification trees for more complicated data structures, as illustrated in Section 3.4). This avoids the need for another model selection step. It is also important to note that, since the estimation of the missingness probabilities is a step *prior to* the envisaged model selection exercise, and hence is common to all candidate models under consideration, it has no effect on the penalization term in the expression of AIC_W. Concerning the second complication: rather than focusing on a theoretical study of the effect of estimating π_i on the expected value of AIC_W (a Taylor expansion immediately shows highly 'untractable' bias expressions), we opted for examining the finite sample performance of AIC_W with estimated weights by a simulation study (see Section 3.4).

In analogy to its expression based on the original data (Hurvich and Tsai, 1989), we define a bias-corrected weighted AIC as

$$\text{AIC}_W^{cor} = \text{AIC}_W + \frac{2K(K+1)}{\sum_{i=1}^{n} w_i - K - 1}. \tag{3.21}$$

This small-sample correction (second-order bias adjustment) has been especially recommended in a setting where there are many parameters in relation to the size of the sample n (for more details see Burnham and Anderson, 2002). Its performance in some simulations is briefly discussed in Section 3.4.1.

3.3.2 Design-Based Samples

Assume a finite population consisting of N units with measurements $\mathcal{M} = \{y_1, \ldots, , y_N\}$. A particular sampling plan leads to the random variable $\delta_i = 1$ if the i-th unit is included in the sample (and 0 otherwise) with $n = \sum_{i=1}^{N} \delta_i$ the total sample size. The selection probabilities are defined as $\pi_i = P(\delta_i = 1)$, for $i = 1, \ldots, N$. The choice $\pi_i = n/N$ corresponds to a simple random sample. In this finite population setting, only the δ_i are to be considered as random; the set $\mathcal{M}$ is to be considered as unknown but fixed.

Supposing that the population $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ is a single realization of a true 'superpopulation' model $f_0(\cdot)$, using the approximating model $f(\cdot; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)$ and treating the sample indicated by the δ_i as a random sample, a KL distance similar to the $I_{CC}(f_0, f)$ measure in (3.9) can be defined as (with now the expectation E with respect to the δ_i's, conditional on the 'realized' population)

$$\begin{aligned} I_{DB}(f_0, f) &= E\{\sum_{i=1}^{N} \delta_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} \\ &= \sum_{i=1}^{N} \pi_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]. \end{aligned} \tag{3.22}$$

For true and approximating models as in (3.4) and (3.5), with now $\Pi = \text{diag}(\pi_1, \ldots, \pi_n)$, $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1; \boldsymbol{\theta}), , \ldots, \mu(\boldsymbol{x}_N; \boldsymbol{\theta}))^T$ and $\boldsymbol{\mu}_0 = (\mu_0(1), \ldots, \mu_0(N))^T$ and with $\boldsymbol{z} = (\boldsymbol{y} - \boldsymbol{\mu}_0)/\sigma_0$ as before, we get

$$\begin{aligned} I_{DB}(f_0, f) &= \frac{\text{trace}(\Pi)}{2} \log(\frac{\sigma^2}{\sigma_0^2}) + \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T \Pi \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2) \\ &\quad + z^T \Pi z \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + z^T \Pi (\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta}))\left(\frac{\sigma_0}{\sigma^2}\right). \end{aligned} \tag{3.23}$$

As an example, consider a simple two-valued true superpopulation model

$$\boldsymbol{\mu}_0 = (\mu_0(1), \ldots, \mu_0(N_1), \mu_0(N_1 + 1), \ldots, \mu_0(N))^T = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2)^T$$

with $\mu_1 \neq \mu_2$, and the incorrect constant model $\boldsymbol{\mu}(\theta) = (\theta, ..., \theta)^T$. For this incorrect model, the minimal distance $I_{DB}(f_0, f)$ is at least as small as its value at $\tilde{\sigma}^2 = \sigma_0^2$ and

$$\tilde{\theta} = \frac{\sum_{i=1}^{N} \pi_i y_i}{n}. \tag{3.24}$$

Using the correct two-parameter mean model with $\sigma^2 = \sigma_0^2$, $I_{DB}(f_0, f)$ is minimized at

$$\tilde{\mu}_1 = \frac{\sum_{i=1}^{N_1} \pi_i y_i}{n_1}, \quad \tilde{\mu}_2 = \frac{\sum_{i=1}^{N_2} \pi_i y_i}{n_2}, \tag{3.25}$$

where $n_1 = \sum_{i=1}^{N_1} \delta_i$ and $n_2 = \sum_{i=N_1+1}^{N} \delta_i$. Now, in the particular case that the selection probabilities induce a bias resulting in $\tilde{\mu}_1 = \tilde{\mu}_2$, the KL distance $I_{DB}(f_0, f)$ is exactly the same for both models and hence the incorrect model is indistinguishable from the correct model.

Identical to the case of missing data, the weighting of the KL distance and corresponding AIC-criterion, with weights as in (3.16), can be used to correct both measures. Note that in general the selection probabilities can depend on both $\boldsymbol{x_i}$ and y_i. In most applications the selection probabilities π_i are determined by the design of the sample and hence are known.

3.3.3 Design-Based Samples with Missing Observations

In typical surveys, as in the cervix cancer screening example introduced in Section 3.2, both complications occur together. In this case δ_i, indicating whether or not the i-th unit is in the sample and is fully observed, can be written as

$$\delta_i = \delta_i^D \delta_i^M, \tag{3.26}$$

where $\delta_i^D = 1$ if the i-th unit is included in the sample (as in Section 3.3.2) and $\delta_i^M = 1$ if the i-th unit is fully observed (as in Section 3.3.1). The weighted AIC (3.17) can now be based on weights $w_i = \delta_i/\pi_i$ where

$$\pi_i = P(\delta_i = 1) = P(\delta_i^M = 1|\delta_i^D = 1)P(\delta_i^D = 1). \tag{3.27}$$

These latter probabilities can be estimated by the product of the (known) probabilities $P(\delta_i^D = 1)$ and the (non-parametrically) estimated probabilities $P(\delta_i^M = 1|\delta_i^D = 1)$.

In the next section, we show how the idea of a weighted AIC can be extended to select a smoothing parameter for non-parametric regression.

3.3.4 Smoothing Parameter Selection using AIC$_W$

Assume

$$y_i = \mu_0(\boldsymbol{x_i}) + \epsilon_i, \;\; i, \ldots, n, \tag{3.28}$$

where $\mu_0(\cdot)$ is an unknown smooth function and $\epsilon_i, i = 1, \ldots, n$, are independent error terms with mean 0 and variance σ_0^2. Different linear smoothers for μ are available: orthogonal series, kernel estimators, splines, ... (see e.g. Simonoff, 1996). The most crucial choice for any smoother is the choice of the smoothing parameter. Hurvich, Simonoff and Tsai (1998) proposed to select this parameter α by minimizing the corrected AIC-criterion

$$\text{AIC}_\alpha^{cor} = n\log(\hat{\sigma}^2) + \frac{n + \text{trace}(S_\alpha)}{1 - \{\text{trace}(S_\alpha) + 2\}/n}, \tag{3.29}$$

where S_α is the smoother matrix for which $\hat{\boldsymbol{y}} = S_\alpha \boldsymbol{y}$.

In case of an incomplete or design-based sample, this criterion can be turned into a weighted version

$$\text{AIC}^{cor}_{\alpha,W} = (\sum_{i=1}^{n} w_i) \log \left(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i} \right) + \frac{\sum_{i=1}^{n} w_i + \text{trace}(S_{W,\alpha})}{1 - \{\text{trace}(S_{W,\alpha}) + 2\}/(\sum_{i=1}^{n} w_i)}, \quad (3.30)$$

where $S_{W,\alpha}$ is the smoother matrix from the weighted fit. Taking $S_{W,\alpha}$ the classical regression 'hat matrix', (3.30) reduces (up to a constant) to (3.21).

To study the effects of weighting more closely, a simulation study in a variety of settings was conducted. The next section summarizes our main findings. All computations were conducted in R 2.0 (R Development Core Team, 2004).

3.4 Simulations

In the first two scenarios, we consider a setting with missing covariate data. The third scenario focuses on design-based samples and the last scenario on the selection of the smoothing parameter in non-parametric regression.

3.4.1 Scenario 1: Parametric Model Selection for Incomplete Data

In the initial setting, the set of candidate models contains the true model.

Initial Setting

In this first scenario, uniform$[0, 10]$ x-values were generated, together with (independently) Bernoulli(0.5) z-values. Given x and z, response y-values were generated from a normal distribution with mean $\mu_0(x, z) = -3 + 3x + 5x^2$ and variance $\sigma_0^2 = \exp(5)$. x-observations were then turned missing with conditional probability (see middle panel in Figure 3.1),

$$\pi(y, z) = 1 - [1 + \exp\{1 - 0.009(y - 300)\}]^{-1}. \quad (3.31)$$

Not depending on unobserved x-values, the missingness process is MAR. Let n denote the total sample size and n_c the number of complete observations. We generated 1000 different samples $\{(x_i, z_i, y_i), i = 1, \ldots, n\}$, with fixed design $\{x_i, z_i, i = 1 \ldots, n\}$. For each sample, 8 different regression models were fit, all submodels of $\mu(x, z) = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 z + \beta_4 xz$.

Four different 'strategies' are compared: i) AIC on the original data, before introducing missingness (what we would get if no values were missing), ii) (unweighted)

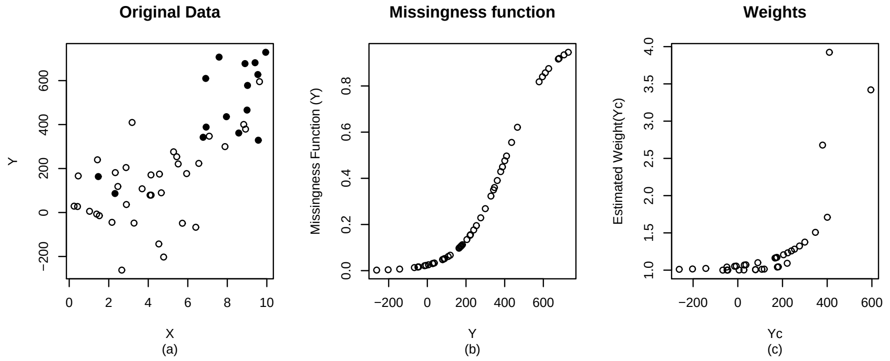


Figure 3.1: Scenario 1, an arbitrary chosen sample: (a) original sample, complete cased (white bullets) and unobserved data (black bullets); (b) missingness probabilities; (c) estimated weights.

AIC on the complete cases only (ignoring missingness), iii) weighted AIC using the true weights (3.16) and iv) weighted AIC, using the estimated weights (3.20). The probabilities (3.31) are estimated by gam estimates $\hat{\pi}(y, z)$ (using the R package `mgcv` 1.8, Wood 2001). On average 35% of the x-values were missing. In Figure 3.1, a typical dataset for Scenario 1 is shown together with the missingness probabilities and the estimated weights. This latter figure shows a double curve, as a consequence of the additive model in x and z (being binary). The upper part of Table 3.1 displays the results for $n = 50$. Each column (from 2 to 9) corresponds to a particular model and the numbers show how often the respective model has been selected by AIC under the four strategies mentioned above. Models more complex than the true quadratic model $\{x, x^2\}$ can be considered as correct models, the others as incorrect models. The last rightmost column shows the total number of times a correct model was chosen. The table shows that for the initial setting, the unweighted AIC applied on the complete cases, very often selects the incorrect simpler model $\{x\}$. This is to be expected since the missingness is mainly located at the larger y-values (which of all response values mostly represent the quadratic effect). The weighted versions correct for that. The one with true weights selects about 10% more often a correct model, though it less often selects the true model, while the one with the estimated weights shows an improvement of about 9% and it selects 1% more often the true model.

We computed the average of the fitted values based on the selected model, together with 95% pointwise confidence intervals, using AIC on the original data, (unweighted) AIC on the complete cases, and weighted AIC on the complete cases.

Table 3.1: Scenario 1: The numbers indicate how often a model has been selected, for the four strategies. The last column shows how often a correct model has been chosen, out of 1000. This scenario is repeated for different settings.

	1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
						xz	z	z, xz	classified
Scenario 1: Initial Setting $n = 50, \sigma_0^2 = \exp(5), \text{slope} = 5, \%(\text{miss}) = 35$									
Original Data	0	272	0	467	55	40	85	81	633
Complete Cases	0	447	0	274	97	53	81	48	403
True Weighted	0	271	0	254	125	99	101	150	505
Est. Weighted	0	329	0	286	100	83	102	106	494
Scenario 1: Variance exp(5.3)									
Original Data	0	396	0	374	65	47	70	48	492
Complete Cases	9	540	2	210	107	56	48	28	286
True Weighted	4	330	3	170	131	140	87	135	392
Est. Weighted	5	372	2	198	130	117	78	103	379
Scenario 1: Missingness 20%									
Original Data	0	275	0	496	38	31	93	67	656
Complete Cases	0	451	0	311	90	54	49	45	405
True Weighted	1	290	0	286	80	104	93	146	525
Est. Weighted	1	355	0	308	79	70	80	109	497
Scenario 1: Smaller Quadratic Effect: slope = 3									
Original Data	0	459	0	297	82	55	63	44	404
Complete Cases	6	548	1	225	87	57	47	29	301
True Weighted	5	414	0	224	107	92	87	71	382
Est. Weighted	4	450	2	245	102	75	74	58	377
Scenario 1: Sample Size 100									
Original Data	0	114	0	666	31	18	106	65	837
Complete Cases	0	312	0	452	65	35	91	45	588
True Weighted	0	199	0	371	67	61	129	173	673
Est. Weighted	0	228	0	416	70	56	110	121	647

The resulting curves are shown in Figure 3.2. The middle figure clearly shows the bias when using the unweighted AIC on the complete cases. The use of the weighted AIC nicely corrects the average best model in the direction of the true underlying curve.

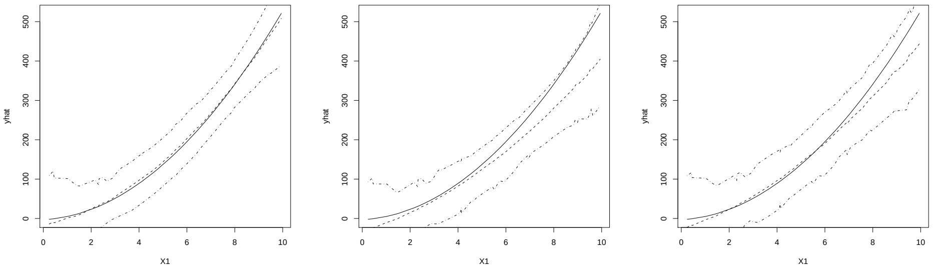


Figure 3.2: Scenario 1: Average best model with 95% pointwise confidence intervals for the original data (left), the complete cases with unweighted AIC (middle) and with weighted AIC (right). The solid curve is the true function $\mu_0(x, z)$

In the other parts of Table 3.1 similar results for variations on Scenario 1 are shown: a larger error variance, less missingness, smaller quadratic effect and larger sample. Figure 3.3 displays the number of correct models as a function of error variance σ_0^2, missingness percentage (by changing the coefficient of y in equation (3.31)), quadratic effect of x in $\mu_0(x, z)$ and sample size n. All curves show the decrease in selecting a correct model when using the unweighted AIC on the complete cases. The difference gets more pronounced for increasing error variance, increasing missingness and increasing quadratic effect of x in $\mu_0(x, z)$. Note that this latter increasing effect implicitly generates more missingness via, on average, increasing response values y (see equation (3.31)).

The use of the weighted version improves the performance of the AIC and the version with known weights is consistently choosing more correct models than with estimated weights. On the other hand the version with estimated weights constantly performs better than with true weights in selecting the only true model. One might argue that the gain by using the weighted AIC is not so spectacular but rather moderate, that it tends to select more complicated models and that, thinking critically further along these lines, always taking the "most complex model"(including x, x^2, z and xz) is actually the best criterion (since it leads to a 100% correct classification according to our definition of a correct model). But first of all, we have to realize that correcting for missing information is often a hard exercise, since information in

Table 3.2: Scenario 1: MASE-values and bias-variance decomposition based on the original fixed design. On the one hand the model is selected using the AIC- and AIC$_W$-criterion, on the other hand the most complex model is chosen.

	Model Selection	bias^2	var	MASE
Original Data	min AIC	39.26	2085.05	2124.32
	most complex	2.25	2253.05	2255.30
Complete Cases	min AIC	2433.37	2485.58	4918.95
	most complex	1986.74	2964.73	4951.47
True Weighted	min AIC$_W$	460.62	3984.71	4445.33
	most complex	404.51	4289.29	4693.81
Est. Weighted	min AIC$_W$	738.53	3153.06	3891.60
	most complex	608.09	3595.19	4203.28

available data might be very scarce. Next, the selection of somewhat more complicated models might be justified in this setting and not just arbitrary. Moreover a needless complex model will be accompanied with larger variability in its estimates. To show that the weighted AIC does not just select more complex models in an arbitrary way, but leads to models with an improved accuracy, Table 3.2 shows, for the initial setting, mean averaged squared errors (together with squared bias-variance decomposition)

$$\text{MASE} = \frac{1}{1000}\sum_{r=1}^{1000}\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^{(r)}(x_i,z_i) - \mu_0(x_i,z_i))^2\right\}, \tag{3.32}$$

for the different AIC selected models together with that of the "most complex model". Here, $\hat{\mu}^{(r)}(x_i, z_i)$ denotes the fitted value within simulation run r. This table shows that choosing the most complex model is not a sensible strategy (as expected) and more importantly that the weighted AIC does lead to a considerable improvement. Also for the original data, choosing the "most complex model"gives an increase in MASE. Just using complete cases has a disastrous effect on the quality of the selected fits (particularly on the bias), whereas the use of the estimated weighted AIC leads to the best results in terms of MASE. Indeed, the latter reduces bias, at the cost of a moderate increase in variance. That the use of estimated rather than true weights leads to the smallest MASE-values is in accordance with known results in related settings (see e.g. Robins *et al.*, 1994; Rotnitzky and Robins, 1995).

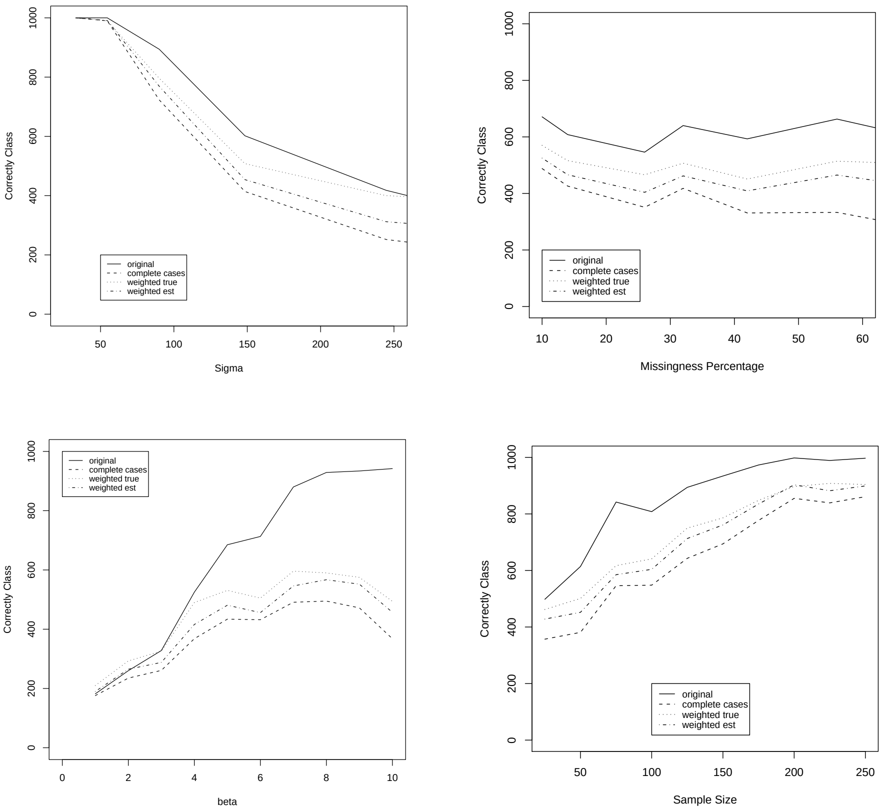


Figure 3.3: Scenario 1: Correctly selected models for different sigma-values (upper left), for different missingness percentages (upper right), for different quadratic effects (lower left) and for different sample sizes (lower right).

Non-parametric Weighting Methods

Different smoothers can be used to estimate the missingness probabilities $\pi(y, z)$. In Scenario 1, equation (3.31) shows that these probabilities only depend on y. In Section 3.4.1, these probabilities were estimated with a gam model, as a function of both y and z. In this section we illustrate how results differ when using different smoothers: gam using y only, Nadaraya-Watson (NW) kernel estimate using both y and z or y only, with fixed or with data-driven bandwidth (cross-validation).

Table 3.3 shows that the best results are obtained when using a penalized spline, especially the one as a function of y only. The other numbers are more or less similar. The fixed bandwidth $h = 150$ for the NW-estimator was chosen by visual inspection

Table 3.3: Scenario 1, initial setting: Model selection using different smoothers to estimate the weights.

	x	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
				xz	z	z, xz	classified
Complete Cases	447	274	97	53	81	48	403
NW h=150 (y, z)	342	270	106	84	102	96	468
NW h=150 (y)	337	288	114	76	96	89	473
NW CV (y, z)	315	257	108	96	103	121	481
NW CV (y)	336	287	114	75	96	92	475
gam CV(y, z)	329	286	100	83	102	106	494
gam CV (y)	278	282	107	109	103	121	506
True Weights	271	254	125	99	101	150	505

of some of the generated samples. Main conclusion is that the choice of smoother and smoothing parameter is not unimportant. It is also recommendable to carefully examine the missingness process, so that accurate estimation of the probabilities is possible.

Corrected AIC

For small sample sizes, the use of the corrected AIC-criterion (3.21) is recommended. The results in Table 3.4 are based on the corrected AIC-criterion for the initial setting of Scenario 1 but with $n = 30$. The improvement is considerable. The true model is chosen most often using the weighted AIC, especially when the weights are estimated (this latter phenomenon was also noticeable in Table 3.1).

3.4.2 Scenario 2: Generating Model Not Included

We now consider the (more realistic) setting that the set of candidate models does not contain the true model. The response y is generated as in Scenario 1, but now with mean function $\mu_0(x, z) = -3 - 3\log(x + 1) + 5x^2$. The same set of candidate models is considered. As before, a generalized additive model was used to estimate the weights. Since now direct comparison with the true model, nor a categorization in correct or incorrect models is possible anymore, the average of the fitted values based on the selected model, together with 95% pointwise confidence intervals, using

Table 3.4: Scenario 1, with sample size 30: Model selection using the corrected AIC-criterion.

	1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
						xz	z	z, xz	classified
Original Data	0	435	0	392	77	31	40	25	457
Complete Cases	16	616	3	217	80	34	26	8	251
True Weights	6	398	1	260	129	77	61	68	389
Est. Weights	8	442	0	275	122	53	56	63	394

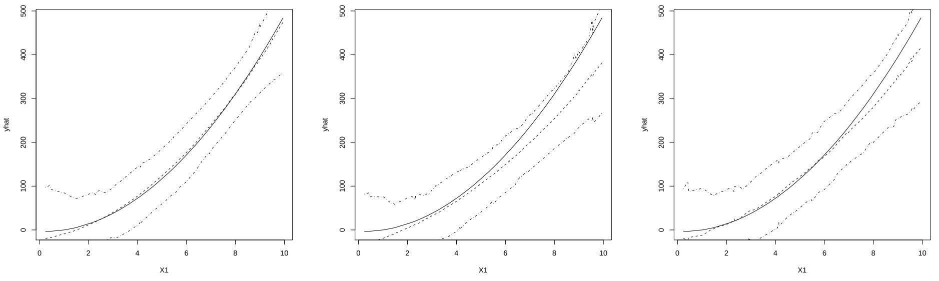


Figure 3.4: Scenario 2: Average best model with 95% pointwise confidence intervals for the original data (left), the complete cases with unweighted AIC (middle) and with weighted AIC (right). The solid curve is the true function $\mu_0(x, z)$

AIC on the original data, (unweighted) AIC on the complete cases, and weighted AIC on the complete cases are shown in Figure 3.4. The resulting curves show a similar behaviour as for Scenario 1. Indeed, also for this scenario the weighted AIC corrects the average best model in the direction of the true underlying curve, while the unweighted AIC on the complete cases results in a considerable bias.

Similarly to Scenario 1, Table 3.5 shows the MASE-values and bias-variance decomposition for the different methods. The benefit in using the AIC_W-criteria is reflected in the MASE-values and the behaviour of the bias and variance components is similar to Scenario 1 (Table 3.2).

3.4.3 Scenario 3: Model Selection for Design-Based Samples

To illustrate the use of the weighted AIC for design-based samples, a population $\{y_1, \ldots, y_N\}$ of size $N = 1500$ was generated, as a single realization from the su-

Table 3.5: Scenario 2: MASE-values and bias-variance decomposition based on the original fixed design. On the one hand the model is selected using the AIC- and AIC$_W$-criterion, on the other hand the most complex model is chosen.

	Model Selection	bias^2	var	MASE
Original Data	min AIC	41.58	2079.93	2121.50
	most complex	2.90	2236.82	2239.72
Complete Cases	min AIC	2040.05	2310.80	4350.85
	most complex	1638.04	2750.06	4388.10
True Weighted	min AIC$_W$	382.79	3516.66	3899.45
	most complex	307.85	3802.61	4110.46
Est. Weighted	min AIC$_W$	439.66	3128.05	3567.70
	most complex	374.15	3447.90	3822.05

perpopulation model f_0, being a normal distribution with variance σ_0^2 and mean $\mu_0(i) = \mu_1$ for $i = 1, \ldots, 500$ (group 1), $\mu_0(i) = \mu_2$ for $i = 501, \ldots, 1000$ (group 2), $\mu_0(i) = \mu_3$ for $i = 1001, \ldots, 1500$ (group 3).

In a first setting 1000 samples were taken by dividing this population into three strata based on the ordered population y values: the 200 smallest y-values, the middle 900 y-values and the 400 largest y-values. The sample was then taken as follows: a population unit i (y_i) is selected for the sample with probability $p_1 f$ when it belongs to the first or third stratum and with probability $p_2 f$ when it belongs to the second stratum. When $p_1 < p_2$, this results in an oversampling of the second stratum.

The (single) population was generated with $\mu_2 = \mu_3 = \kappa = -\mu_1$ with $\kappa > 0$. The simulation parameters κ, σ_0, f, p_1 and p_2 were set to different values as shown in Table 3.6. For each of the samples, 5 different models were fit: (1) $\mu_i = \mu, i = 1, \ldots, 3$, (2) $\mu_1 = \mu_2 \neq \mu_3$, (3) $\mu_1 \neq \mu_2 = \mu_3$, (4) $\mu_1 = \mu_3 \neq \mu_2$, and (5) $\mu_i \neq \mu_j$ for $i \neq j$. Model (3) is the true model, model (5) is another correct model. The other models assume $\mu_1 = \mu_2$ or $\mu_1 = \mu_3$ and are incorrect (for $\kappa \neq 0$).

In a first setting, where $\{\kappa, \sigma_0, f\} = \{0.5, 3, 0.5\}$, sampling was done according to different choices of (p_1, p_2), ranging from simple random sampling $p_2/p_1 = 1$ to highly unequal stratified sampling $p_2/p_1 = 11$. The results in Table 3.6 show an improved selection for the AIC$_W$-criterion compared to the AIC-criterion. Models (3) and (5) are chosen more frequently by the AIC$_W$-criterion.

Table 3.6: Scenario 3, first setting: The number of models chosen by AIC and AIC$_W$, for different variations of the basic setting and different choices of p_1 and p_2.

		AIC						AIC$_W$					
p_1	p_2	1	2	*3*	4	*5*	*Cor*	1	2	*3*	4	*5*	*Cor*
Basic													
0.05	0.55	321	110	445	107	17	462	128	192	277	133	270	547
0.10	0.50	284	101	498	92	25	523	155	146	424	136	139	563
0.20	0.40	191	116	594	63	36	630	156	132	572	60	80	652
0.30	0.30	133	108	639	64	56	695	125	108	648	63	56	704
$\sigma_0 = 4$													
0.05	0.55	467	108	301	115	9	310	134	205	281	189	191	472
0.10	0.50	428	117	325	118	12	337	209	199	328	161	103	431
0.20	0.40	331	121	450	75	23	473	259	144	471	72	54	525
0.30	0.30	305	136	445	86	28	473	295	137	455	86	27	482
$\kappa = 1$													
0.05	0.55	13	31	817	25	114	931	27	89	397	62	425	822
0.10	0.50	6	8	841	11	134	975	9	23	604	20	344	948
0.20	0.40	2	5	850	2	141	991	2	6	786	2	204	990
0.30	0.30	0	1	842	0	157	999	0	1	840	0	159	999
$f = 0.2$													
0.05	0.55	494	113	249	133	11	260	116	211	240	204	229	469
0.10	0.50	481	142	241	128	8	249	227	193	280	189	111	391
0.20	0.40	440	130	304	112	14	318	351	158	321	129	41	362
0.30	0.30	364	133	360	123	20	380	368	130	364	118	20	384

Increasing σ_0 (more noise) results in model (1) to be chosen more frequently. Also to be expected, a larger choice of κ (group 1 more different) leads more often to correct model choices. The fraction parameter f was initially chosen 0.5, resulting in a sample of size 225. To reflect the behaviour for a smaller sample, f was set to 0.2, resulting in a larger variability due to the smaller sample size (= 90). For all variations of the basic setting, AIC$_W$ improves the selection from slightly to

Table 3.7: Scenario 3, second setting: The number of models chosen by AIC and AIC$_W$, for different variations of the basic setting and different choices of p_1 and p_2.

		AIC						AIC$_W$					
p_1	p_2	1	2	*3*	4	*5*	*Cor*	1	2	*3*	4	*5*	*Cor*
						Basic							
0.05	0.55	92	120	56	596	136	192	66	175	510	52	197	707
0.10	0.50	189	19	392	381	19	411	46	171	590	12	181	771
0.20	0.40	126	131	651	31	61	712	60	197	615	7	121	736
0.30	0.30	133	108	639	64	56	695	125	108	648	63	56	704
						$\sigma_0 = 4$							
0.05	0.55	162	266	27	389	156	183	156	307	377	56	104	481
0.10	0.50	370	59	215	349	7	222	144	276	475	28	77	552
0.20	0.40	289	168	472	44	27	499	137	283	500	14	66	566
0.30	0.30	305	136	445	86	28	473	295	137	455	86	27	482
						$\kappa = 1$							
0.05	0.55	0	0	316	599	85	684	0	0	613	3	384	997
0.10	0.50	0	0	757	64	179	936	0	0	709	0	291	1000
0.20	0.40	0	3	845	1	151	996	0	2	775	0	223	990
0.30	0.30	0	1	842	0	157	999	0	1	840	0	159	999
						$f = 0.2$							
0.05	0.55	336	138	108	385	33	141	243	254	356	77	70	426
0.10	0.50	439	64	219	270	8	227	263	236	395	62	44	439
0.20	0.40	359	167	381	76	17	398	250	240	439	46	25	464
0.30	0.30	364	133	360	123	20	380	368	130	364	118	20	384

substantially (according to the ratio p_2/p_1), except for $\kappa = 1$.

In a second setting, the same population was taken but now design-based sampling was based on two strata, the 300 largest y-values of the third group and the remaining 1200 y-values. Sampling was done as follows: a population unit i is selected with probability $p_1 f$ when it belongs to the first stratum and with probability $p_2 f$ when it belongs to the second stratum. If $p_1 < p_2$ this results in an undersampling of units in the third group with the larger y values. The results for 1000 such samples are shown in Table 3.7, again for the same basic setting and variations thereof. One can see that the AIC-criterion very often chooses the incorrect model (4) $\mu_1 = \mu_3 \neq \mu_2$ and the AIC$_W$-criterion corrects this choice to model (3) $\mu_1 \neq \mu_2 = \mu_3$, which is the true model. For all variations of this setting, the AIC$_W$ outperforms AIC in all cases. The differences are much more pronounced than in the previous setting. One can also observe that the number of times a correct model is selected by the AIC$_W$-criterion is more or less the same for all different choices of (p_1, p_2). When sampling probabilities are equal and thus a simple random sample is taken, the choices made using AIC and AIC$_W$ are essentially the same.

3.4.4 Scenario 4: Smoothing Parameter Selection in Nonparametric Regression for Incomplete Data

For this scenario, $n = 200$ x-values were generated from uniform$[0, 1]$, and corresponding y-values from a normal distribution with mean $\mu_0(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ and variance $\sigma_0^2 = 0.4$ Range(y). This corresponds to one of the simulation settings used in Hurvich *et al.* (1998). Next, x observations were turned missing with probability

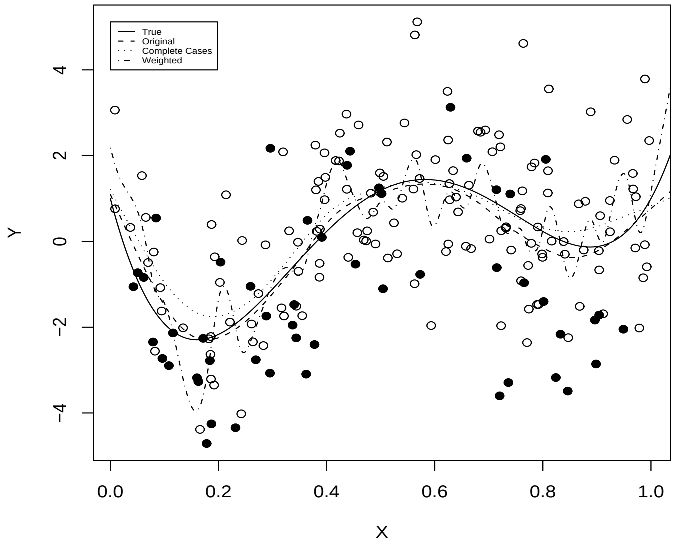
$$\pi(y) = [1 + \exp\{2 - 0.1(y - 2)^2]^{-1}. \tag{3.33}$$

For each of the 1000 generated samples $\{Y_i, i = 1, \ldots, n\}$ with a fixed design $\{x_i, i = 1 \ldots, n\}$, a smoothing spline was fitted (using `smooth.spline` in R) according to three methods, and with smoothing parameter selected by AIC (as introduced by Hurvich *et al.*, 1998). The first method is based on the original data, while the second method is based on the complete cases only and finally the third method weights the complete cases (at the model selection stage and at the final fitting stage) with $\hat{w}_i = 1/\hat{\pi}_i$ where $\hat{\pi}_i$ is the estimated probability for a complete case to be observed. The estimation of π_i is also based on a smoothing spline with smoothing parameter again determined by AIC.

The left panel in Figure 3.5 displays an arbitrary sample together with the fitted splines. The white dots indicate the observed data, while the black dots show the unobserved or missing data. The spline using the weights tends to severely undersmooth.

In this context, Wahba (1990) uses the unbiased variance estimator

$$\hat{\sigma}_U^2 = \frac{y^T(I - S_\alpha)^2 y}{\text{trace}(I - S_\alpha)}, \tag{3.34}$$



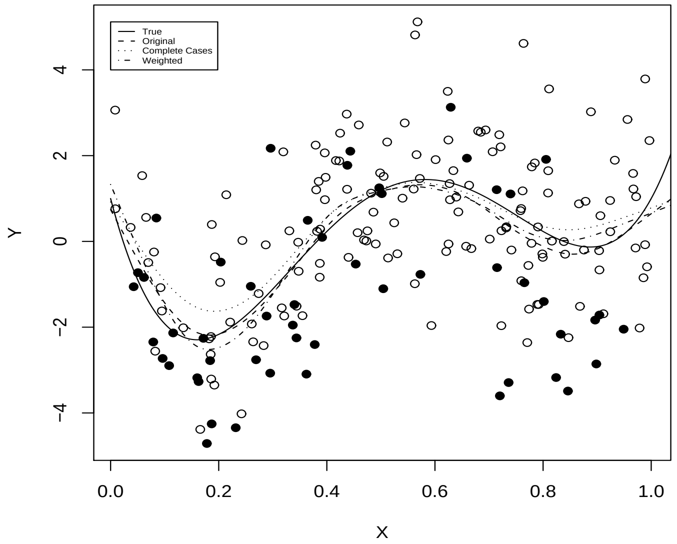


Figure 3.5: Scenario 4: Simulated dataset with spline curves according to the different methods together with the true function, using the ML variance estimator $\hat{\sigma}^2_{ML}$ (upper panel) and the unbiased variance estimator $\hat{\sigma}^2_U$(lower panel).

Table 3.8: Scenario 4: The average number of parameters using variance estimator $\hat{\sigma}^2_{ML}$ or $\hat{\sigma}^2_U$.

	$\hat{\sigma}^2_{ML}$	$\hat{\sigma}^2_U$
Original Data	8.33	6.99
Complete Cases	7.55	6.31
Weighted	18.31	9.00

where S_α is the smoother matrix. The use of $\hat{\sigma}^2_U$ instead of $\hat{\sigma}^2_{ML}$ is equivalent to an extra penalization of $-n\log(\text{trace}(I - S_\alpha))$, which corrects for undersmoothing, as can be seen for the fit of a random sample in the right panel of Figure 3.5. This is also confirmed by Table 3.8. It shows the simulation average of the equivalent number of parameters, selected by the three methods (rows) and for both variance estimators (columns). The models using the unbiased estimator are generally smoother and this reduction in equivalent number of parameters is very substantial for the weighted analysis. Other simulations confirmed this and therefore we certainly recommend the use of the unbiased estimator $\hat{\sigma}^2_U$ for the weighted method.

In Figure 3.6, the true curve (the solid curve) and the simulation average of the fitted curves for all three methods and both variance estimators, together with 95% pointwise confidence intervals, are shown. Again, the beneficial effect on the smoothing when using the unbiased variance estimator is illustrated. The middle panels show that there is substantial bias at both minima, when using the complete cases without weighting. The weighted AIC does correct for bias, as shown in the right panels.

To assess the goodness of fit quantitatively for each of the fits, MASE-values were calculated for each method and each variance estimator. The boxplots in Figure 3.7 show again that the weighted AIC method is not resulting in an improvement when using $\hat{\sigma}^2_{ML}$, but that it does when using $\hat{\sigma}^2_U$.

3.5 Other Model Selection Criteria

Next to the AIC, several other model selection criteria have been developed and can be extended to a weighted version to handle incomplete and design-based samples. For a model M with p regression parameters, the Mallows' C_p-criterion, developed as an estimator of the relative mean squared error, is very popular for least squares regression. Its definition Cp $= n\hat{\sigma}^2(M)/\hat{\sigma}^2(F) - (n - 2p)$ where $\hat{\sigma}^2(M)$ ($\hat{\sigma}^2(F)$) is

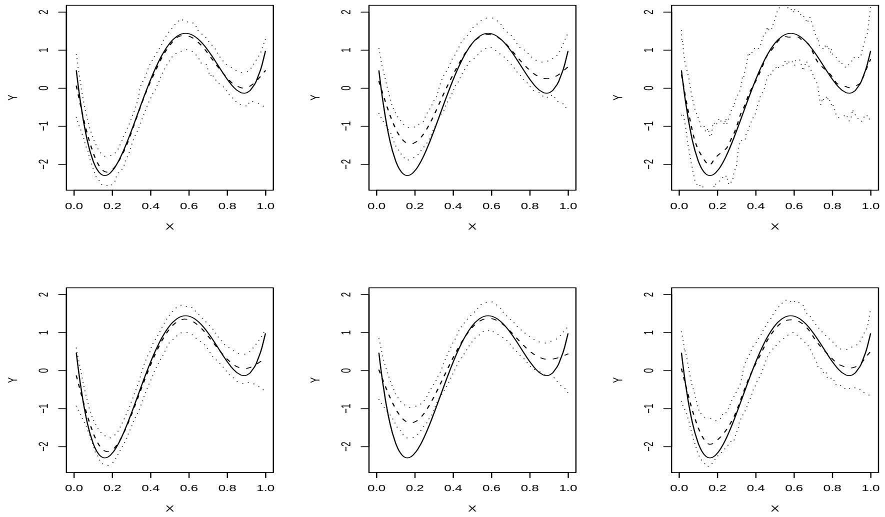


Figure 3.6: Scenario 4: Average of the fitted values (dashed curve) based on the chosen models over simulation runs together with the true function (solid curve) and 95% confidence intervals (dotted curves). From left to right: the original data, the complete cases and the weighted complete cases, using either $\hat{\sigma}^2_{ML}$ (upper row) or $\hat{\sigma}^2_U$ (lower row).

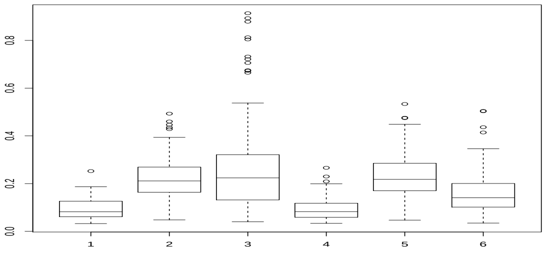


Figure 3.7: Scenario 4: Boxplots of simulated MASE-values for the different methods: original data, $\hat{\sigma}^2_{ML}$ (1), complete cases, $\hat{\sigma}^2_{ML}$ (2), weighted complete cases, $\hat{\sigma}^2_{ML}$ (3), original data, $\hat{\sigma}^2_U$ (4), complete cases, $\hat{\sigma}^2_U$ (5), weighted complete cases, $\hat{\sigma}^2_U$ (6).

the estimated variance based on a reduced model M (respectively full model F), can be modified in the weighted version

$$\mathrm{Cp}_W \quad = \quad \left(\sum_{i=1}^{n} w_i\right) \frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i e_i^{*2}} - \left(\sum_{i=1}^{n} w_i - 2p\right),$$

where e_i and e_i^* are the residuals based on reduced model and full model, respectively. Analogously, the Bayesian information criterion BIC $= n(\log \hat{\sigma}^2_{ML}) + \log(n)K$ (for classical regression) can be modified in a weighted version

$$\mathrm{BIC}_W = \sum_{i=1}^{n} w_i \left(\log \frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i} \right) + \log \left(\sum_{i=1}^{n} w_i \right) K.$$

Table 3.9: Scenario 1, basic setting: The number of chosen models by the Cp- and BIC-criteria.

	1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
						xz	z	z, xz	classified
Scenario 1: Basic Setting, Cp.									
Original Data	0	259	0	465	55	43	97	87	643
Complete Cases	0	424	0	280	104	57	84	51	415
True Weights	0	337	0	257	119	80	97	110	464
Est.Weights	0	375	0	289	104	69	89	80	458
Scenario 1: Basic Setting, BIC.									
Original Data	0	536	0	374	27	13	38	12	424
Complete Cases	1	702	1	196	54	17	24	5	225
True Weights	2	578	2	224	89	38	38	29	291
Est. Weights	2	651	0	211	68	27	27	20	258

We also investigated the performance of these alternative model selectors in a simulation study. As an illustration, Table 3.9 shows some results for the initial simulation setting of the first scenario. Up to expected differences, like the BIC-criterion selecting more simple models, a similar improvement is realized by the weighted selection criteria.

3.6 Discussion

The naive use of model selection criteria in case of incomplete and design-based samples can lead to the selection of inappropriate or non-optimal models. In this chapter we introduced a weighted Akaike information criterion. The weights are inversely proportional to the selection probabilities and if unknown, can be estimated non-parametrically. For incomplete data, the method can be seen as an implicit non-parametric imputation approach and its application is straightforward. Simulations show that the use of this weighted AIC-criterion results in an improved model selection for design-based samples. For incomplete data, the model-selection performance of the weighted AIC-criterion is somewhat less pronounced. But missing data are more problematic than design related complications. Moreover, the simulated MASE results are showing the improved accuracy of the AIC_W-selected models.

In case covariates are complete and Y-values are missing at random, valid and efficient parameter estimates are obtained using the complete cases only (Little, 1992). These results however do not apply in the case of missing covariates where the missingness probability depends on the completely observed response (Robins *et al.*, 1994; Zhao *et al.*, 1996). In the context of model selection, there is no need to distinguish between missing covariates and missing responses as such. The underlying motivation to use weights in either case however is different. When covariates are missing and missingness depends on y, the distribution of $Y|X$ is distorted and weights are used to correct for this. In case of nonresponse, it is not the distribution that is distorted but it is the finite sample behaviour that causes inadequate model selection. In the latter situation weights are used, as for design-based samples, to correct for this. It is not yet clear whether the use of a weighted AIC is more beneficial in the situation of missing covariates compared to nonresponse. This is topic of further research.

The other options to deal with missingness in the context of model selection are full likelihood methods, that models both measurement and missingness part simultaneously. This approach needs an additional model to be selected and is not extendable to the analogous setting of design-based samples. Another approach is to first impute missing observations and then select the model based on the augmented dataset. When the imputation model is flexible, as for example a generalized additive model can be, one can consider the choice of the imputation model to be a separate preliminary step in the model selection process. The latter approach is the topic of next chapter.

Chapter 4

Imputation-based Model Selection

4.1 Introduction

In the previous chapter, model selection for incomplete data relied on using a weighted AIC-criterion where the weights are the inverse probabilities for an observation to be observed. A natural alternative for this implicit imputation is an explicit imputation of missing data. Selecting an appropriate model is then done using the AIC-criterion on the augmented dataset. One might argue that model selection based on augmented data or on complete cases is not directly comparable because of the use of different samples. On the other hand, they all have the observed data in common and from there one is interested in which models are selected by AIC, not the comparison of AIC-values over the different methods as such.

In a recent paper by Carpenter and Kenward (2005) the use of inverse probability weighting and the use of multiple imputation to handle "missing at random"data are contrasted. In some situations data are missing not only for one variable but for several variables, possibly continuous and categorical. In those situations, an adequate imputation method is hard to find and the use of inverse probability weighting has some major advantages. However, when dealing with only a few fully observed subjects, i.e., none of the variables for that subject are missing, weighting shows major deficiencies since the complete cases do not contain enough information to justify an implicit imputation. Both methods as such, can be seen as complementary tools to handle missing data.

Several methods to impute missing data are known, ranging from 'naive' procedures such as unconditional mean imputation towards proper imputation methods such as multiple imputation (Rubin, 1978). Single and thus also mean imputation is improper as pointed out in Chapter 2, but its use as a first step in selecting a model is more than satisfactory as will be shown in this chapter. When exploiting the relation between X and Y by inverse regression to impute missing X-values, biased regression estimates result. Afifi and Elashoff (1969a,b) propose to use bias-corrected versions. Nielsen (2001) shows that the use of non-parametric (rather than parametric) conditional mean imputation results in consistent estimators. He uses a local linear regression method to impute the data. In the same line of thinking, we will use penalized regression splines to impute missing covariate values. To allow the use of more than one predictor variable, a generalized additive model using penalized regression splines as described by Wood (2001) will be used (see Section 1.3.1).

Although the results presented in this chapter address the missing covariate situation, similar results were obtained in case of nonresponse.

In a first section, a short simulation study shows the performance of the imputation-based model selection. Some limitations of imputation-based model selection will be addressed in Section 4.3. In Section 4.4, focus is on model selection after smoothing and we conclude with a discussion in Section 4.5.

4.2 Imputation-based Model Selection

In this section the simulation study of Section 3.4.1 is repeated, now using an imputation-based model selection. A first imputation uses a penalized regression spline of Y

$$X \sim s(Y). \tag{4.1}$$

However, in practice, one would include all possible information and thus Z to improve upon imputation. For this second imputation, we used a generalized additive model with penalized regression splines (Wood, 2000)

$$X \sim Z + s(Y) + Z * s(Y). \tag{4.2}$$

Let us focus on Scenario 1 and 2 of previous chapter.

4.2.1 Scenario 1

In Table 4.1, the results for the initial setting for Scenario 1, as introduced in Section 3.4.1, are shown together with those based on both imputation-based model

Table 4.1: Scenario 1: The numbers indicate how often a model has been selected, for the eight strategies. The last column shows how often a correct model has been chosen, out of 1000.

	1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
						xz	z	z, xz	classified
Initial Setting: $(n = 50, \sigma_0^2 = \exp(5), \text{slope} = 5, \%(\text{miss}) = 35)$									
Original Data	0	272	0	467	55	40	85	81	633
Complete Cases	0	447	0	274	97	53	81	48	403
True Weighted	0	271	0	254	125	99	101	150	505
Est. Weighted	0	329	0	286	100	83	102	106	494
Imputation Based (4.1)	0	173	0	533	34	34	117	109	759
Imputation Based (4.2)	0	184	0	471	38	34	139	134	744

selection methods, (4.1) and (4.2). It is seen from the correctly chosen models that both imputation methods improve selection even beyond the selection based on the original data. Imputation-based model selection using (4.2) performs about as well, i.e., selects the true model, as model selection based on the original data. Using (4.1) to impute the data gives an additional improvement of about 6% in choosing the true model. It can be seen that imputation based on (4.2) chooses more overly complex models compared to the imputation based on (4.1). This shift is caused by the creation of a Z-effect by imputing data based on (4.2). Selecting a correct model using the imputation-based selection techniques has by far the best results over all methods.

In Table 4.2, different settings of Scenario 1 were considered. When using a larger variance ($\sigma_0^2 = \exp(5.3)$), both imputation methods give an increase of more than 10% in the selection of the true underlying model compared to the complete cases (Table 3.1). Using (4.1), the selection is almost as good as based on the original data (0.8% less true models). For a lower missingness percentage (20%), both imputation-based model selection methods outperform model selection based on the original data. Using a smaller quadratic slope, only a moderate improvement is noticed when using imputation-based model selection compared to model selection based on the complete cases only and the weighting methods perform considerably better in selecting a correct model, although the true underlying function is not selected as often as for the imputation-based methods. A larger sample size, $n = 100$, shows again that the model-based imputation outperforms the weighting methods.

The weighted AIC is clearly outperformed by the use of imputation-based AIC except for the case of a small quadratic effect, where the weighting methods perform better in selecting a correct model as well as a true model. The use of a penalized regression spline of Y to impute X (4.1) gives a substantial increase over the use of a bivariate generalized additive model of Y and Z (4.2). It sometimes even overshoots the selection using the AIC on the original data. It seems that the semi-parametric imputation model in most situations captures the true underlying function extremely well and thus selection based on the imputed data reflects this. We will come back to this peculiar phenomenon in Section 4.4. Table 4.3 shows the MASE-values and bias-variance decomposition for both imputation methods. For reasons of comparison, the results of Table 3.2 are repeated in this table. Here, MASE-values were calculated, on the one hand, using the original fixed design (OD) and, on the other hand, using the observed part of the fixed design over simulations (CC). Looking at the OD-results, the bias reduces when using the imputation-based AIC compared to the true- and estimated-weighted AIC on the complete cases. The variance however increases substantially when using the imputation-based methods, resulting in a larger MASE(OD)-value. If we look at the CC-results, the variance reduces substantially, while the bias increases moderately. This shows that the increase in variance for the imputation-based methods comes from the imputed data

Table 4.2: Scenario 1: Selected models using both imputation-based selection methods for different settings of Scenario 1.

		1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
							xz	z	z, xz	classified
Initial Setting: $(n = 50, \sigma_0^2 = \exp(5), \text{slope} = 5, \%(\text{miss}) = 35)$										
$\sigma_0^2 = \exp(5.3)$	(4.1)	0	376	0	366	61	58	69	70	505
	(4.2)	0	363	0	319	78	66	96	78	493
%(miss)=20	(4.1)	0	170	0	605	26	18	92	89	786
	(4.2)	0	180	0	511	29	19	154	107	772
slope = 3	(4.1)	1	491	1	251	91	58	61	46	358
	(4.2)	1	482	1	242	91	69	67	47	356
n=100	(4.1)	0	46	0	711	15	12	129	87	927
	(4.2)	0	40	0	666	19	15	171	89	926

and not from the complete cases and so does the moderate increase in bias.

Table 4.3: Scenario 1: Imputation-based model selection: MASE and bias-variance decomposition based on the available and complete cases. On the one hand the model is selected using the AIC-criterion, on the other hand the most complex model is chosen.

	Model	bias^2		var		MASE	
	Selection	OD	CC	OD	CC	OD	CC
Original Data	min AIC	39.26	113.11	2085.05	2075.49	2124.32	2188.60
	most complex	2.25	58.18	2253.05	2226.20	2255.30	2284.38
Compl. Cases	min AIC	2433.37	2436.31	2485.58	2303.38	4918.95	4739.69
	most complex	1986.74	2063.29	2964.73	2676.31	4951.47	4739.60
True Weighted	min AIC_W	460.62	460.83	3984.71	3710.22	4445.33	4171.05
	most complex	404.51	402.50	4289.29	3927.89	4693.80	4330.39
Est. Weighted	min AIC_W	738.53	876.27	3153.06	2824.43	3891.60	3700.70
	most complex	608.09	772.77	3595.19	3140.29	4203.28	3913.06
Imputation	min AIC	140.27	119.16	5168.71	2200.35	5308.98	2319.21
(4.1)	most complex	85.12	106.90	5407.64	2312.62	5492.75	2419.53
Imputation	min AIC	147.46	109.75	5045.66	2352.39	5193.12	2462.15
(4.2)	most complex	80.80	100.34	5250.13	2453.09	5330.93	2553.42

Figure 4.1 shows the resulting curves of the average of the fitted values based on the selected model, together with 95% pointwise confidence intervals for the complete cases with weighted AIC, augmented data using (4.1) and (4.2), respectively. These figures show an increasing variability when using the imputation-based AIC-criteria compared to the weighted AIC-criterion. The increase in variability comes from the region with higher missingness probability where the main data-imputation takes place. This confirms our previous findings.

In the next section we investigate the performance of imputation-based model selection when the true underlying model is not part of candidate set of models, i.e., Scenario 2 of Chapter 3.

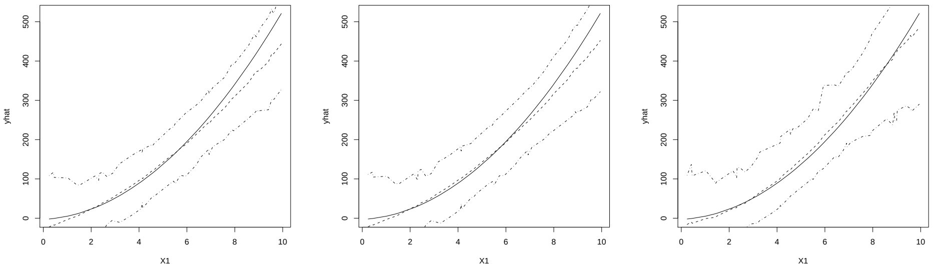


Figure 4.1: Scenario 1: Average best model with 95% pointwise confidence intervals for the complete cases with weighted AIC (left), augmented data using (4.1) (middle) and augmented data using (4.2) (right). The solid curve is the true function $\mu_0(x, z)$

4.2.2 Scenario 2

Let us consider Scenario 2 were $\mu_0(x, z) = -3 - 3\log(x + 1) + 5x^2$ as described in Section 3.4.2. Similarly to Table 4.3, Table 4.4 shows the MASE-results for the imputation-based methods. The conclusions from this table are similar to those for Scenario 1.

Similar to Figure 4.1, Figure 4.2 shows the resulting curves of the average of the fitted values based on the selected model, together with 95% pointwise confidence intervals for the three different methods. These figures indicate a similar behaviour as for Scenario 1.

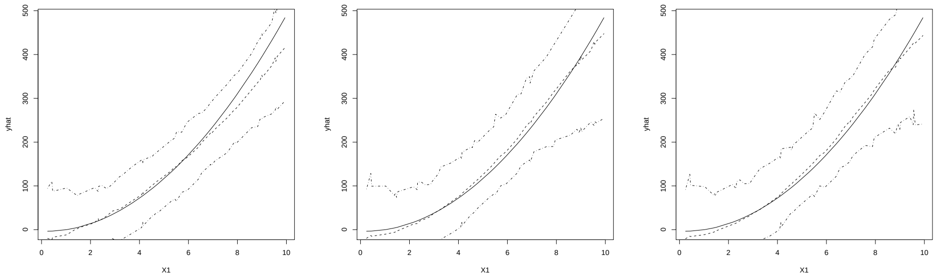


Figure 4.2: Scenario 2: Average best model with 95% pointwise confidence intervals for the complete cases with weighted AIC (left), augmented data using (4.1) (middle) and augmented data using (4.2) (right). The solid curve is the true function $\mu_0(x, z)$

These simulations show an improved model selection when using imputation-based methods. The imputation, based on a generalized additive model seems to be

Table 4.4: Scenario 2: Imputation-based model selection: MASE and bias-variance decomposition based on the available and complete cases. On the one hand the model is selected using the AIC-criterion, on the other hand the most complex model is chosen.

	Model	bias^2		var		MASE	
	Selection	AC	CC	AC	CC	AC	CC
Original Data	min AIC	41.58	92.88	2079.93	2054.75	2121.50	2147.63
	most complex	2.90	40.05	2236.82	2208.54	2239.72	2248.59
Compl. Cases	min AIC	2040.05	1973.02	2310.80	2179.40	4350.85	4152.42
	most complex	1638.04	1661.70	2750.06	2527.05	4388.10	4188.76
True Weighted	min AIC$_W$	382.79	349.28	3516.66	3323.59	3899.45	3672.87
	most complex	307.85	297.16	3802.61	3524.97	4110.46	3822.13
Est. Weighted	min AIC$_W$	439.66	485.19	3128.05	2865.52	3567.70	3350.71
	most complex	374.15	421.72	3447.90	3109.23	3822.05	3530.96
Imputation	min AIC	145.64	126.04	4591.55	2063.58	4737.20	2189.62
(4.1)	most complex	101.65	113.26	4886.85	2246.59	4988.50	2359.85
Imputation	min AIC	151.83	118.20	4492.10	2195.02	4643.93	2313.22
(4.2)	most complex	98.60	111.89	4747.08	2375.32	4845.68	2487.21

capable of restoring the true underlying model. Therefore, a closer look at model selection using smoothed data will be provided in Section 4.4. Let us first point out several limitations towards imputation-based model selection.

4.3 Limitations to Imputation-based Model Selection

The applicability and the performance of imputation-based model selection highly depends on the (unknown) missingness process and on the imputation model used.

In the previous section a generalized additive model was used to impute the data. The flexibility of a generalized additive model with penalized regression splines (Section 1.3.1) allows the choice of the imputation model to be considered as a preliminary step to the selection of a model. However, the choice of the imputation technique is not unimportant and a careful examination of the missingness process

Table 4.5: Imputation-based model selection: The number of chosen models by the different methods based on the AIC-criterion.

Scenario 1: $\pi(y) = 1 - [1 + \exp\{1 - 0.009\lvert y - 300\rvert\}]^{-1}$.									
Bandwidth\ Model	1	X	Z	X, X^2	X, Z	$X, X^2,$	$X, Z,$	$X, X^2,$	correctly
						Z	XZ	Z, XZ	classified
Original Data	0	121	0	631	26	19	130	74	835
Complete Cases	3	469	0	251	100	75	63	40	354
True Weights	0	64	1	166	54	129	198	389	753
Est. Weights	0	103	1	255	73	137	142	290	687
Imputation (4.2)	0	205	0	187	153	242	84	129	400

is recommendable.

If $E(x|y)$ is not well defined, imputation using a generalized additive model fails since the relationship between x and y cannot be captured by the complete cases only.

Indeed, as a first example, let us reconsider Scenario 1 with $n = 100$, but now adapting the conditional missingness probability in equation (3.31) to

$$\pi(y) = 1 - [1 + \exp\{1 - 0.009|y - 300|\}]^{-1}. \tag{4.3}$$

This gives the results shown in Table 4.5. In this situation, the imputation spline deviates from the true underlying function and therefore the imputation-based model selection does not perform as well as the weighted imputation (Figure 4.3).

In a second example uniform$[0, 10]$ x-values were generated. Given x, response y-values were generated from a normal distribution with mean $\mu_0(x) = 1.5(x-5)^2$ and variance $\sigma_0^2 = \exp(3.5)$. x-observations were then turned missing with conditional probability (see left bottom panel in Figure 4.4),

$$\pi(y, z) = 1 - [1 + \exp\{1 - 0.0009y^2\}]^{-1}. \tag{4.4}$$

The lower left panel of Figure 4.4 shows the augmented data based on an imputation using (4.2). It is clear from this figure that imputation-based model selection cannot be used when $E(x|y)$ is not well defined.

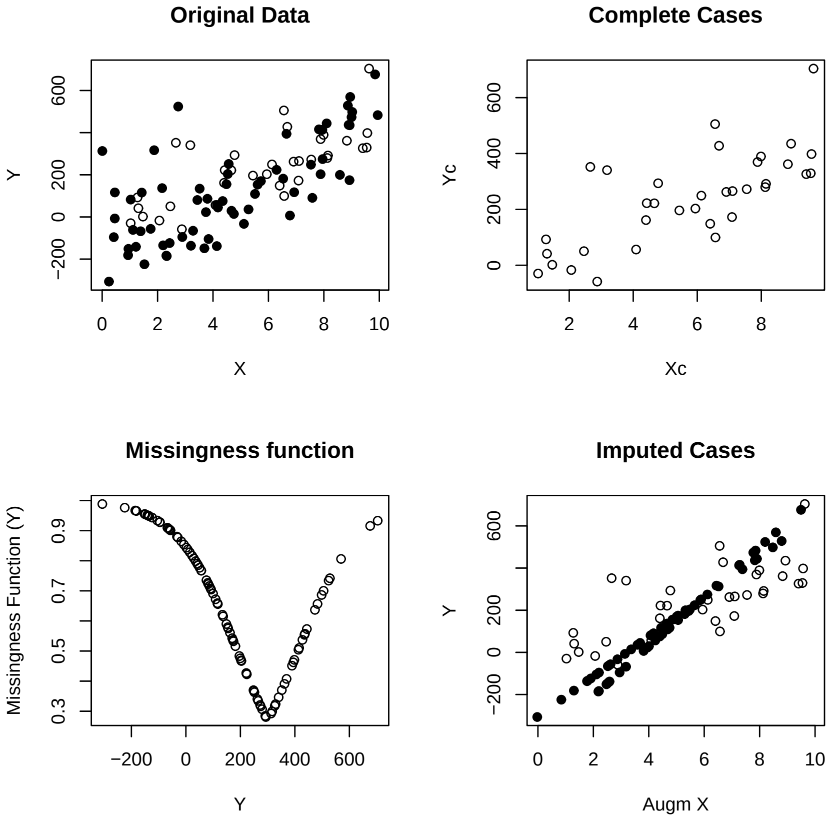


Figure 4.3: A first example: In the upper left panel, the observed data (white dots) and the missing data (black dots), in the upper right panel the complete cases. The lower left panel gives the missingness function and the lower right panel an augmented dataset using (4.2) (imputed data indicated by black dots).

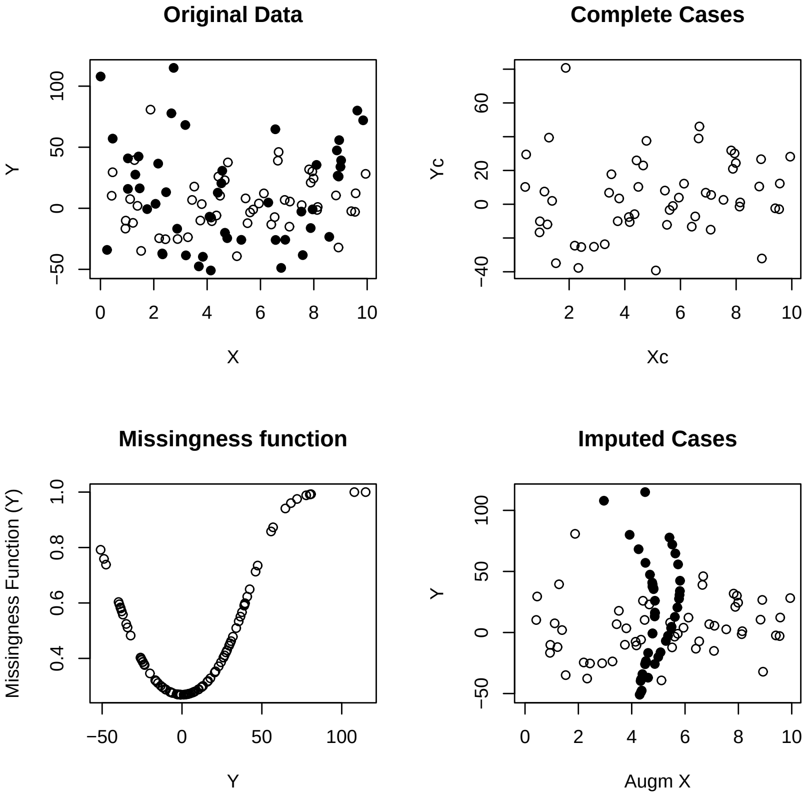


Figure 4.4: A second example: In the upper left panel, the observed data (white dots) and the missing data (black dots), in the upper right panel the complete cases. The lower left panel gives the missingness function and the lower right panel an augmented dataset using (4.2) (imputed data indicated by black dots).

4.4 Model Selection after Smoothing

In Section 4.2, it was shown that the success of the imputation-based model selection lies in the modelling of the complete cases. This raises an intuitive feeling that it is the smooth nature of the spline which is responsible for the improved performance of imputation-based methods. If the spline captures the true underlying curve, then the imputed data will also suggest a model close to the true underlying curve to be most adequate.

Therefore it is interesting to examine the effect of smoothing on model selection, even if missingness is not an issue. The algorithm proposed is the following. (1) Fit a penalized regression spline through the data (x, y), resulting in (x, y_s), where y_s are the predicted responses based on the spline fit; (2) use the AIC-criterion on (x, y_s) to select a model; (3) use this model on (x, y)-data. Figure 4.5 shows a graphical representation of this algorithm.

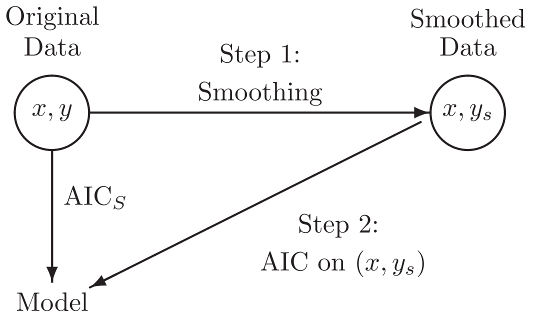


Figure 4.5: Illustration of the “model selection after smoothing”algorithm.

Let us consider some simulations to explore the performance of model selection after smoothing.

4.4.1 Scenario A

In a first scenario, uniform$[0, 10]$ x-values were generated, together with (independently) Bernoulli(0.5) z-values. Given x and z, response y-values were generated from a normal distribution with mean $\mu_0(x, z) = -3 + 3x + 5x^2$ and variance $\sigma_0^2 = \exp(5)$. (This is similar to Scenario 1 without missing values). 1000 different samples $\{(x_i, z_i, y_i), i = 1, \ldots, n\}$, with fixed design $\{x_i, z_i, i = 1 \ldots, n\}$ and $n = 50$ were generated. The candidate set of models is the same as for Scenario 1, i.e., all submodels of $\mu(x,z) = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 z + \beta_4 xz$.

Now, three smoothing methods were considered: (1) a gam model with penalized splines $Y \sim s(X) + Z + Z * s(X)$; (2) a gam model with penalized splines built from $Y \sim s(X) + Z + Z * s(X)$ according to Wood and Augustin (2002), and (3) a penalized regression spline $Y \sim s(X)$.

Wood and Augustin (2002) propose a 3-step ad hoc method to drop terms in a generalized additive model.

(1) Are the estimated degrees of freedom for the term close to their lower limit (e.g., 1 for a univariate smooth)?

(2) Does the confidence region for the smooth include zero everywhere?

(3) Does the GCV score for the model go down if the term is removed from the model?

If the answer to all 3 of these question is 'yes' then the term should be dropped. If the answer to 2 is 'no' then it probably should not be. Other cases require judgement.

Table 4.6: Scenario A: The numbers indicate how often a model has been selected, for the four strategies. The last column shows how often a correct model has been chosen, out of 1000.

	1	x	z	x, x^2	x, z	$x, z,$	$x, x^2,$	$x, x^2,$	correctly
						xz	z	z, xz	classified
Original Data	0	114	0	666	31	18	106	65	837
(1)	0	3	0	42	5	24	97	829	968
(2)	0	9	0	757	3	7	106	118	981
(3)	0	12	0	892	2	0	60	34	986

The results in Table 4.6 show that model selection after smoothing based on a gam model with penalized splines $Y \sim s(X) + Z + Z * s(X)$ results in the selection of the most complex model. In Figure 4.6, an arbitrary chosen sample, with the associated smoothed data (x, y_s) is shown. The smoothing invokes an apparent Z-effect, not only as a main effect but also as an interaction as can be seen from this figure. This was not a major problem for the imputation-based model selection, since there was still a considerable amount of complete cases indicating the Z-effect to be merely a nuisance.

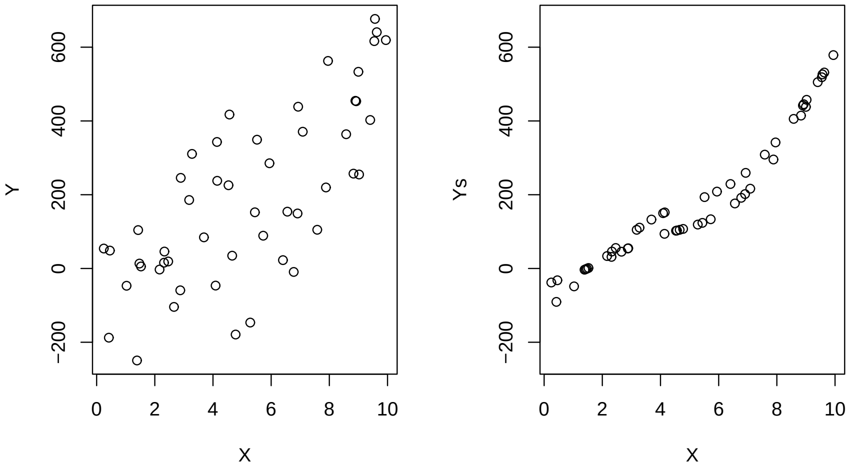


Figure 4.6: Method 1: An arbitrary sample for scenario A together with the associated smoothed data (x, y_s) based on smoothing method (1).

Table 4.7: Scenario A: model selection after smoothing: MASE and bias-variance decomposition.

	bias^2	var	MASE
Original Data	4.26	971.59	975.85
(1)	0.50	1080.38	1080.88
(2)	0.48	939.64	940.12
(3)	0.64	855.40	856.04

The use of the 3-step method of Wood and Augustin (2002) to determine the smoothing model improves model selection considerably. It selects a true model 10% more often than the AIC-criterion on the original data. As an other smoothing method, one can use a penalized regression spline of X only to obtain smoothed data (x, y_s). This results in additional improvement of 11% in selecting the true model and in 98.6% correct models to be chosen.

In Table 4.7, MASE-values together with bias-variance decomposition confirm the performance of the different methods. There is a large decrease in bias and applying methods (2) and (3) reduces the variability, while using method (1) results in an increased variability compared to model selection based on the original data. Model selection after smoothing, AIC$_S$ (methods (2) and (3)), seems to be able to reveal the true underlying function by reducing the noise in the data.

4.4.2 Scenario B

In a second scenario the same setting as Scenario A is considered. We will use the AIC-criterion and AIC$_S$ to select the power(s) of a fractional polynomial (Section 1.3.1) of degree 1(2) from a grid $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Note that the true model is a fractional polynomial of degree 2 with $p_1 = 1$ and $p_2 = 2$.

Table 4.8 shows an overview of the powers chosen by AIC- and AIC$_S$. Powers not chosen by any of the selection criteria were omitted from the table. All fractional polynomials selected using the AIC- and AIC$_S$-criterion are of degree 2. It can be seen that the generating model, which is contained in the set of the candidate models $(p_1, p_2) = (1, 2)$, was chosen 16 times by the AIC-criterion while it was chosen 297 times using AIC$_S$.

Table 4.8: Scenario B: Selected powers using the AIC-criterion (left) and AIC$_S$-criterion (right).

$p_1 \backslash p_2$	0.0	0.5	1.0	2.0	3.0
-2.0	1	4	28	134	87
-1.0	2	3	10	29	36
-0.5	2	13	14	11	39
0.0		11	14	10	36
0.5		29	42	15	31
1.0			49	*16*	31
2.0				12	8
3.0					118

$p_1 \backslash p_2$	0.5	1.0	2.0	3.0
-2.0			7	7
-1.0		4	9	8
-0.5	2	1	16	19
0.0	7	11	28	24
0.5	11	26	68	43
1.0		48	*297*	159
2.0			40	45
3.0				120

In Figure 4.7, a smoothed density plot of the ratio MASE(AIC)/MASE(AIC$_S$) is given. About 85% of the ratios are larger than 1 indicating an improved model choice when using the AIC$_S$-criterion.

The success of model selection based on presmoothing is in accordance with known results in related settings (Faraldo and Gonzalez Manteiga, 1987; Christóbal Christóbal *et al.*, 1987; Janssen *et al.*, 2001).

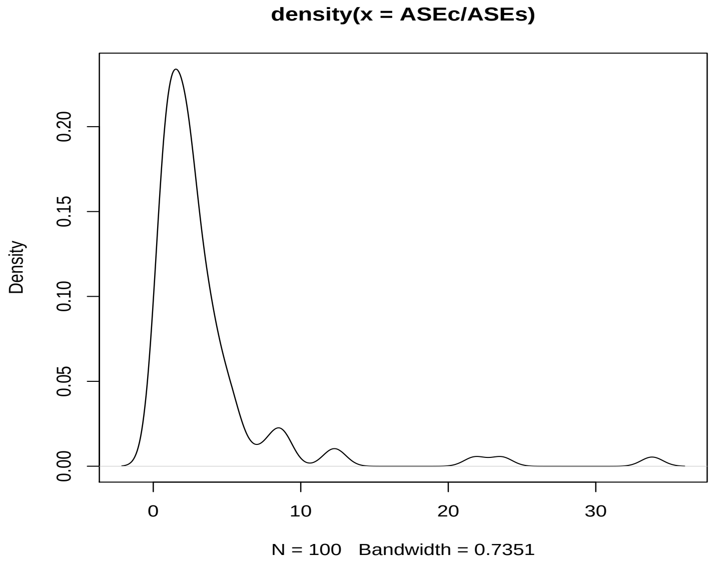


Figure 4.7: Scenario B: Smooth density plot of the fraction of MASE-values according to the model chosen on the original data (ASEc) and the model chosen on the smoothed data (ASEs).

4.5 Discussion

In this chapter, a small simulation study was performed to investigate the performance of imputation-based model selection, i.e., selecting a model based on augmented data. The performance of the method was compared to the performance of the weighted AIC-criterion presented in Chapter 3. The simulations in this chapter show that model selection based on non-parametric mean imputation is able to capture the true underlying model in fairly simple applications. If however information about the true model is scarce or if the true underlying function is not reversible, i.e., $E(x|y)$ is not well defined, it does not succeed in improving the model selection.

From the discussion on the mean average squared errors and bias-variance decomposition of the several methods it is clear that the price to pay using imputation-based model selection is the increase in variability, whereas the weighting methods perform better to that respect. Therefore possible further research lays in the combination of imputation and weighting, i.e., to select the model using an imputation-based AIC and to fit the model using inverse probability weights.

In exploring the improved behaviour of the imputation-based model selection, a concept of model selection after smoothing was briefly investigated by means of a small simulation study. In the preliminary smoothing step, model selection is necessary to avoid overly complex models to be chosen. Since the method reduces the noise in the data, it reveals the true underlying function and an improved model selection is the result. It remains a question whether model selection after smoothing is applicable in higher dimensions and generally applicable for different types of distributions.

Chapter 5

Cervix Cancer Screening in the Belgian Health Interview Survey 1997

5.1 Introduction

To outline an evidence-based health policy, one is often interested in the profiles of persons who are at risk to obtain certain diseases or who do not respond to prevention programs as, e.g., cervix cancer screening via smears. Statistical modelling can provide a tool to discover such profiles.

In the Belgian Health Interview Survey (HIS) of 1997, one of the questions investigated is in what respect the group of women, aged 25-64, not having a smear is different from the group of women that did have a smear taken in the past three years. For this purpose discrimination based on civil status, drug consumption, age, educational level and financial status was of interest.

Statistical modelling of surveys often has to deal with design issues as the sampling in the HIS was based on a combination of stratification, multistage sampling and clustering (Kish, 1995). Moreover it is not unlikely that one or more covariates for the variables of interest are missing, possibly due to numerous reasons or just by chance. In this dataset about 30% of the 2893 women had one or more missing covariates. Together with the design issues, statistical modelling has to deal with the missing values.

Table 5.1: Cervix Cancer Screening: Variables used in the candidate models.

Variable	Abbreviation	Coding
Screening Status	SC	binary
Civil Status	CS	nominal
Drug Consumption	DR	ordinal
Age	Age	continuous
Educational Level	EL	nominal
Financial Status	FS	nominal

In this chapter, different parametric and non-parametric modelling techniques will be applied to the Cervix Cancer Screening data. It is already in the first step, the model selection step, that one has to account for the design and the occurrence of missing values as pointed out in Chapters 3 and 4. In a second step the selected model is then used, while accounting for the design and missing values, to discover the profiles of persons who are at risk to obtain cervix cancer but who do not respond to prevention programs. We will assume data to be missing at random (see Section 1.2).

As a parametric technique, logistic regression will be used while as a non-parametric technique, the method of classification trees is described. Both logistic regression and classification trees have advantages but also limitations with respect to their application in the survey domain.

In a first section, focus is on model selection using logistic regression and in a second section, focus is on the final tree selection for a classification tree analysis, dealing with both design and missingness issues. We end with a discussion in Section 5.4.

5.2 Model Selection using Logistic Regression

Let us first apply the model selection procedures as introduced in Chapters 3 and 4 in a logistic regression setting.

Based on the variables given in Table 5.1, twelve different models as shown in Table 5.2 were considered. In Table 5.3, the AIC-criterion based on the complete cases (second column) is given together with five modified AIC-criteria. The models are ranked according to their AIC-criterion based on the complete cases. For all other

Table 5.2: Cervix Cancer Screening: Overview of the candidate models.

Model	Structure
(1)	SC$\sim$ Age+Age^2+log(DR)+CS
(2)	SC$\sim$ Age+Age^2+log(DR)+EL+DR*EL
(3)	SC$\sim$ Age+Age^2+DR+EL+EL*DR
(4)	SC$\sim$ Age+Age^2+log(DR)
(5)	SC$\sim$ Age+Age^2+log(DR)+log(Age)
(6)	SC$\sim$ Age+Age^2+DR
(7)	SC$\sim$ Age+Age^2+CS+CS*Age
(8)	SC$\sim$ CS+Age+EL+DR+Age*EL
(9)	SC$\sim$ Age+Age^2
(10)	SC$\sim$ CS+Age+EL+DR+Age*EL+DR*EL
(11)	SC$\sim$ FS+CS+DR+Age+EL
(12)	SC$\sim$ FS+CS+DR+AGe+Age*FS

columns, the three models with lowest AIC-values are indicated by their ranks.

In the third column, a first weighted version, AIC$_{W_1}$, takes into account the complex design. Individual weights, W_1, reflecting the stratification at provincial level and the differential selection probabilities within households were available. This results in a somewhat different ordering of the models. The best model now is the one with original rank 8.

Similarly, the fourth column shows the modified AIC-value, AIC$_{W_2}$, incorporating missing covariate data (assuming MAR). Because of the high dimensional covariate space, a classification tree with surrogate splitting (Section 5.3) was used to obtain estimates of the missingness probabilities and thus the weights W_2. This leads to only minor changes, as compared to the second column. The best model now is model 2.

In the fifth column both complications have been taken into account by multiplying both weights in AIC$_{W_1,W_2}$. Again the same models appear to be the best ones; model 8 showing up again, now as the third best model, while model 3 is having the lowest value.

To contrast these weighting methods to model selection after imputing missing covariate values, the AIC-criterion was applied to an imputed data set (AIC$_I$).

The imputation of missing values for the Cervix Cancer Screening data is not at all straightforward since the missing values are spread over several variables. The 'Random Forest'-methodology of Breiman (2001), introduced in Section 1.3.1, provides a flexible iterative imputation method. The algorithm starts by a rough imputation of missing values where for continuous variables, missing values are replaced with their median and for factor variables, missing values are replaced with the most frequent class breaking ties at random. Then a random forest is built with this augmented dataset. The proximity matrix from the random forest is used to update the imputation of the missing values. For continuous predictors, the imputed value is the weighted average of the non-missing observations, where the weights are the proximities. For categorical predictors, the imputed value is the category with the largest average proximity. This process is iterated 10 times.

A comparison between the imputation-based and weighted AIC-criteria shows that model (11), not chosen by the AIC and AIC_{W_1} on the complete cases, has the third lowest AIC_I-, AIC_{I,W_1}-value. Model (8), not chosen by the AIC_{W_2}-criterion which ignores the design but corrects for the missingness is now chosen by the AIC_I-criterion while it was chosen by the AIC_{W_1,W_2}-criterion and not by the AIC_{I,W_1}-criterion, both dealing with design and missing values. Model (2), the model with the lowest AIC_{W_2} and second-lowest AIC_{W_1,W_2}-value had the lowest AIC_I- and AIC_{I,W_1}-value.

From this data example, we see that weighted and imputation-based model selection opt for different models compared to model selection on the complete cases, but also compared to each other. There is a general tendency to opt for model (2) and (3) to be the better models. This example illustrates that differently weighted or imputation-based AIC-criteria can select different models as best ones. Since the choice of the final model or the set of final models used for e.g., model averaging is affected by missing data and by the design, we recommend in general the use of the weighted and imputation-based criteria (at least as a sensitivity tool).

Let us now focus on classification trees as a non-parametric alternative to logistic regression.

5.3 Model Selection using Classification Trees

The classification tree methodology was briefly introduced in Section 1.3.1. More details can be found in Breiman *et al.* (1984), Zhang and Singer (1999) and Hastie *et al.* (2001).

Table 5.3: Cervix Cancer Screening: The different (weighted/imputation-based) AIC-values and, between brackets, the rank of the three best models.

Model	AIC	AIC_{W_1}	AIC_{W_2}	AIC_{W_1,W_2}	AIC_I	AIC_{I,W_1}
(1)	1489.02*(1)*	975.31	2614.04*(2)*	1451.19	3626.56	3781.04
(2)	1489.81*(2)*	969.04	2606.71*(1)*	1441.53*(2)*	3556.07*(1)*	3706.76*(1)*
(3)	1490.70*(3)*	963.26*(2)*	2617.82*(3)*	1440.44*(1)*	3585.19	3713.98*(2)*
(4)	1492.39	965.66*(3)*	2625.36	1445.89	3654.47	3779.14
(5)	1494.10	967.60	2625.73	1447.96	3656.45	3781.30
(6)	1495.86	967.64	2632.11	1449.03	3660.89	3787.37
(7)	1496.19	984.37	2631.01	1461.50	3648.50	3806.66
(8)	1496.84	961.57*(1)*	2628.85	1441.77*(3)*	3556.28*(2)*	3737.76
(9)	1496.97	969.54	2636.47	1451.42	3665.03	3796.02
(10)	1502.31	967.35	2632.49	1447.34	3559.81	3747.50
(11)	1504.01	970.94	2648.48	1460.69	3559.48*(3)*	3733.11*(3)*
(12)	1516.75	980.92	2676.15	1477.45	3658.80	3839.60

One attractive feature of tree-based methods is the ease with which missing values can be handled. The appropriateness of these methods is however not straightforward (Ripley, 1996).

A first approach is prediction on complete observations suggested by Quinlan (1986). He suggests replacing missing values using the distribution within the class at that node when computing the expected value of a split. In his paper of 1993, Quinlan multiplies the impurity gain calculated on known observations by the proportion of missing values. This method has a major disadvantage when the number of complete observations in the node is quite small. Another disadvantage is that other available variables for this observation are neglected while they are possibly highly correlated with the missing one.

A second approach, Ripley (1996) discusses, is the missing together approach (MT). Suppose that we attempt to split a node by a variable and that the measurement for that variable is missing for a number of observations. The MT approach forces all of these subjects to the same daughter node. If it is a nominal variable with several levels, the missing value is regarded as an additional level, so the variable has one more level. On the other hand, when the variable has a natural order, two copies are made. If a component is missing, the component in the first copy will be set on plus infinity and the corresponding component in the second copy will be given the value minus infinity. In this way, replacing the variable by its two variants, results in two possible splits such that the observations with missing measurement are sent to the same daughter node. The variant that gives the best split is chosen. This is the key idea of the MT approach. The advantages of the MT approach are that it is very easy to implement and that a recursive partition algorithm that assumes no missing data can still be used without modification when the raw data contain missing values. Also the observations with missing information can easily be located in the tree structure. In contrast, both daughter nodes may contain some of these subjects by using surrogate splits instead. A major disadvantage of the MT approach is that imputation relies on the assumptions of simultaneous behaviour for subjects with a missing observation for the covariate of interest. Moreover, the most favourable split is chosen to be the best split, without considering the information in the other covariates. This can be circumvented by surrogate splits.

The third approach of surrogate splits is analogous to replacing a missing value in a linear model by regressing on the explanatory variable with a non-missing value most highly correlated with it. However it is more robust because of no model assumptions. The surrogate split approach attempts to utilize the information in the other predictors to assist in making the decision to send a observation to the left or the right daughter node. One looks for the predictor that is most 'similar' to the original predictor in classifying the observations. Similarity is measured by a measure of association. It is not unlikely that the predictor yielding the best surrogate split may also be missing. Then we have to look for the second best surrogate, and so on. In this way all available information is used. If surrogate splits are used, the user should take full advantage of them. In particular, a thorough examination of the best surrogate splits may reveal other important predictors that are absent from the final tree structure, and it may also provide alternative tree structures that in principle can have a lower misclassification cost than the final tree, because the final tree is selected in a stepwise manner and is not necessarily a local optimizer in any sense. This problem arises also in the case of selection procedures for parametric models.

A fourth possibility is to take missing as a further level of the attribute. This method allows multi-way splits which are not appealing because making some values missing can increase the gain in impurity. This can be circumvented by allowing only binary splits, or by penalizing multi-way splits.

As a conclusion one can say that in most approaches tree construction is based on the observations without any missing values. Where missing values are very frequent; as in large scale surveys, this may be unacceptable or even impossible.

The practical implementation of the previous methods, handling missing data, is not an issue. The appropriateness of the chosen approach however is. Especially the use of all available information by surrogate splits is appealing. Substantial improvements upon this method can be thought of, although the practical implementation can be a drawback. All of these ideas have merits and demerits, depending on how common missing values are and whether or not they are missing at random (Little and Rubin, 1987).

Prediction using either the complete observations, the missing together approach or multi-way splits has several disadvantages compared to the use of surrogate splits. Therefore we will not discuss them. As an illustration we focus on four classification tree analyses: (1) using complete cases only, (2) using complete cases with inverse probability weighting, (3) using augmented data, i.e., missing values are imputed in the original dataset, and (4) using available data with surrogate splits. These four methods were applied with and without design weights.

Let us describe the pruning process, using cross-validation, for the analysis without design weights using the complete cases only. Calculation of the cross-validation relative error was based on subsets of size 10. In Figure 5.1, the cross-validation relative error is shown as a function of the size and cost complexity parameter (cp). In Table 5.4, the cross-validation relative error (xerror) is shown together with the cost complexity (CP), the number of splits (nsplit), the relative error (rel error) and the cross-validation relative error standard deviation (xstd). The size of the tree is defined as the number of terminal nodes. From this table we can see that a minimum is obtained for size 3 (nsplit=2). The 1 SE-error rule, i.e., selecting the smallest tree of which the cross-validation error is within one standard error from the minimal cross-validation error was not applied because of the simplicity of the resulting tree. Similarly, the selection of the final tree for the other methods occurred with and without design weights. In Figure 5.2 and 5.3, the resulting trees are shown without and with design weights, respectively.

Looking at Figure 5.2, small trees were selected by all four methods. There is a small difference in the final trees using the complete cases and inverse probability weighted complete cases. They both have age as a primary split, while based on the complete cases a second split based on educational level is found. There is no difference in the final tree using the augmented data and the available cases with surrogate splits.

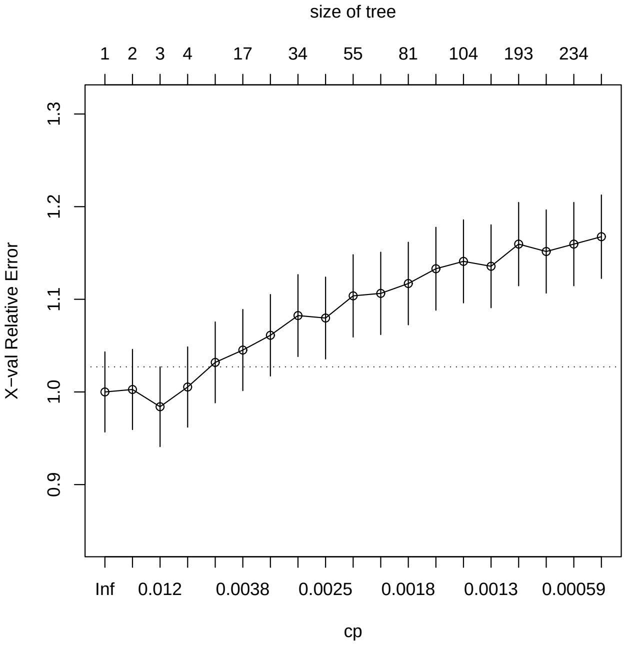


Figure 5.1: Cervix Cancer Screening: cross-validation relative error as a function of the size (upper axis) and cost complexity (lower axis) using complete cases only without design weights.

Table 5.4: Cervix Cancer Screening: cross-validation relative error using complete cases only without design weights.

CP	nsplit	rel error	xerror	xstd
0.03723404	0	1.00000	1.00000	0.043247
0.01861702	1	0.96277	1.00266	0.043280
0.00797872	2	0.94415	*0.98404*	0.043045
0.00531915	3	0.93617	1.00532	0.043313
0.00398936	12	0.88564	1.03191	0.043635
0.00354610	16	0.86968	1.04521	0.043790
0.00341945	19	0.85904	1.06117	0.043971
0.00265957	33	0.80319	1.08245	0.044205
0.00227964	47	0.76596	1.07979	0.044176
0.00199468	54	0.75000	1.10372	0.044429
0.00182846	58	0.74202	1.10638	0.044457
0.00177305	80	0.69947	1.11702	0.044565
0.00159574	92	0.67819	1.13298	0.044723
0.00132979	103	0.65957	1.14096	0.044800
0.00122750	170	0.56915	1.13564	0.044748
0.00088652	192	0.53723	1.15957	0.044975
0.00066489	225	0.50798	1.15160	0.044900
0.00053191	233	0.50266	1.15957	0.044975
0.00000000	238	0.50000	1.16755	0.045048

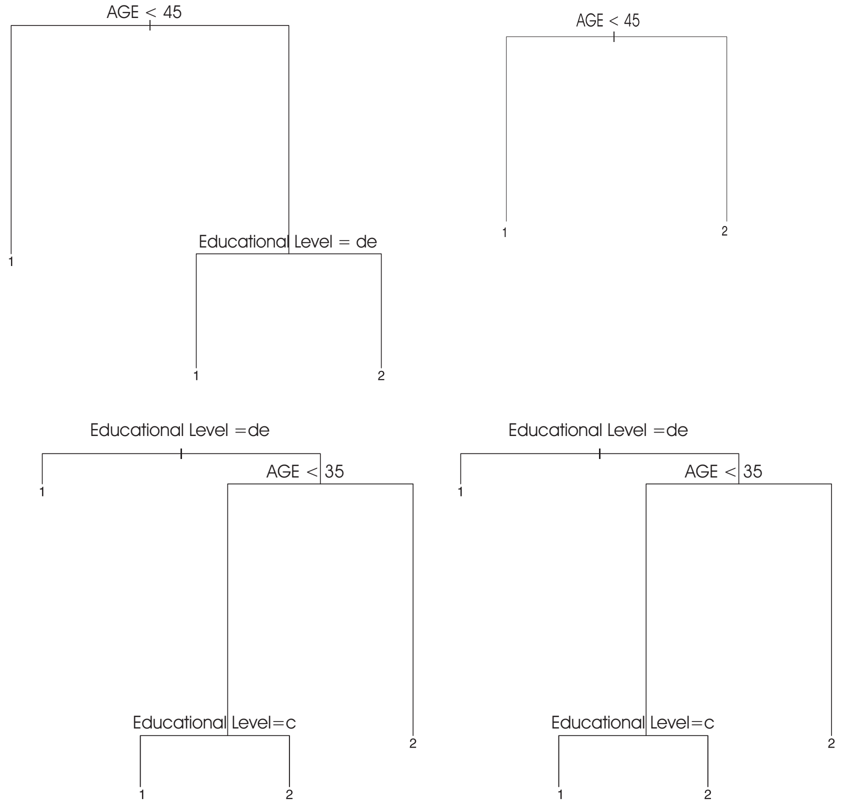


Figure 5.2: Cervix Cancer Screening: final trees without design weights based on the complete cases (upper left), complete cases with inverse probability weights (upper right), augmented cases (lower left) and available cases with surrogate splits (lower right).

However, there is a difference between the latter two and the first two trees, i.e., the latter two methods result in a tree with primary split educational level. A secondary split is based on age while educational level shows up again as a last split.

Comparing these results with the final trees using design weights, we see that more complex trees are chosen for those trees resulting from using the augmented data and the available cases with surrogate splits, while the use of complete cases and inverse probability weighted complete cases result in a tree of size 1. The final trees based on the augmented data and available data using surrogate splits show some differences from the seventh layer onwards. The basis of the trees, using design weights are the same as when ignoring the design but the complexity is different.

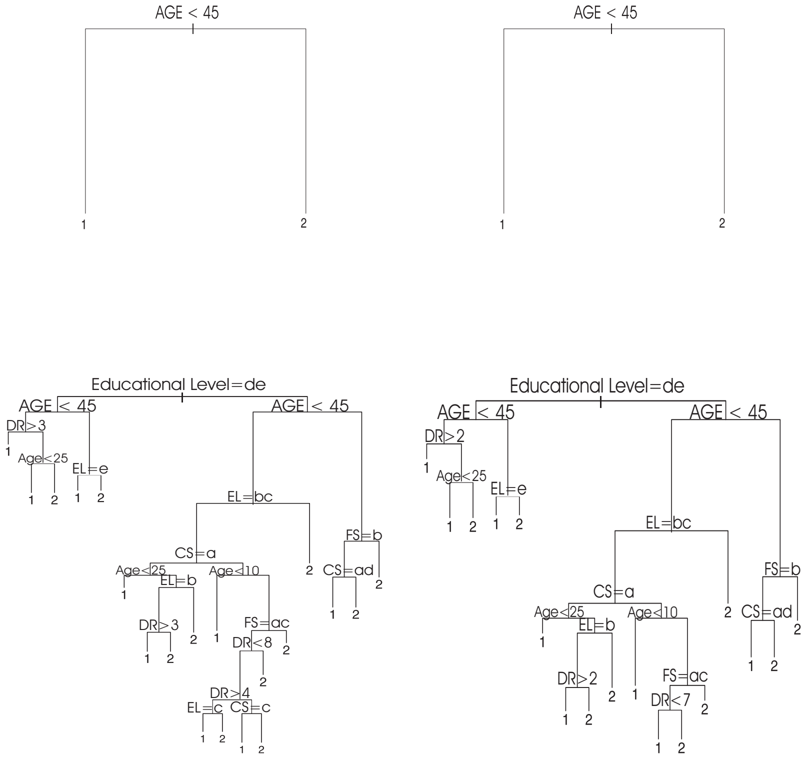


Figure 5.3: Cervix Cancer Screening: final trees with design weights based on the complete cases (upper left), complete cases with inverse probability weights (upper right), augmented cases (lower left) and available cases with surrogate splits (lower right).

To validate the final trees, we calculated the cross-validation misclassification error for each method. First, ignoring the design, all methods give a similar misclassification error of about 29%. Secondly, correcting for the design, we find a misclassification error of 27% for both the complete cases and inverse probability weighted complete cases, while 31% was found for both the augmented data and available cases with surrogate splits. This shows that more complex (non-nested) trees are not necessarily equivalent with an improved classification.

Looking at tree-based methods as variable reduction methods, variables age and education level are the only important variables when ignoring the design while the other variables show up when using design weights. We refer to Hens *et al.* (2002) for a more thorough discussion of both a classification tree analysis and logistic regression analysis of cervix cancer screening in the Belgian HIS of 1997.

5.4 Discussion

The results of the analyses on the Cervix Cancer Screening data, presented in this chapter, show that dealing with missing data highly depends on the method used. Feelders (2000) showed that using imputation outperforms the use of surrogate splits. In our analysis, they seem to perform equally well. It is appealing to use imputation because of the preliminary imputation step which is not a part of the model building process as opposed to the use of surrogate splits. Carpenter and Kenward (2005) contrast the use of inverse probability weighting with multiple imputation to handle data missing at random in a parametric setting. A similar comparison for non-parametric techniques such as classification trees has, to our knowledge, not been done before.

Dealing with the design affects the final model chosen and is therefore not to be ignored. A sensitivity analysis is recommended to be a part of the data analysis.

The use of a classification tree as a prediction and classification tool is known to be highly unstable. Ensemble methods as bagging, boosting and random forests have been developed to obtain more accurate predictions (see Section 1.3.1). Instead of applying a single classification tree, one could opt for the use of one of such ensemble methods. However, in practice, one is often interested in the effects of one or more covariates on the response variable and it is often difficult to derive these from an ensemble method. Therefore, we advise to use these methods to validate the predictions made by a single classifier.

The motivation to use a parametric or non-parametric modelling technique depends on the aim of the analysis, the underlying assumptions and other sometimes subjective criteria. Ye (1998) introduced the concept of generalized degrees of freedom which can be used to compare different model techniques as, e.g., classical parametric models and tree-based methods and thus could be used as a basis for further research in the field of survey data.

Chapter 6

Kernel Weighted Influence Measures

6.1 Introduction

When dealing with longitudinal data, it is not unlikely for measurements to drop out. In Section 1.2 several modelling techniques to represent dropout under different missingness mechanisms have been introduced. These models, trying to represent a non-random dropout mechanism, rely on strong and untestable assumptions. Therefore, there is a clear need for a sensitivity analysis. A sensitivity analysis can be defined as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools. Examples include, Crouchley and Ganjali (2002) who used a multivariate generalization of the Heckman model as an alternative to selection models on the Mastitis data; Baker *et al.* (2003) who use different missing data models and compared the resulting goodness of fit statistics and parameter estimates in a selection modelling framework; and Daniels and Hogan (2000) who performed a sensitivity analysis for pattern mixture models under informative dropout.

In this chapter, we will focus on the selection model proposed by Diggle and Kenward (1994). Using a selection model, not only the assumed distributional form can be misspecified but also the presence of influential observations can be of great importance to select an appropriate model. A well known method to investigate the influence of individual cases is case deletion (Cook and Weisberg, 1982; Lawrance, 1995; Zhu *et al.*, 2001; Cavanaugh and Oleson, 2001). This results in the global influence approach. A quite different approach is to perturb the model a bit and study the stability of the model, as is done by Lesaffre and Verbeke (1998) as an application of the local influence approach introduced by Cook (1986). In Thijs *et al.* (2000), Verbeke *et al.* (2001), Molenberghs *et al.* (2001, 2003) and Jansen *et al.* (2003), this method was used to investigate the influence of non-random missingness as part of a sensitivity analysis in the selection modelling framework. A thorough discussion can also be found in Verbeke and Molenberghs (2000) and Jansen *et al.* (2005).

One of the datasets discussed in the literature is the Mastitis dataset. These data were initially used by Kenward (1998) for an informal sensitivity analysis. They were analyzed extensively with the local influence approach by Molenberghs *et al.* (2001). The influence analyses on the Mastitis data and other datasets, make it clear that the allocation of the possibly different sources of influence is still a burden. The related question on when to call a case influential (i.e., well defined cut-off values) is still an open problem. In view of obtaining new insight in this matter, we introduce kernel weighted influence measures. We will illustrate the techniques on the Mastitis dataset throughout this chapter.

Our proposal is an extension of the two approaches of global and local influence. Instead of looking at cases, we are interested in looking at the influence of types of observations. To know why an observation is influential, one has to consider the characteristics of that observation. So, instead of wondering why this particular observation is influential, the question becomes which characteristics of this observation makes this type of observation influential. Therefore we will look at observations in the neighbourhood of a case. This new exploratory and graphical tool supplements many other tools for sensitivity analysis and can contribute in obtaining further insights in the mechanisms generating missing data (Hens *et al.*, 2005b; Jansen *et al.*, 2005).

Kenward (1998) introduced a statistical model to analyze the Mastitis data, a model that fits in the selection modelling framework. We briefly describe the selection model of Diggle and Kenward and the global and local influence in Section 6.3. The development and motivation of the kernel weighted influence measures is given in Section 6.4. This approach will be extended to a grid analysis (Section 6.5) and a small simulation study is carried out (Section 6.6). In this chapter, we restrict attention to the case of two measurements for each subject. How the method can be extended to the general case of more than two time points is briefly sketched in Section 6.7.

6.2 A Descriptive View on the Mastitis Data

Consider the Mastitis data as introduced in Section 1.5.3. Looking at the different profiles in Figure 1.2, cows #4, #5 and #66 have a large increase in milk yield compared with the other cows. Cow #89 appears to have the largest decrement. Next to cow #66, cows #54, #69 and #53 are high yielding cows in both consecutive years.

Because some cows have a large reduction in milk yield and others exhibit a substantial increase, it is useful to look at the increments, i.e., the difference between the milk yield in the second year and the first year. In Figure 1.3, a scatterplot of the original data is given together with a plot of the increments against the first measurement.

If we take a closer look at the scatterplots in Figure 1.3, we can see that the cows mentioned above, are located at the border of the data region. Are these specific cows having a large influence on a statistical analysis and are there any other cases with high influence? Getting more insights in such questions is the purpose of a sensitivity analysis. Special attention goes to cows #54 and #69, having almost identical measurements (both very high). It is known that, in classical regression models without missingness, most influence measures are not able to detect such cases, because they mask each other (Ryan, 1997). One of the main objectives is to study to which extent the influence measures introduced by Molenberghs *et al.* (2001) suffer from the same deficiency; and to propose modified versions of these influence measures which deal with it. Another objective is to extend the methodology to measure the influence of 'types' of observations, not really included in the sample but represented by clusters of neighbouring observations.

6.3 Influence Measures

This section summarizes parametric approaches to sensitivity analysis within the framework of selection models.

6.3.1 A Selection Model for Non-random Dropout

Let us assume that for subject $i = 1, \cdots, N$, a sequence of responses Y_{ij} is measured at two occasions $j = 1, 2$. Let R_i be a missingness indicator and assume that Y_{i1} is always observed. Then, $R_i = 1$ if Y_{i2} is observed and $R_i = 0$ if Y_{i2} is missing. The measurement part of the model of Diggle and Kenward (1994), applied to the

Table 6.1: Parameter estimates (and standard errors) of the selection model fitted on the Mastitis dataset.

Effect	Parameter	Random Dropout	Non-random Dropout
Measurement Model			
Intercept	μ	5.77(0.09)	5.77(0.09)
Time effect	Δ	0.72(0.11)	0.33(0.14)
First variance	σ_1^2	0.87(0.12)	0.87(0.12)
Second variance	σ_2^2	1.30(0.20)	1.61(0.29)
Correlation	ρ	0.58(0.07)	0.48(0.09)
Dropout Model			
Intercept	ψ_0	-2.65(1.45)	0.37(2.33)
First measurement	ψ_1	0.27(0.25)	2.25(0.77)
Second measurement	ψ_2	0	-2.54(0.83)
-2 loglikelihood		280.02	274.91

Mastitis data, is characterized by, for $i = 1, \ldots, N$,

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ \mu+\Delta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right], \tag{6.1}$$

where the covariance matrix expresses a serial correlation between the measurements at the two occasions. The missingness process is described by

$$\text{logit}[Pr(R_i = 0|y_{i1}, y_{i2})] = \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}, \tag{6.2}$$

where $Pr(R_i = 0|y_{i1}, y_{i2})$ is the probability for the i-th subject to have a missing measurement at the second occasion, under the posited model. If ψ_2 differs from zero, the missingness process is non-random as in the terminology of Rubin (1987).

The fit of this model on the Mastitis data based on the assumption that the dropout process is MAR on the one hand and MNAR on the other hand (Diggle and Kenward, 1994) is summarized in Table 6.1.

Testing $H_0 : \psi_2 = 0$ by means of a likelihood ratio test gives the value $G^2 = 5.11$, indicating some evidence against the MAR assumption. The appropriateness of using the $\chi^2(1)$-distribution in this situation is postponed to Chapter 7. The high value of the test statistic does not at all mean that there are observations in the dataset which are missing not at random. It is also possible that this high value is due to misspecification of the distribution or even just the missingness process. An important question is then, whether some particular subjects are responsible for this behaviour. Cook and Weisberg (1982) introduced a case deletion approach to investigate the influence of subjects. From their approach, several other methods were developed. The next two sections discuss global and local influence measures as applied on the Mastitis data.

6.3.2 Global Influence

Let us introduce a weighted loglikelihood

$$l(\boldsymbol{\gamma}; \boldsymbol{w}) = \sum_{j=1}^{N} w_j l_j(\boldsymbol{\gamma}), \tag{6.3}$$

where $\boldsymbol{w} = (w_1, \ldots, w_N)$ is a vector of subject specific weights such that $\sum_{i=1}^{N} w_i = N$ (reflecting an effective total sample of size N) and $l_j(\boldsymbol{\gamma})$ represents the loglikelihood contribution of the j-th subject with $\boldsymbol{\gamma}$ the parameter vector containing all unknown parameters (from measurement and dropout model). Denote $\hat{\boldsymbol{\gamma}}$ the maximum likelihood (ML) estimator of the unweighted likelihood, corresponding to the weight vector $\mathbf{1} = (1, \ldots, 1)$, and $\hat{\boldsymbol{\gamma}}_w$ the ML estimator based on the weighted likelihood (6.3).

Define

$$CD(\boldsymbol{w}) = 2\{l(\hat{\boldsymbol{\gamma}}; \mathbf{1}) - l(\hat{\boldsymbol{\gamma}}_w; \mathbf{1})\}, \tag{6.4}$$

as a measure for the distance between the ML estimator $\hat{\boldsymbol{\gamma}}$ and the weighted ML estimator $\hat{\boldsymbol{\gamma}}_w$. The global influence measure (Molenberghs *et al.*, 2001)

$$CD_i = CD(\boldsymbol{w}_{(-i)}), \tag{6.5}$$

compares $\hat{\boldsymbol{\gamma}}$ to $\hat{\boldsymbol{\gamma}}_{(-i)}$; the latter is the weighted ML estimator using weight vector $\boldsymbol{w}_{(-i)} = N/(N-1) \times (1, \ldots, 1, 0, 1, \ldots, 1)$ with the 0 at the i-th entry.

A global influence analysis on the Mastitis data, leads to influential cows #4, #5, #66 and #89, as shown in Figure 6.1. This is not surprising since cows #4, #5 and #66 have the largest increases in milk yield from year 1 to year 2 and cow #89 has the largest decrease in milk yield. Their behaviour is thus different from the other cows. A full discussion is given by Molenberghs *et al.* (2001). But apparently cows #54 and #69 are not suggested to be influential by the global influence measure CD_i.

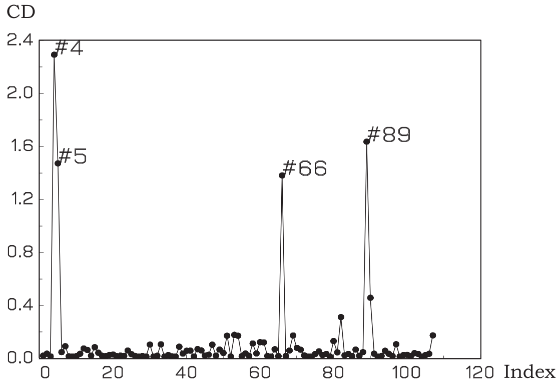


Figure 6.1: Influential subjects of the Mastitis data based on the global influence measure.

A main disadvantage of global influence measures is that the influence that can be ascribed to a specific case is hard to assess, since by deleting a subject all sources of influence are lumped together, with little hope to disentangle them. This was the main motivation to look at local influence methods.

6.3.3 Local Influence

The principle is to investigate how the results of an analysis are changed under infinitesimal perturbations of the model. Based on knowledge about Mastitis, the increments appear to be important. A thorough motivation is given in Molenberghs *et al.* (2001). Therefore a missingness process of the following form is considered.

$$\text{logit}[P(R_i = 0|Y_{i1}, Y_{i2})] = \psi_0 + \psi_1(Y_{i1} + Y_{i2}) + \phi_i(Y_{i2} - Y_{i1}), \tag{6.6}$$

where ϕ_i is a subject-specific weight, allowing the investigator to determine the local influence of one subject on the dropout model.

Let $l_i(\boldsymbol{\gamma}|\phi_i)$ denote the i-th loglikelihood contribution of the i-th subject, associated with missingness process (6.6) and let $l(\boldsymbol{\gamma}|\boldsymbol{\phi}) = \sum_{i=1}^{N} l_i(\boldsymbol{\gamma}|\phi_i)$ denote the total loglikelihood with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$. The vector $\boldsymbol{\phi}_0 = (0, \ldots, 0)$ corresponds to a MAR process. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\phi}}$, the ML estimator based on $l(\boldsymbol{\gamma}|\boldsymbol{\phi})$ and $\widehat{\boldsymbol{\gamma}}_0$, the ML estimator based on $l(\boldsymbol{\gamma}|\boldsymbol{\phi}_0)$, by the so-called likelihood displacement, defined by

$$LD(\boldsymbol{\phi}) = 2\{l(\hat{\boldsymbol{\gamma}}_0|\boldsymbol{\phi}_0) - l(\hat{\boldsymbol{\gamma}}_{\boldsymbol{\phi}}|\boldsymbol{\phi}_0)\}. \tag{6.7}$$

This approach takes into account the variability of $\widehat{\boldsymbol{\gamma}}$. The geometric surface formed by the values of the graph $\xi(\boldsymbol{\phi}) = (\boldsymbol{\phi}, LD(\boldsymbol{\phi}))$ gives the essential information about the influence of the perturbation scheme. Because of graphical limitations in dimensions higher than 2, Cook (1986) proposed to look at the normal curvatures $C_{\boldsymbol{h}}$ of $\xi(\boldsymbol{\phi})$ at $\boldsymbol{\phi}_0$, in the direction of some N-dimensional vector $\boldsymbol{h}$ of unit length.

Cook (1986) has shown that $C_{\boldsymbol{h}}$ can easily be calculated by

$$C_{\boldsymbol{h}} \;=\; 2 \left| \, \boldsymbol{h}^T \; \boldsymbol{\Delta}^T \; \ddot{\boldsymbol{L}}^{-1} \; \boldsymbol{\Delta} \; \boldsymbol{h} \, \right|, \tag{6.8}$$

where $\boldsymbol{\Delta}$ is a $(s \times N)$ matrix with $\boldsymbol{\Delta}_i$ as its i-th column, $\boldsymbol{\Delta}_i$ being the s dimensional vector defined by

$$\boldsymbol{\Delta}_i = \left. \frac{\partial^2 l_i(\boldsymbol{\gamma}|\phi_i)}{\partial \phi_i \partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}_0, \phi_i=0}. \tag{6.9}$$

Further, $\ddot{\boldsymbol{L}}$ denotes the $(s \times s)$ matrix of second order derivatives of $l(\boldsymbol{\gamma}|\phi_0)$ with respect to $\boldsymbol{\gamma}$, also evaluated at $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}_0$. One evident choice for $\boldsymbol{h}$ is the vector $\boldsymbol{h}_i$ containing 1 in the i-th position and 0 elsewhere, corresponding to a perturbation from the MAR model for the i-th subject in (6.7) only. The measure $C_{\boldsymbol{h}_i}$ reflects the influence of allowing the i-th subject to drop out non-randomly, while the others can only drop out at random.

Calculating the local influences of the cows in the Mastitis data, cows #4, #5 and #66 appear to be influential (see Figure 6.2). This is in agreement with the global influence analysis. Because the local influence looks at perturbations of the MNAR-parameter, while the global influence is based on case deletion, this was not to be expected a priori (Molenberghs *et al.*, 2001). Kenward (1998) observed that cows #4 and #5, which show up in both analyses, are substantially different from the other cows by their large increment.

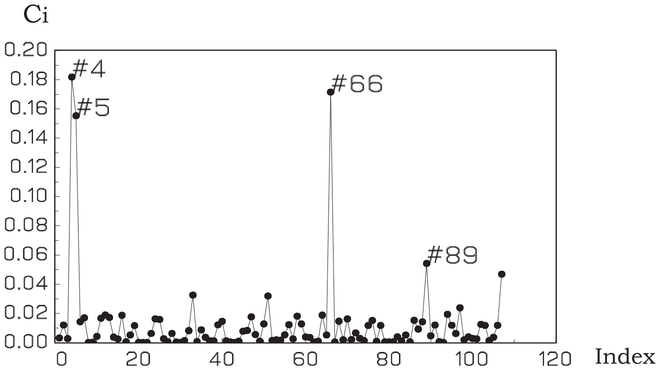


Figure 6.2: Influential subjects of the Mastitis data based on the local influence measure.

If the dropout probabilities are considered, then cow #66 seems to have a large dropout probability compared with the other cows. Therefore, a perturbation of the MNAR-parameter will reflect this.

From both the global and local influence analysis it is clear that the location of the data is of great interest. Therefore, a method to analyze sensitivity of types of observations might lead to a better comprehension of the influence measures and sensitivity analyses.

6.4 Kernel Weighted Influence Measures

The basic idea is to study the influence of types of observations, which are defined by neighbourhoods centered at the observations (y_{1i}, y_{2i}, r_i). Here techniques from non-parametric smoothing methods can be used. Inspired by the well-known kernel estimators for density and regression estimation (Wand and Jones 1995, Section 1.3.1), we propose the use of a kernel based choice for the weight vector $\boldsymbol{w}$ in the global measure (6.4) and for the direction vector $\boldsymbol{h}$ in the local measure (6.8).

6.4.1 Kernel Weighted Global Influence

Influence measures such as the global influence and local influence approach are essentially based on the influence of single cases. The global measure CD_i quantifies the change in the parameter estimates when including or excluding the i-th case; the local measure $C_{\boldsymbol{h}_i}$ reflects the influence of allowing the i-th subject to drop out non-randomly. We extend these two approaches by considering a neighbourhood $N(i)$ of (y_{1i}, y_{2i}, r_i) defined by kernel functions (see e.g. Wand and Jones, 1995). Let K be a density function and g_1 and g_2 two so-called bandwidth parameters.

The neighbourhood $N(i)$ of observation i is characterized by the values of the product (or multiplicative) kernel

$$K(\frac{y_{1j} - y_{1i}}{g_1})\{K(\frac{y_{2j} - y_{2i}}{g_2})\}^{r_i} I(r_j = r_i), \tag{6.10}$$

for $j = 1, ..., N$, where $I(r_j = r_i)$ equals 1 if $r_j = r_i$ and 0 otherwise. The first two factors in the definition of (6.10) are typical kernels for continuous variables and the indicator function can be considered as a kernel for a categorical variable. Taking the product of one-dimensional kernels is a typical simple way to characterize multivariate observations in a the neighbourhood of a certain observation (see e.g. Wand and Jones 1995). The exponent of the second factor expresses the possible missingness of the second measurement y_2.

First consider the case observation i is complete ($r_i = 1$). Complete observations ($r_j = r_i = 1$) with values close to $K^2(0)$ (the upper limit) are close neighbours of the i-th observation; observations at a further distance have values for (6.10) close to 0 (the lower limit). Incomplete observations ($r_j = 0$) get value 0. In case observation i is incomplete ($r_i = 0$), the interpretation is essentially the same focusing on the first factor, now having a maximal value $K(0)$ for the closest neighbours (identical observations).

This leads to the following definition: the kernel based weight vector $\boldsymbol{w}_{(-N(i))}$ is the vector of length N with elements, for $j = 1, \dots, N$,

$$\begin{aligned}(\boldsymbol{w}_{(-N(i))})_j &= [K(0)\{K(0)\}^{r_i} \\ &\quad -K(\frac{y_{1j}-y_{1i}}{g_1})\{K(\frac{y_{2j}-y_{2i}}{g_2})\}^{r_i} I(r_j = r_i)\Big] /D. \end{aligned} \qquad (6.11)$$

The denominator D is a normalization constant assuring that

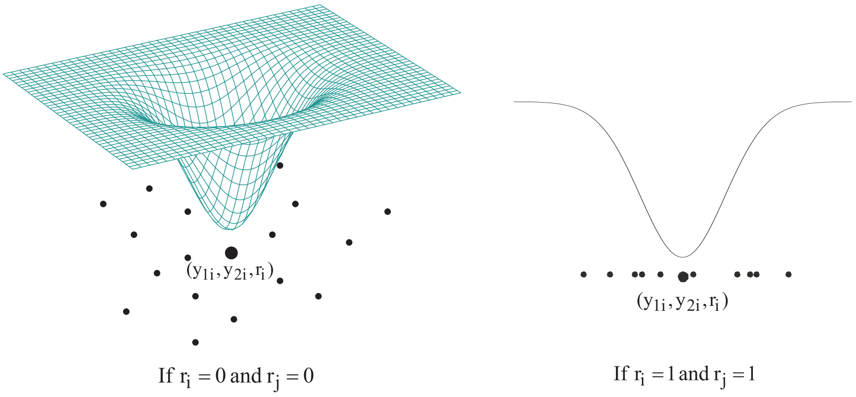
$$\sum_{i=1}^{N} (\boldsymbol{w}_{(-N(i))})_j = N.$$

Define the kernel weighted global influence measure of the i-th observation (y_{1i}, y_{2i}, r_i) as

$$CD_{N(i)} = CD(\boldsymbol{w}_{(-N(i))}). \qquad (6.12)$$

It measures the discrepancy between the ML-parameter estimator including or excluding the neighbourhood $N(i)$ as indicated by the weight vector $\boldsymbol{w}_{(-N(i))}$. The weights are shown graphically in Figure 6.3. For bandwidths g_1 and g_2 tending to 0 and in case all observations are different (no ties), the weight vector $\boldsymbol{w}_{(-N(i))}$ converges to $\boldsymbol{w}_{(-i)}$. In case there are ties (or very close neighbours), the method contrasts the parameter estimates including or excluding these particular ties (or very close neighbours) for bandwidths tending to 0 (or very small). So the kernel weighted influence measure (6.12) is able to allocate groups of influential cases with similar outcomes, thus avoiding the problem of *masking*. Masking refers to the existence of a close cluster of influential data points such that deleting a single point will cause little effect (see e.g. Ryan 1997).

As the method is intended as an exploratory and graphical tool, the influence of neighbourhoods $N(i)$, characterizing a certain type of observation, is explored by considering a series of bandwidth values. But, from our experience, the bandwidth needs to be adjusted to the data density at the observation i under consideration. We suggest to use a density adaptive bandwidth $g = g_1 = g_2$. Let (y_{1i}, y_{2i}, r_i) be



If $r_i = 0$ and $r_j = 0$

If $r_i = 1$ and $r_j = 1$

Figure 6.3: Shape of the weights. On the left hand side the weights are shown for the situation $r_i = r_j = 1$ (completers), while on the right hand side the weights are shown for the situation $r_i = r_j = 0$ (incompleters).

the observation of interest. If $r_i = 1$, the bandwidth g is taken as

$$g(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j, r_j = 0} K(\frac{y_{1j} - y_{1i}}{\tilde{g}_1}) K(\frac{y_{2j} - y_{2i}}{\tilde{g}_2})}. \tag{6.13}$$

If $r_i = 0$, the bandwidth is taken to be

$$g(y_{1i}, y_{2i}, r_i) = \frac{CK^2(0)}{\sum_{j, r_j = 1} K(\frac{y_{1j} - y_{1i}}{\tilde{g}_1}) K(0)}, \tag{6.14}$$

where C is a constant and $\tilde{g}_1$ and $\tilde{g}_2$ are two initially chosen bandwidths. Throughout the chapter we used the standard normal density function as the kernel function K.

A kernel weighted global influence analysis with initial bandwidths $\tilde{g}_1 = \tilde{g}_2 = 0.2$ and $\tilde{g}_1 = \tilde{g}_2 = 1.5$ on the Mastitis data leads to Figures 6.4 and 6.5, respectively. For both bandwidths the types of cows corresponding to #4, #5, #54, #66, #69 and #89 seem to have a large influence. From Figure 1.3 it is clear that these cows are the ones, lying at the border of the region. Cows #54 and #69 were not found with the global influence. The profiles of these two cows are practically the same (Figure 1.2). The global influence did not identify these cows as influential due to masking. The ML estimators $\hat{\gamma}_{(\mathbf{-54})}$, $\hat{\gamma}_{(\mathbf{-69})}$ as defined in Section 6.3.2 do not differ very much from $\hat{\gamma}$. In the kernel weighted global influence both cows get low weight and therefore, the shift in likelihood is detected. If we have a closer look at Figure 6.5, a second group of observations seems to be influential. This group corresponds to types of observations #7, #47 and #58, which are incomplete observations. These incomplete observations have the three highest y_1-values among the incompleters

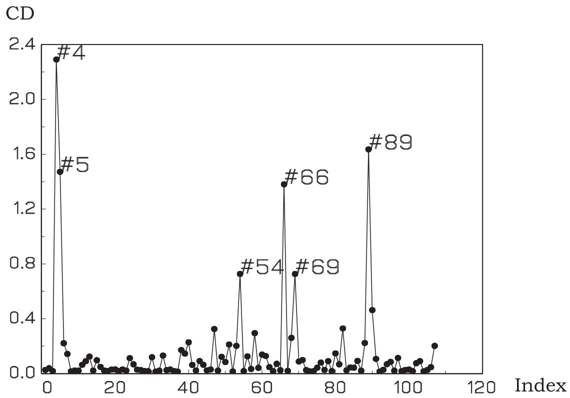


Figure 6.4: Influential subjects of the Mastitis data for the kernel weighted global influence with initial bandwidths $\tilde{g}_1 = \tilde{g}_2 = 0.2$.

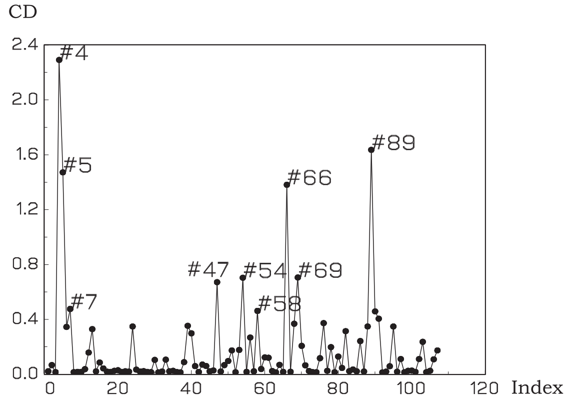


Figure 6.5: Influential subjects of the Mastitis data for the kernel weighted global influence with initial bandwidths $\tilde{g}_1 = \tilde{g}_2 = 1.5$.

(Figure 1.2) and thus can also be seen as outlying observations with substantial influence. A comparison of Figures 6.4 and 6.5 in this respect clearly shows the role of the bandwidth as a tuning parameter in an explorative sensitivity analysis. Both figures show the same influential complete cases but Figure 6.5 with the larger bandwidth adds to these some incomplete influential cases.

6.4.2 Kernel Weighted Local Influence

The local influence approach can be extended by looking at the direction determined by the neighbourhood $N(i)$. First, note that from the discussion in Section 3 it follows that $\boldsymbol{h}_i = (1 - \boldsymbol{w}_{(-i)})/D$ where D is a normalizing constant such that $\boldsymbol{h}_i$ has unit length. This motivates the definition of the kernel weighted local influence of the i-th observation (y_{1i}, y_{2i}, r_i) as

$$C_{\boldsymbol{h}_{N(i)}} \;=\; 2 \left| \boldsymbol{h}_{N(i)}{}^T \, \boldsymbol{\Delta}^T \, \ddot{\boldsymbol{L}}^{-1} \, \boldsymbol{\Delta} \, \boldsymbol{h}_{N(i)} \right|, \tag{6.15}$$

where

$$\boldsymbol{h}_{N(i)} = (1 - \boldsymbol{w}_{(-N(i))})/D, \tag{6.16}$$

with $\boldsymbol{w}_{(-N(i))}$ as defined in (6.11) and D a normalizing constant. The choice $\boldsymbol{h}_{N(i)}$ reflects the influence of allowing subjects in the neighbourhood of the i-th subject to drop out non-randomly, while others, not within this neighbourhood, can only drop out at random. This method provides new insights in the local influence of types of observations.

It is again interesting to compute the kernel weighted local influence for a series of bandwidths. Because the vector $\boldsymbol{h}_{N(i)}$ is normalized, there is no need to have a density-adaptive bandwidth as in Section 6.4.1.

In the weighted local influence approach, applied on the Mastitis data, one is interested in whether the probability of occurrence of mastitis is related to the yield that would have been observed had mastitis not occurred for a cow with certain characteristics. In Figure 6.6, a kernel weighted influence analysis for 6 different bandwidths is shown for the local influence analysis.

For a larger bandwidth the left upper panel in Figure 6.6 suggests two groups of observations. The group with the highest influence is the group of completers, while the other group is the group of incompleters. If the bandwidth decreases, #66 shows up, as is shown in the right upper panel in Figure 6.6. For further decreasing bandwidths, #66 remains influential, while two other observations, #4 and #5, show up. The fact that #66 is dominantly present at several choices for the bandwidth, stresses the high degree of influence for this type of observations. The profile of #66 (Figure 1.2) is special in the sense that the milk yield in year 1 and year 2 are very high and so is the increase in milk yield. Types of observations with such a profile have a high dropout probability (Table 6.1) and, if they do not drop out, they are highly influential. This again illustrates the usefulness to examine the kernel weighted influence measures over a range of bandwidth values. The kernel weighted influence approach has the additional advantage to allow for a grid-based influence analysis as explained in the next section.

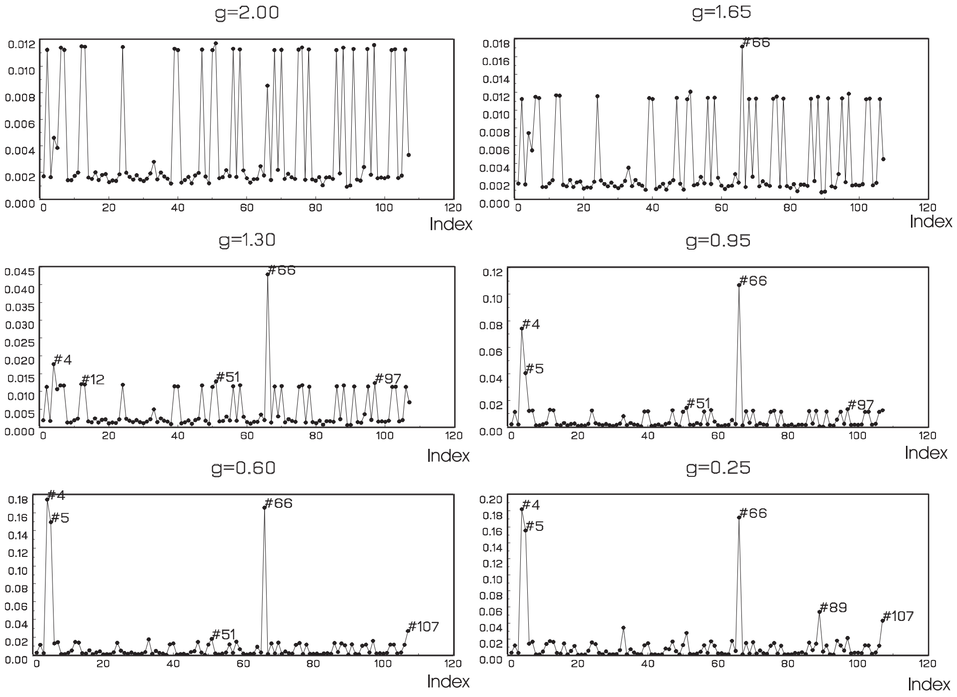


Figure 6.6: Influential subjects of the Mastitis data for the kernel weighted local influence (increments) with different bandwidths $g = g_1 = g_2$.

6.5 Grid-based Influence Measures

Instead of considering weights, centered at the datapoints (y_{1i}, y_{2i}, r_i), $i = 1, \ldots, N$, we now consider weights centered at points (y_1, y_2, r) on a one- or two-dimensional grid (for $r = 0$ and $r = 1$, respectively) enclosing the full observed data range. Define, in analogy with definition (6.11), the kernel based weight vector $\boldsymbol{w}_{(-N(y_1,y_2,r))}$ as the vector of length N with elements, for $j = 1, \ldots, N$,

$$(\boldsymbol{w}_{(-N(y_1,y_2,r))})_j = \quad \Big[K(0)\{K(0)\}^r \\ -K(\frac{y_{1j} - y_1}{g_1})\{K(\frac{y_{2j} - y_2}{g_2})\}^r I(r_j = r)\Big] /D, \quad (6.17)$$

where, as before, D is a normalization constant such that $\sum_{i=1}^{N}(\boldsymbol{w}_{(-N(y_1,y_2,r))})_j = N$, and define the kernel weighted global influence measure on the grid points (y_1, y_2, r) as

$$CD_{N(y_1,y_2,r)} = CD(\boldsymbol{w}_{(-N(y_1,y_2,r))}). \quad (6.18)$$

Examining the graph of $CD_{N(y_1,y_2,r)}$ as a function of y_1 (incompleters) or y_1 and y_2 (completers) allows us to identify influential regions over a grid, not only centered at the observed data points.

The kernel weighted local influence can be calculated over a grid in a similar way. With $\boldsymbol{h}_{N(y_1,y_2,r)} = (1 - \boldsymbol{w}_{(-N(y_1,y_2,r))})/D$ (D a normalizing constant), define the grid based weighted local influence as $C_{\boldsymbol{h}_{N(y_1,y_2,r)}}$. A plot of the weighted local influence values can be constructed and can lead to additional insights.

The two plots in Figure 6.7 show kernel weighted global influence values over a (y_1, y_2)-grid $[1, 9] \times [2, 12]$ in steps of 0.2. Again, as in Section 6.4.1, we used a density-adaptive bandwidth. The initial bandwidths $\tilde{g}_1$ and $\tilde{g}_2$ in (6.13) and (6.14) were chosen equal to 0.2 and 1.5, respectively.

These plots show that, using the available information in the Mastitis sample, certain types of observations are highly influential when modelled missing not at random in stead of missing at random. The peaks shown in Figure 6.7 confirm the results from Section 6.4.1. Indeed, a closer inspection of the first plot in Figure 6.7 reveals that the four highest peaks correspond to types of observations with characteristics similar to cows #4 and #5, to #54 and #69, to #66 and to #89.

The main structure of the second plot in Figure 6.7, based on a larger initial bandwidth, is essentially the same but the influence of observations at the border of the ellipsoidal area of datapoints gets more pronounced. Especially observations on that border, with Y_2 large, seem to be highly influential. A similar grid analysis for the incompleters didn't show any highly influential types of observations.

The construction of such a grid-based global influence graph is very computer intensive due to the calculation of the numerous (weighted) ML estimates. This is not the case for a grid analysis based on kernel weighted local influence, which is computationally much simpler. So, for the local influence measures, based on the directions $\boldsymbol{h}_{N(y_1,y_2,r)}$, we used a wider range, a finer grid and tried several bandwidth choices. Figure 6.8 shows a selection of weighted local influence graphs, for four different bandwidths. The main structure is essentially the same in each graph. If we have a closer look to the graphs for smaller bandwidths, the non-influential region is concentrated at the first principal component axis. The correlation between Y_1 and Y_2 is strongly positive, as can be seen in Figure 1.3. The types of observations which do not follow this main structure of the data, can be seen as potential outlying types of observations. Especially, types of observations with low values for Y_1 and high values for Y_2 seem to be influential. The highest influence for each of the graphs in Figure 6.8 for decreasing bandwidths is reached for (y_1, y_2) equal to $(2.93, 9.34)$;$(2.93, 8.49)$;$(3.08, 7.72)$;$(3.62, 7.41)$;$(3.78, 7.18)$ and $(3.93, 7.10)$, respectively. A closer look at these highly influential types of observations and to the Mastitis data shows that they are of the same type as observations #4 and #5.

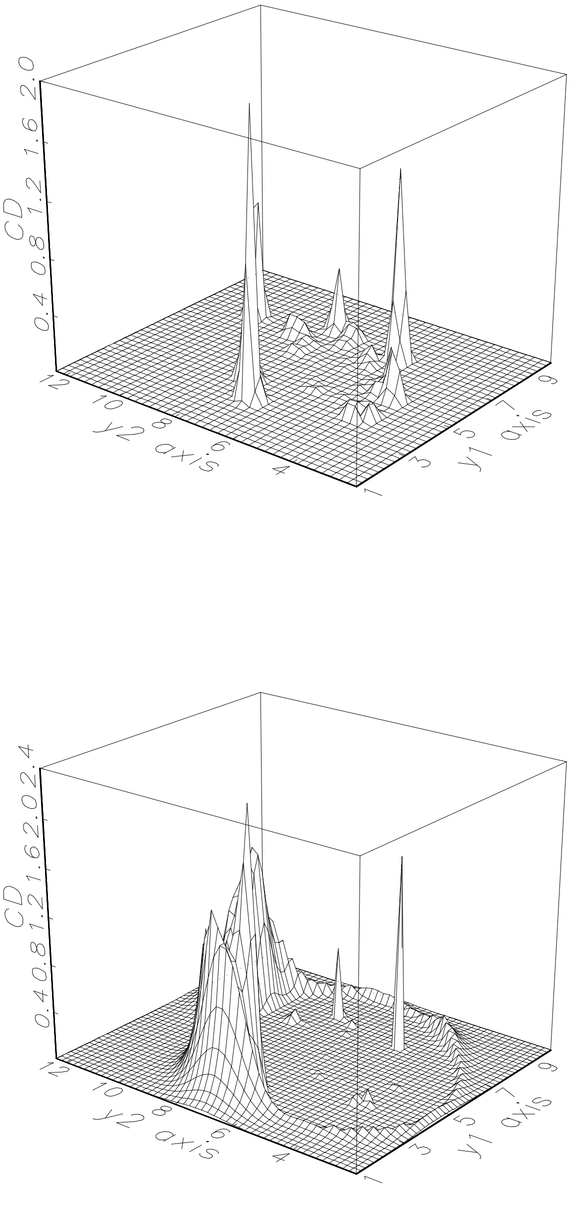


Figure 6.7: Kernel weighted global influence graph over a grid of completers with density-adaptive bandwidths initially equal to 0.2 (upper panel) and 1.5 (lower panel).

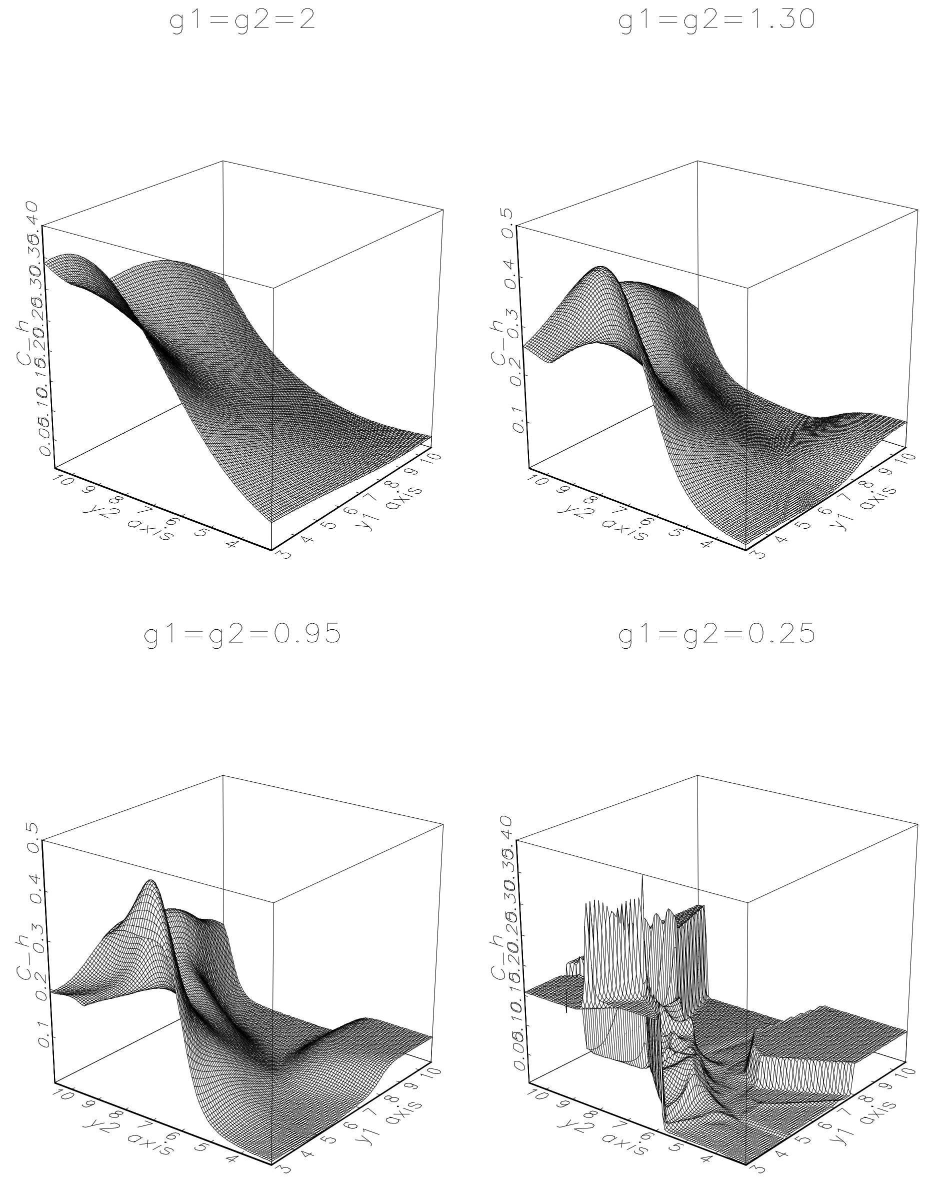


Figure 6.8: Kernel weighted local influence graphs over a grid of completers for several bandwidths $g_1 = g_2$.

This confirms our findings in Section 6.4.2.

A plot (omitted from the text) of the grid-based kernel weighted local influence for different bandwidths for types of incomplete observations showed little influence compared with the types of complete observations. The influential types of incomplete observations, when present, are located in the center of the first measurement-range $(3.5, 7.5)$.

A simulation study for the kernel weighted influence measures can give us a better insight in the source of influence for both complete and incomplete types of observations. Computationally, it is not feasible to carry out a simulation study for the grid-based kernel weighted global influence. Therefore, we restrict ourselves to a simulation study for the grid-based kernel weighted local influence.

6.6 A Simulation Study

A small simulation study is carried out in order to obtain new insights in the different sources of influence. For this simulation study, 100 similar datasets were generated. Each dataset consists of 107 subjects, each with two measurements generated from a bivariate normal distribution. Consider the following bivariate normal distribution, based on a compound symmetry covariance matrix:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 6.426 \\ 7.095 \end{pmatrix}, \begin{pmatrix} 2.865 & 2.324 \\ 2.324 & 2.865 \end{pmatrix}\right]. \tag{6.19}$$

The dropout process was generated according to the following model

$$\text{logit}[P(R_i = 0|Y_{i1}, Y_{i2})] = -3.379 + 0.387Y_{i1} + \psi_2 Y_{i2}, \tag{6.20}$$

where ψ_2 is the MNAR-parameter. The choice for the parameters in both the measurement model and dropout process was based on a fit of this model with $\psi_2 = 0$ (MAR) on the Mastitis data.

6.6.1 A First Setting

In a first simulation setting, 104 of the 107 subjects in each dataset were generated according to the process described above with ψ_2 equal to 0 (MAR). Three subjects however were generated with $\psi_2 = -0.5$, so three observations were allowed to be missing not at random. To compare the additional influence of generating 3 subjects which are allowed to be missing not at random versus the situation where all subjects are allowed to be missing at random, an average influence measure was plotted in Figure 6.9 for the completers and in Figure 6.10 for the incompleters. This average influence measure is the difference between the average grid-based influence of 100 datasets with 3 subjects allowed to be missing not at random and the average grid-based influence of 100 datasets, where none of the subjects were allowed to be missing not at random.

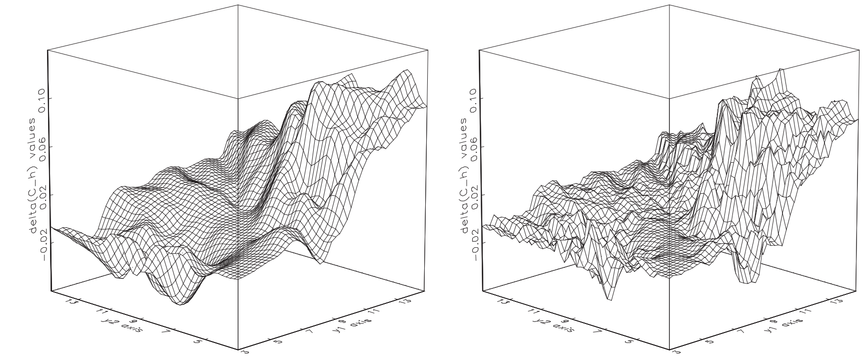


Figure 6.9: A figure of the relative average gain in influence of the completers when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5.

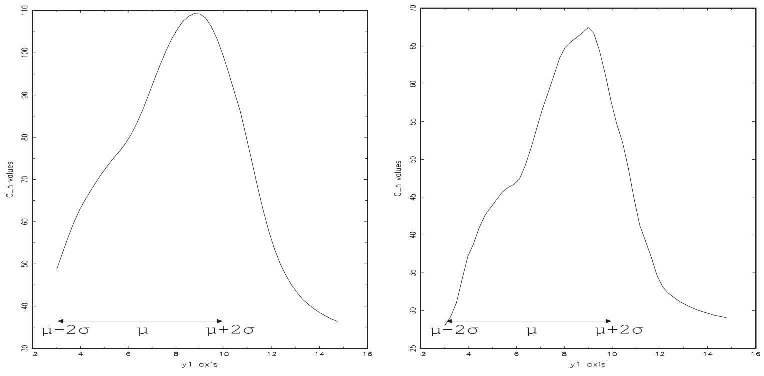


Figure 6.10: A figure of the relative average gain in influence of the incompleters when generating 3 subjects under MNAR. The bandwidths used are respectively equal to 1 and 0.5. μ and σ denote the mean and standard deviation of Y_1.

If we consider the dropout structure in Figure 6.11 for both MAR ($\psi_2 = 0$) and MNAR ($\psi_2 = -0.5$) and relate this to the results shown in Figure 6.9, it becomes clear that completers which tend to have a large probability of dropping out under the MNAR model, but do not, appear to be influential.

For the types of observations with a missing second measurement the largest influence is located at higher y_1 values as can be seen in Figure 6.10. A high value of y_1 often goes with a higher value of y_2 (correlation 0.8), a combination which has, under the MNAR model, a small probability to drop out. If it then drops out nevertheless, it is highly influential.

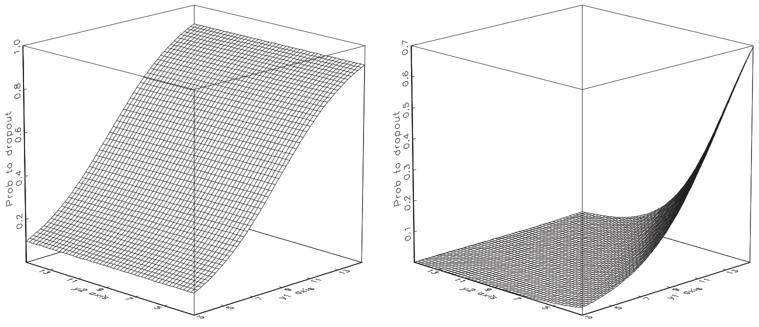


Figure 6.11: Plot of the probability of dropout. On the left hand side the dropout probability under MAR is shown, while on the right hand side the dropout probability is shown under MNAR.

6.6.2 A Second Setting

In a second simulation setting, the presence of subjects missing not at random is invoked by taking 100 datasets generated under MAR ($\psi_2 = 0$) as above, but now all data, with a second measurement higher than 8.5, are set to be missing.

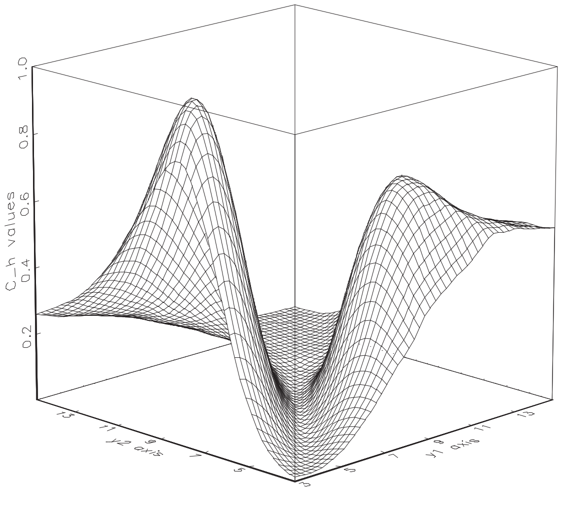


Figure 6.12: The average kernel weighted local influence for the completers of the 100 reference datasets

In Figure 6.12, the average influence measure of the completers of 100 datasets is shown. We will refer to these datasets generated under MAR as the reference datasets. The plot of the average influence of the completers of the reference datasets versus the grid has a particular shape. There is very low or no influence for data along the first principal component axis due to the high correlation ($\rho_{Y_1,Y_2} = 0.80$) between Y_1 and Y_2. When we move away from this axis the average influence increases. This indicates that outlying types of observations, not following the main pattern in the data, are influential. To see what the effect of invoking MNAR-dropout is on the completers, we leave out all observations in these datasets with a Y_2-measurement higher than 8.5 and calculate the average kernel weighted local influence again.

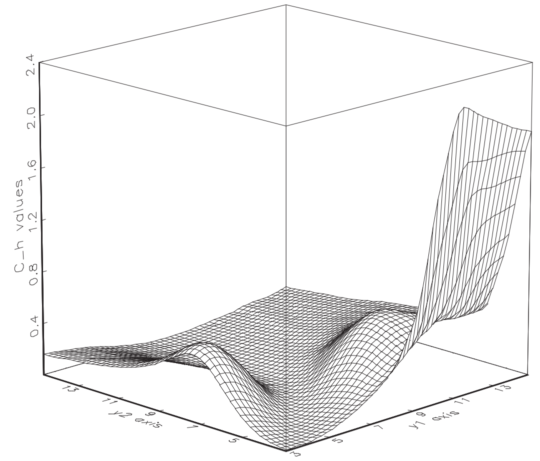


Figure 6.13: Kernel weighted local influence for the completers of the 100 complete datasets with MNAR dropouts for $Y_2 > 8.5$

The average influence of the completers under such a MNAR dropout process is shown in Figure 6.13, which indicates that dropout due to this MNAR mechanism has a large change in influence for types of completers with a high Y_1-measurement and a low Y_2-measurement. The larger influence of observations with a high Y_1-measurement and a low Y_2-measurement is not surprising. In Figure 6.14, a scatterplot of the completers is given. If we consider the structure of the data, we know that observations with a high value for Y_1 are more likely to be missing due to the underlying MAR-mechanism (Figure 6.11). Combined with the MNAR-mechanism we invoked in this setting, we especially obtain complete observations with a low Y_2-measurement. The correlation indicates that, among these types of observations, the ones with a low Y_1-measurement follow the correlation structure of the data. The ones with a high Y_1-measurement do not follow this structure and therefore they can be seen as outlying types of observations. Their influence is rather high

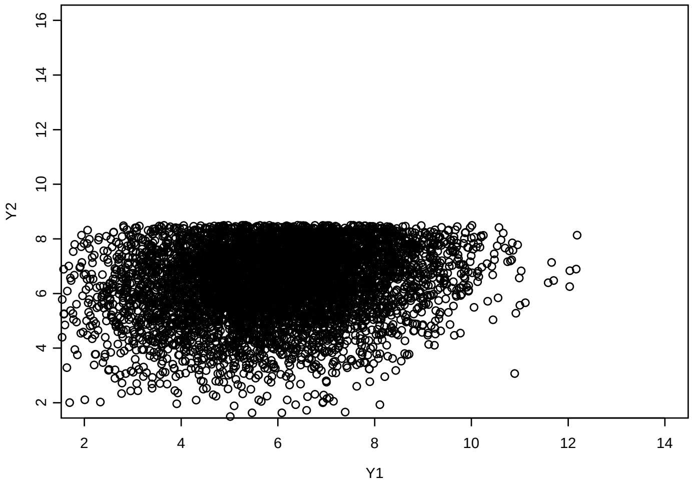


Figure 6.14: A scatterplot for all simulated datasets with MNAR dropouts for $Y_2 > 8.5$

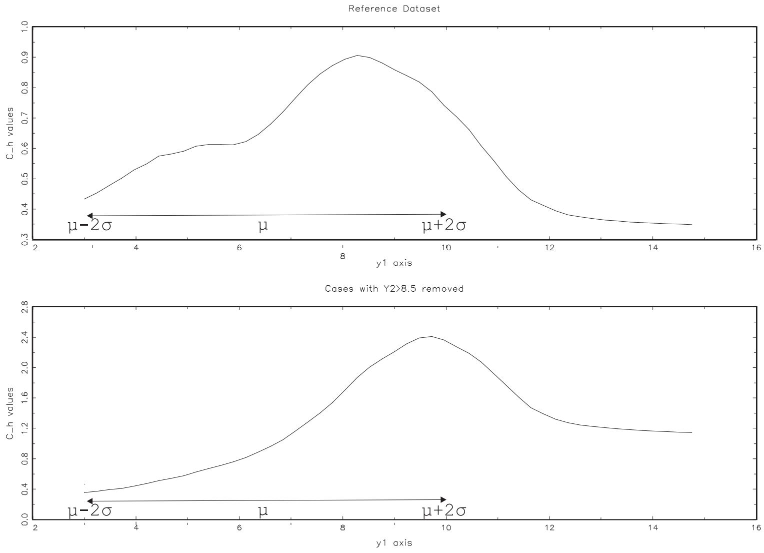


Figure 6.15: The figures of kernel weighted local influence for the incompleters of the complete dataset and the incompleters of the datasets with MNAR dropouts for $Y_2 > 8.5$

compared with the other types of observations.

Looking at the incompleters in Figure 6.15 one can see that there is a large change in influence on the incompleters. The highest average influence for the incompleters of the reference datasets was reached for $Y_1 = 8.5$, considering the MNAR-mechanism there is a shift towards $Y_1 = 9.75$. Not only this shift can be seen, but also the overall average influence increases. This indicates that the presence of types of observations which are left out non-randomly seem to have a large influence.

Other simulation settings (such as larger sample sizes) confirm these results. The main idea is illustrated here and therefore these other simulations are omitted from this chapter.

6.7 Discussion

In this chapter we introduced some new exploratory and graphical techniques, supplementing existing tools for sensitivity analysis. These methods combine parametric global and local influence measures with non-parametric smoothing weights. They provide new insights in the influence of certain types of observations and offer a nice solution to the problem of masking. The discussion here has been focusing on the setting of two (repeated) measurements. In case of three or more measurements, the kernel based weights (6.11) and (6.21) can be based on higher dimensional kernels. Alternatively, one can first determine the Euclidean distance between two observations (belonging to the same pattern) in combination with a one-dimensional kernel function. This latter option leads to the following extension of the weights (6.11), to any number of measurements.

Let $(\boldsymbol{y}_i, \boldsymbol{r}_i)$ denote the data where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in}) = (\boldsymbol{y}_i^o, \boldsymbol{y}_i^m)$ is the vector of observed components $\boldsymbol{y}_i^o$ and missing components $\boldsymbol{y}_i^m$ and where $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{in})$ is the vector grouping the missingness indicators

$$r_{i\ell} = \begin{cases} 1 & \text{if } y_{i\ell} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

For a neighbourhood $N(i)$ of outcome $(\boldsymbol{y}_i, \boldsymbol{r}_i)$, define the weights

$$(\boldsymbol{w}_{(-N(i))})_j = \{K(0) - K(\|\boldsymbol{y}_j^o - \boldsymbol{y}_i^o\|/g)I(\boldsymbol{r}_j = \boldsymbol{r}_i)\}/D, \qquad (6.21)$$

where K is for instance a Gaussian kernel function, g is the bandwidth and D a normalizing constant, as before. So, similar to the weights (6.11), the weights (6.21) are constant for all observations with a different missingness pattern ($\boldsymbol{r}_j \neq \boldsymbol{r}_i$) and assign low weights to all observations $\boldsymbol{y}_j$ in the close neighbourhood of $\boldsymbol{y}_i$ and with an identical missingness pattern ($\boldsymbol{r}_j = \boldsymbol{r}_i$). Note that this definition is not restricted to monotone dropout missingness mechanisms.

As a further generalization one could extend the concept of the neighbourhood of a particular observation $(\boldsymbol{y}_i, \boldsymbol{r}_i)$ to all observations with not only an identical missingness pattern $\boldsymbol{r}_i$ but also with a similar pattern, in this way including, for example, observations which drop out one time point earlier or later. This could be an interesting option in order to enlarge the number of effective observations in the neighbourhood of $(\boldsymbol{y}_i, \boldsymbol{r}_i)$ which is, especially in case of several measurements and in view of the curse of dimensionality, not unimportant.

A deeper study of the properties and the applicability of this extension to more than two measurements is beyond the scope of this chapter.

The local influence methodology and the proposed weighted influence methodology are both tools for sensitivity analysis. It has been shown that these tools pick up a lot of different anomalies in the data, not just deviations of the MNAR mechanism. One possible tool to assess the appropriateness of including the MNAR-parameter in the model of Diggle and Kenward (1994) is the Likelihood Ratio Test to test for MAR versus MNAR. Many authors have noted that there is very little information available for the MNAR-parameter, in addition to the information available for all other parameters. If this were to be true, this ought to show in the behaviour of the likelihood ratio test statistic, as well as in the structure of the information matrix for the vector of model parameters. This will be explored in the next chapter.

Chapter 7

Behaviour of the Likelihood Ratio Test for MAR versus MNAR

7.1 Introduction

Recall the selection model of Diggle and Kenward (1994) as introduced in Section 6.3.1 for two occasions. Rubin's (1976) classification of missing data into three types, missing completely at random, missing at random and missing not at random can be translated into the presence or absence of specific parameters in the drop out part of the model. Opposing the different missingness mechanisms to each other can be done using a likelihood ratio test. In classical theory, the asymptotic distribution of the likelihood ratio test is a chisquare distribution with degrees of freedom equal to the difference in number of parameters. Careful considerations have to be made when using this result to test for missing not at random as shown by Rotnitzky *et al.* (2000) and by Bottai (2003) in a simpler setting.

In Section 7.2, we will formally introduce the framework in which we work. Section 7.3 gives an informal look at theoretical considerations regarding the distribution. An overview of different simulation settings to illustrate the finite sample behaviour of the likelihood ratio test will be given in Section 7.4. In an attempt to generate the null distribution for a given data set, two bootstrap methods will be introduced in Section 7.5. Finally, a discussion is provided in Section 7.6.

7.2 The Selection Model by Diggle and Kenward (1994)

The selection model of Diggle and Kenward (1994) was already introduced in Section 6.3.1 for two occasions. The terminology here, is repeated for J occasions. The different missingness mechanisms according to Rubin (1976) can easily be expressed in the selection modelling framework as described in Section 1.2.2.

Let us assume that for subject i, $i = 1, \cdots, N$, a sequence of responses Y_{ij} is measured at several occasions $j = 1, 2, \ldots, J$. Let R_{ij} be a missingness indicator and assume that y_{i1} is always observed. Then $r_{ij} = 0$ if y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is observed. The measurement part of the model of Diggle and Kenward (1994) is given by

$$\mathbf{Y_i} = (Y_{i1}, \ldots, Y_{iJ}) \sim N(X_i\boldsymbol{\beta}, \Sigma_i), \qquad i = 1, \ldots, N, \tag{7.1}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, X_i is a matrix containing covariate values and Σ_i is a covariance matrix. The missingness process is described by

$$\text{logit}[Pr(R_{ij} = 0|y_{i,j-1}, y_{ij})] = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}, \tag{7.2}$$

where $Pr(R_{ij} = 0|y_{i,j-1}, y_{ij})$ is the probability for the i-th subject to drop out at time j. If ψ_2 differs from zero, the missingness process is non-random. Let us denote

$$g(\mathbf{h_{id}}, y_{id}) = Pr(R_{id} = 1|y_{i,d-1}, y_{id}),$$

with d the time of dropout and $\mathbf{h_{id}} = (y_{i1}, \ldots, y_{i,d-1})$ the history of y_{id}, which we now restrict to depend on the previous measurement only. The total loglikelihood has the form

$$\ell = \sum_{i=1}^{N} [r_i \ell_i^c + (1 - r_i)\ell_i^i],$$

with ℓ_i^i the contribution for an incompleter

$$\ell_i^i = \ln f(\mathbf{h_{id}}) + \sum_{j=2}^{d_i - 1} \ln[1 - g(\mathbf{h_{ij}}, y_{ij})] + \ln \int f(y_{id}|\mathbf{h_{id}}) g(\mathbf{h_{id}}, y_{id})\ dy_{id},$$

and ℓ_i^c the contribution for a completer

$$\ell_i^c = \ln f(\mathbf{y_i}) + \sum_{j=2}^{J} \ln[1 - g(\mathbf{h_{ij}}, y_{ij})].$$

The likelihood ratio test statistic for testing $\psi_2 = 0$, and thus MNAR versus MAR, is then given by

$$G = -2[\ell_\gamma(\hat{\gamma}) - \ell_{\gamma^*}(\hat{\gamma^*})],$$

where $\gamma = (\beta_1, \beta_2, \sigma, d, \psi_0, \psi_1, \psi_2)$ and $\gamma^* = (\beta_1, \beta_2, \sigma, d, \psi_0, \psi_1, 0)$. Due to the difference in only one parameter, the distribution of this statistic can be misleadingly expected to be $\chi^2(1)$. Based on this statistic Kenward (1998) and Molenberghs *et al.* (2001) rejected the null hypothesis of missing at random on a value of 5.11, which corresponds to a p-value of 0.02 for their data example (Mastitis in dairy cattle). They compared this result with the Wald test (p-value of 0.002) and concluded that the asymptotic approximations are not very accurate. Rotnitzky *et al.* (2000) state that the regular assumptions of the likelihood ratio test statistic do not hold in case of a singular information matrix. In the next paragraph, we will have an informal look at the theoretical aspects of the distribution.

7.3 An Informal Look

In a first subsection we will focus on an example used by Rotnitzky *et al.* (2000) to motivate the need for a careful use of the likelihood ratio test statistic in the context of the selection model introduced in Section 7.2.

7.3.1 The Example of Rotnitzky *et al.* (2000)

Suppose that $y_1, \ldots, y_n$ are observations from a normal distribution with mean β and variance σ^2. Suppose there is the possibility that the value of y_i is missing with probability

$$P_c(y; \alpha_0, \alpha_1) = 1 - exp\{H(\alpha_0 + \alpha_1(y - \beta)/\sigma)\}, \tag{7.3}$$

where α_0 and α_1 are unknown parameters and $H(\cdot)$ is a known function assumed to have its first three derivatives at α_0 non-zero. Interest may lie in small values of α_1 and especially in testing the null hypothesis $\alpha_1 = 0$. The likelihood contribution of an individual is given by

$$\begin{aligned} L_n(\beta, \sigma, \alpha_0, \alpha_1) &= \left[\frac{1}{\sqrt{2\pi}\sigma} e^{(\frac{(y-\beta)^2}{2\sigma^2})} e^{H\{\alpha_0 + \alpha_1(y-\beta)/\sigma\}}\right]^r \\ &\cdot \left[\int (1 - e^{H\{\alpha_0 + \alpha_1(y-\beta)/\sigma\}}) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y-\beta)^2}{2\sigma^2}} dy\right]^{(1-r)}, \end{aligned}$$

such that the loglikelihood contribution equals

$$\begin{aligned} \ell(\beta, \sigma, \alpha_0, \alpha_1) &= r\left[-\ln(\sqrt{2\pi}) - \ln(\sigma) + \frac{(y-\beta)^2}{2\sigma^2} + H\{\alpha_0 + \alpha_1(y-\beta)/\sigma\}\right] \\ &+ (1-r)\left[\ln(\int (1 - e^{H\{\alpha_0 + \alpha_1(y-\beta)/\sigma\}}) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y-\beta)^2}{2\sigma^2}} dy)\right]. \end{aligned}$$

From this, the score equations can be calculated

$$\begin{aligned}
S_1^* = \frac{\partial \ell}{\partial \beta}(\beta, \sigma, \alpha_0, 0) &= \frac{r(y-\beta)}{\sigma^2}, \\
S_2^* = \frac{\partial \ell}{\partial \sigma}(\beta, \sigma, \alpha_0, 0) &= rH'(\alpha_0) - (1-r)\frac{H'(\alpha_0)e^{H(\alpha_0)}}{1-e^{H(\alpha_0)}}, \\
S_3^* = \frac{\partial \ell}{\partial \alpha_0}(\beta, \sigma, \alpha_0, 0) &= \frac{rH'(\alpha_0)(y-\beta)}{\sigma}, \\
S_4^* = \frac{\partial \ell}{\partial \alpha_1}(\beta, \sigma, \alpha_0, 0) &= \frac{-r}{\sigma} + r\frac{(y-\beta)^2}{\sigma^3}.
\end{aligned}$$

We can see that S_1^* and S_3^* are proportional and so this set of equations is degenerate at this particular parameter point. Equivalently, the information matrix calculated from expected second order derivatives is singular at this parameter point.

7.3.2 Likelihood-based Inference with Singular Information Matrix

The key feature is the singularity of the information matrix. Rotnitzky *et al.* (2000) formulated two basic theorems who give us the asymptotic distribution of the likelihood ratio test statistic when dealing with a singular information matrix. We formulate them in a multidimensional setting which relies on several regularity conditions. Similar theorems were formulated relaxing upon some of the conditions, they can be found in Rotnitzky *et al.* (2000).

Let us first introduce some terminology. Let $Y_1, \ldots, Y_n$ denote n independent copies of a random variable Y with density $f(y; \boldsymbol{\theta}^*)$ where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)$ is an unknown parameter. Let us assume some regularity conditions on $f(y; \boldsymbol{\theta})$ are fulfilled. These essentially consist of the usual smoothness assumptions that guarantee uniqueness and consistency of the ML estimator and in addition the existence in a neighbourhood of $\boldsymbol{\theta}^*$ of $2s+1$ derivatives with respect to $\boldsymbol{\theta}$ of $\ell(Y; \boldsymbol{\theta}) = \log f(Y; \boldsymbol{\theta})$ for some positive integer s with absolute values uniformly bounded by functions of Y that have finite mean (see Rotnitzky *et al.* 2000 for more details). s is a positive integer for which

$$\partial^j \ell(Y; \boldsymbol{\theta})/\partial\theta_1^j|_{\boldsymbol{\theta}^*} = 0, \; 1 \leq j \leq s-1 \qquad \text{and} \qquad \partial^s \ell(Y; \boldsymbol{\theta})/\partial\theta_1^s|_{\boldsymbol{\theta}^*} \neq 0. \quad (7.4)$$

Let $S_j(\boldsymbol{\theta})$ denote the score equation with respect to θ_j, $1 \leq j \leq p$ and $S_j = S_j(\boldsymbol{\theta}^*)$. Let I denote the covariance matrix of $(S_1^{(s)}/s!, S_2, \ldots, S_n)$, where $S_1^{(s)} = \partial^s \ell(Y; \boldsymbol{\theta})/\partial\theta_1^s|_{\boldsymbol{\theta}^*}$. The rank of the information matrix at $\boldsymbol{\theta}^*$ is $p-1$ if and only if $p-1$ elements of the score vector, say the last $p-1$ scores are linearly independent and the remaining score is equal to a linear combination of them. Let us denote

$\rightsquigarrow$ for convergence in distribution under $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Denote $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ a mean-zero normal random vector with variance equal to I^{-1} and B a Bernoulli variable with success probability equal to 1/2 that is independent of Z. Let I^{jk} denote the (j,k)-th entry of I^{-1}. The following theorems hold.

Theorem 2. *Under regularity conditions, when s is odd:*

(a) the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ exists when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, it is unique with a probability tending to 1*, and it is a consistent estimator of $\boldsymbol{\theta}$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,*

(b)

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}(\hat{\theta}_2 - \theta_2^*) \\ \vdots \\ n^{1/s}(\hat{\theta}_p - \theta_p^*) \end{bmatrix} \rightsquigarrow \begin{bmatrix} Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}; \tag{7.5}$$

(c)

$$2\{L_n(\hat{\boldsymbol{\theta}}) - L_n(\boldsymbol{\theta}^*)\} \rightsquigarrow \chi_p^2.$$

Theorem 3. *Under regularity conditions, when s is even:*

(a) the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ exists when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, it is unique with a probability tending to 1*, and it is a consistent estimator of $\boldsymbol{\theta}$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,*

(b)

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}(\hat{\theta}_2 - \theta_2^*) \\ \vdots \\ n^{1/s}(\hat{\theta}_p - \theta_p^*) \end{bmatrix} \rightsquigarrow \begin{bmatrix} (-1)^B Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix} I_{(Z_1>0)} + \begin{bmatrix} 0 \\ Z_2 - (I^{21}/I^{11})Z_1 \\ \vdots \\ Z_p - (I^{p1}/I^{11})Z_1 \end{bmatrix} I_{(Z_1<0)}; \tag{7.6}$$

(c)

$$2\{L_n(\hat{\boldsymbol{\theta}}) - L_n(\boldsymbol{\theta}^*)\} \rightsquigarrow \sum_{j=1}^{p} Z_j^{*2} I_{(Z_1^*>0)} + \sum_{j=2}^{p} Z_j^{*2} I_{(Z_1^*<0)},$$

where Z_j^*, $j = 1, 2, \ldots, p$, are independent $N(0,1)$ random variables. That is, the asymptotic distribution of the likelihood ratio test statistics is a mixture of a χ_{p-1}^2 and χ_p^2 random variable, with mixing probabilities equal to 1/2*, where $I_{(A)}$ is an indicator variable which takes the value 1 if A is true and 0 if not.*

Rotnitzky *et al.* (2000) show that the difference between the likelihood ratio test statistic and its limiting random variable is of order $O_p(n^{-1/(2s)})$.

For the example shown in Section 7.3.1, $s = 3$ and so the limiting distribution according to Theorem 1 is a χ^2_1-distribution.

In the next section, we will show that the set of score equations for the likelihood ratio test opposing different missingness mechanisms for a simplified selection model are degenerate.

7.3.3 Testing Hypotheses in the Selection Modelling Framework

Let us consider the selection model introduced in Section 7.2. To simplify the general derivations, we will only derive the full expressions for the case of a covariance matrix expressing compound symmetry and for the case $J = 2$. The derivations are analogues for other association structures.

The full likelihood is given by

$$\ell = \sum_{i=1}^{N} [r_i \ell_i^c + (1 - r_i)\ell_i^i],$$

with

$$l_i^c = \ln f(y_{i1}, y_{i2}) + \ln[1 - g(h_{i2}, y_{i2})],$$

and

$$l_i^i = \ln f(y_{i1}) + \ln \int f(y_{i2}) g(h_{i2}, y_{i2}) dy_{i2}.$$

If

$$\Sigma = \begin{pmatrix} \sigma^2 + d & d \\ d & \sigma^2 + d \end{pmatrix},$$

we know that

$$\begin{aligned} f(y_{i1}, y_{i2}) &= \frac{1}{2\pi} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\beta)^T \Sigma^{-1} (\mathbf{y}-\beta)} \\ &= \frac{1}{2\pi\sigma(\sigma^2 + 2d)^{1/2}} e^{-\frac{(y_{i1}-\beta_1)^2(\sigma^2+d) - 2(y_{i1}-\beta_1)(y_{i2}-\beta_2)d + (y_{i2}-\beta_2)^2(\sigma^2+d)}{2\sigma^2(\sigma^2+d)}}, \end{aligned}$$

and

$$g(h_{i2}, y_{i2}) = \frac{e^{\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}}}{1 + e^{\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}}} = \frac{1}{1 + e^{-\psi_0 - \psi_1 y_{i1} - \psi_2 y_{i2}}}.$$

So the expression for an individual contribution to the loglikelihood is the following one:

$$\begin{aligned} \ell_i &= r_i\Bigg[-\ln(2\pi) - \ln(\sigma) - \frac{1}{2}\ln(\sigma^2+2d) \\ &- \frac{(y_{i1}-\beta_1)^2(\sigma^2+d) - 2(y_{i1}-\beta_1)(y_{i2}-\beta_2)d + (y_{i2}-\beta_2)^2(\sigma^2+d)}{2\sigma^2(\sigma^2+2d)} \\ &- \ln(1+e^{\psi_0+\psi_1 y_{i1}+\psi_2 y_{i2}})\Bigg] \\ &+ (1-r_i)\Bigg[-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2+d) - \frac{(y_{i1}-\beta_1)^2}{2(\sigma^2+d)} \\ &+ \ln\int \frac{1}{\sqrt{2\pi(\sigma^2+d)}}\frac{e^{-\frac{(y_{i2}-\beta_2)^2}{2(\sigma^2+2d)}}}{1+e^{-\psi_0-\psi_1 y_{i1}-\psi_2 y_{i2}}}dy_{i2}\Bigg]. \end{aligned}$$

From this expression, we can calculate the score equations S_k, $k = 1, \ldots, 7$, where

$$S_k = \frac{\partial \ell_i}{\partial \gamma_k},$$

with γ_k the $k^{\mathrm{t}h}$ component of $\gamma = (\beta_1, \beta_2, \sigma, d, \psi_0, \psi_1, \psi_2)$. In the general form, these score equations are fairly complicated. Let us therefore look at a simpler setting.

Let us consider the specific situation that we test for $\psi_2 = 0$ while $\psi_1 = 0$ in equation (7.2). Denote $\boldsymbol{\gamma}^* = (\beta_1, \beta_2, \sigma, d, \psi_0, 0)$, which corresponds to the null hypothesis: $\psi_2 = 0$. Assume furthermore that the correlation between the two measurements is 0 ($d = 0$) and standardize y_{i2} in the dropout model by $(y_{i2} - \beta_2)/\sigma$.

The score equations S_k, $k = 1, \ldots, 5$ where S_k corresponds to the score equation of the $k^{\mathrm{t}h}$ component of $\gamma = (\beta_1, \beta_2, \sigma, \psi_0, \psi_2)$ are given by

$$\begin{aligned} S_1(\gamma^*) &= \frac{y_{i1}-\beta_1}{\sigma^2}, \\ S_2(\gamma^*) &= \frac{r_i(y_{i2}-\beta_2)}{\sigma^2}, \\ S_3(\gamma^*) &= \frac{-\sigma^2 - 2r_i y_{i2}\beta_2 + r_i\beta_2^2 + r_i y_{i2}^2 - 2y_{i1}\beta_1 + y_{i1}^2 - r_i\sigma_i^2 + \beta_1^2}{\sigma^3}, \\ S_4(\gamma^*) &= \frac{-(r_i e^{-\psi_0} - e^{-\psi_0} + r_i)}{1+e^{-\psi_0}}, \\ S_5(\gamma^*) &= \frac{r_i(-y_{i2}+\beta_2)}{(1+e^{-\psi_0})\sigma^2}. \end{aligned}$$

Again similar to the example of Rotnitzky *et al.* (2000), these score equations are degenerate. $S_5(\gamma^*)$ is proportional to $S_3(\gamma^*)$ and therefore the information matrix is singular.

This and the more general situation, however, have not been studied before. Indeed, Rotnitzky *et al.* (2000) focus on simple null hypotheses while in this situation the parameter vector can be divided into a parameter of interest and several nuisance parameters. Even if the result of Rotnitzky *et al.* (2000) holds, the question remains, whether it is applicable for finite samples.

Another remark is that the calculations for the more general case become more complicated due to the presence of correlations between the two measurements and due to a more complex dropout mechanism including the ψ_1-term. However, one can expect similar issues to occur.

In the next paragraph, we will illustrate the behaviour of the likelihood ratio test statistic for the different missingness parameters in a simple setting by means of a simulation study.

7.4 Simulating the Likelihood Ratio Test Statistic for the Different Missingness Processes

The presented simulation study focuses on the asymptotic distribution of the likelihood ratio test statistic for the different missingness processes. In Section 7.5, we discuss two bootstrap approaches to implement the likelihood ratio test statistic for testing missingness not at random. We will restrict our discussion and derivations to the situation $J = 2$.

400 similar datasets were generated in 4 different settings. Each data set consists of 200 subjects, each with two measurements generated from a bivariate normal distribution. Consider the following bivariate normal distribution, based on a compound symmetry covariance matrix:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}\right]. \tag{7.7}$$

The dropout process was generated according to the following model

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = \psi_0 + \psi_1 Y_{i1} + \psi_2 Y_{i2}, \tag{7.8}$$

where $\psi_0 = -2$ and ψ_1 and ψ_2 were chosen according to four different settings. An overview of the settings is given in the following table. In the situations where $\psi_1 \neq 0$, ψ_1 was chosen to be 1. Figure 7.1 shows plots of the simulated null-distributions together with approximating χ^2-distribution. In Table 7.2, the critical values for each of the four settings are shown together with the critical values of

Table 7.1: Overview of the different simulation settings.

Setting	Missingness Process $\text{logit}[P(R_i = 1 \mid Y_{i1}, Y_{i2})] =$	Null Hypothesis
1	$\psi_0 + \psi_1 Y_{i1}$	$\psi_1 = 0$
2	$\psi_0 + \psi_1 Y_{i1} + \psi_2 Y_{i2}$	$\psi_1 = 0$
3	$\psi_0 + \psi_2 Y_{i2}$	$\psi_2 = 0$
4	$\psi_0 + \psi_1 Y_{i1} + \psi_2 Y_{i2}$	$\psi_2 = 0$

the $\chi^2(1)$-, $\chi^2(2)$- and $\chi^2(3)$-distribution. From Figure 7.1 and the critical values in Table 7.2, it is clear that the distribution corresponding to setting 1 is close to the asymptotically expected $\chi^2(1)$. The distribution for setting 2 is closer to a χ^2-distribution with 2 degrees of freedom, while setting 3 lies in between. Setting 4 seems to correspond to a $\chi^2(1)$-distribution but the critical values in Table 7.2 do not correspond to those of a $\chi^2(1)$-distribution but to those of a $\chi^2(3)$-distribution. This is in contrast with the findings of Rotnitzky *et al.* (2000).

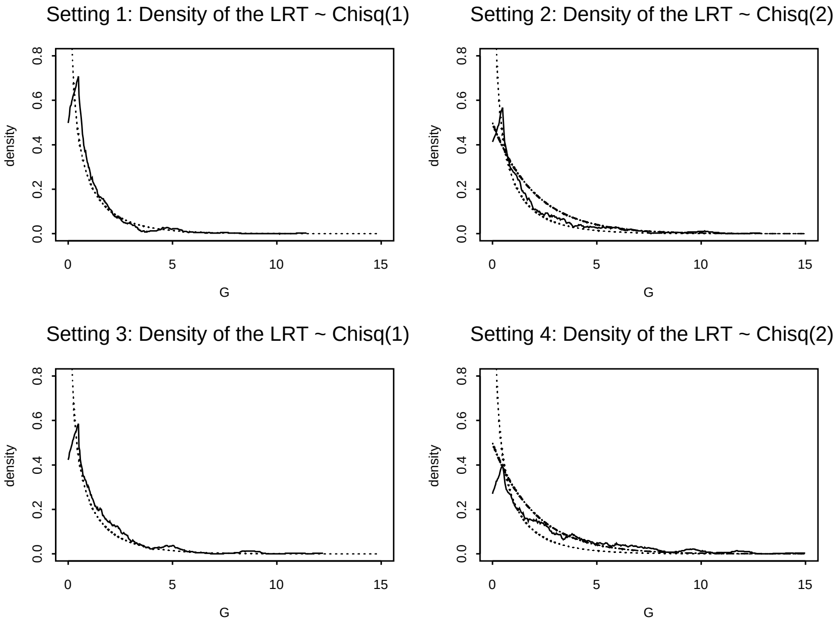


Figure 7.1: Density plots (solid curve) of the different settings with $\chi^2(1)$-distribution (dotted curve) and $\chi^2(2)$-distribution (dashed-dotted curve).

Table 7.2: Table of critical values for the four different settings and some χ^2-distributions.

Method	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
Setting 1	2.23	3.27	6.04
Setting 2	4.07	5.74	9.85
Setting 3	3.18	4.75	8.74
Setting 4	6.38	8.29	11.95
$\chi^2(1)$	2.71	3.84	6.63
$\chi^2(2)$	4.61	5.99	9.21
$\chi^2(3)$	6.25	7.82	11.35

7.5 Bootstrap Approaches

From previous simulation results, it appears that the classical asymptotic theory is not applicable. To decide whether the hypothesis of missing at random holds for any given data set, one has to be able to generate the null distribution. A well known method is the parametric bootstrap, first introduced by Efron (1979), which heavily relies on distributional assumptions. To relax these assumptions we propose to use a semi-parametric bootstrap method. In the next subsections both bootstrap procedures are introduced.

7.5.1 Parametric Bootstrap

The parametric bootstrap scheme is as follows:

1. fit the initial data under the null and the alternative hypothesis resulting in $(\hat{\theta}_{\mathbf{0}}, \hat{\psi}_{\mathbf{0}})$ and $(\hat{\theta}_{\mathbf{1}}, \hat{\psi}_{\mathbf{1}})$, respectively, where θ denotes the parameter vector for the measurement part and ψ for the missingness part; compute the LRT for the hypotheses under consideration,

2. generate a 'bootstrap sample' from the selection model, reflecting the null hypothesis by using the estimates $(\hat{\theta}_{\mathbf{1}}, \hat{\psi}_{\mathbf{0}})$,

3. compute the LRT test for the bootstrap sample,

4. repeat step 2 and 3 B times and determine the bootstrap p-value as the proportion of bootstrap LRT values larger than its value for the original data from step 1.

Alternatively, step 2 could be based on the estimates $(\hat{\theta}_{\mathbf{0}}, \hat{\psi}_{\mathbf{0}})$. But some exploratory simulations showed that both choices resulted in essentially the same p-values. The parametric bootstrap heavily depends on the quality of the estimates $(\hat{\theta}_{\mathbf{1}}, \hat{\psi}_{\mathbf{0}})$. In case the initial data are generated under the alternative, one can expect that bias disturbs the procedure. This would lead to the generation of bootstrap data in step 2 which would obey the null constraint but which would be substantially different from the initial data in many other respects. A semi-parametric model based on resampling and less depending on the estimates from the initial sample might perform better.

7.5.2 Semi-Parametric Bootstrap

Given the data, a semi-parametric bootstrap procedure for testing hypotheses in the selection model can be implemented using the following algorithm:

1. fit the initial data under the null and the alternative hypothesis resulting in $(\hat{\theta}_{\mathbf{0}}, \hat{\psi}_{\mathbf{0}})$ and $(\hat{\theta}_{\mathbf{1}}, \hat{\psi}_{\mathbf{1}})$, respectively; compute the LRT for the hypothesis under consideration,

2. impute the missing data, conditionally on the observed outcomes at the previous occasion, and based on the probability model for the measurement part using the estimate $\hat{\theta}_{\mathbf{1}}$ (this is a parametric part),

3. draw (complete) observations from the augmented data set (resulting from step 2), with replacement, yielding a new sample of the same size N (this resampling is the non-parametric part),

4. observations at time $t \geq 2$ are deleted with a probability according the logistic dropout model using the estimate $\hat{\psi}_0$ (thus reflecting the null hypothesis; this is again a parametric part); this is the final bootstrap sample,

5. compute the LRT test for the bootstrap sample,

6. repeat step 2 and 5 B times and determine the bootstrap p-value as the proportion of bootstrap LRT values larger than its value from the initial data from step 1.

For more details about similar semi-parametric bootstrap implementations in other settings, see Davison and Hinkley (1997).

In the next section, the two bootstrap methods are illustrated on a simulated data example.

7.5.3 Simulated Data Example

In this section the methods introduced in the previous section are investigated further. Setting 2 in Section 7.4 is not of practical use and therefore omitted from the simulation study. Let us consider hypothesis 1, 2 and 3, which correspond to settings 1, 3 and 4, respectively.

Hypothesis 1: MAR vs MCAR

In Table 7.3, the situation of MAR vs MCAR is given. In this setting, the initial data set of size $N = 200$ is generated according to the following model

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}\right]. \tag{7.9}$$

and the dropout process is given by

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -2 + \psi_1 Y_{i1} \tag{7.10}$$

We generated 400 parametric bootstrap values and 400 semi-parametric bootstrap values in three different situations.

- Scenario 1: all N observations generated under the null hypothesis,
- Scenario 2: all N observations generated under the alternative,
- Scenario 3: 190 observations generated under the null hypothesis and 10 observations under the corresponding alternative.

In this situation the asymptotic distribution is known to be $\chi^2(1)$. The simulations confirm this. Table 7.3 shows that both bootstrap methods perform well. Fitting the selection model, obtaining the maximum likelihood estimates and computing the LRT is a nontrivial iterative computing exercise, not lending itself for intensive simulations. A full simulation study based on, e.g., 100 initial samples was computationally not feasible. The '*optmum*' procedure in Gauss 3.2.32 was used for computations. The optimization method used the Broyden-Fletcher-Goldfarb-Shanno procedure (Shanno, 1985) to obtain starting values for the Newton Raphson procedure and it took about one week to obtain the results of one of the 18 combinations.

Table 7.3: Hypothesis 1: Critical points based on the parametric and semi-parametric bootstrap procedure (400 bootstrap runs) for two initial data sets. Lower lines show the critical points of the simulated null distribution based on 800 samples, together with those of the $\chi^2(1)$ distribution.

		quantiles			p-value
		0.10	0.05	0.01	
Scenario 1	Parametric	3.04	4.17	6.12	0.7556
		2.53	3.35	6.76	0.3566
	Semi-Parametric	2.96	3.83	6.22	0.7890
		2.46	4.16	6.60	0.3616
Scenario 2	Parametric	2.55	3.39	6.36	<0.0025
		2.83	3.68	7.02	<0.0025
	Semi-Parametric	2.41	3.39	6.49	<0.0025
		2.68	3.68	6.37	<0.0025
Scenario 3	Parametric	2.35	3.72	7.91	0.9352
		3.00	3.93	6.48	<0.0025
	Semi-Parametric	2.83	4.13	8.00	0.6085
		2.70	4.40	6.49	<0.0025
simulated H_0		2.23	3.27	6.04	
$\chi^2(1)$ distribution		2.71	3.84	6.63	

Nevertheless, we think that our limited results do reveal the main characteristics of the performance of both bootstrap procedures.

For Hypothesis 1, Table 7.3 shows that, for all scenarios, the $\chi^2(1)$ approximation and the bootstrap approximation to the null distribution are consistent and in line with our expectations. Note that the results for the two initial data sets under Scenario 3 are not in agreement: one of them clearly rejects the hypothesis and the other clearly not. Since only 5% of the initial data are generated under the alternative, a less clear rejection pattern is to be expected here.

Hypothesis 2: MNAR vs MCAR

In a third simulated data example the initial data set is generated according to the same measurement model as in Section 7.5.3 but the dropout model is adjusted to

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -2 + 2Y_{i2}, \tag{7.11}$$

This hypothesis corresponds to setting 3 in Section 7.4. We generated 400 parametric bootstrap values and 400 semi-parametric bootstrap values. As in the previous section three different scenarios are considered. In Table 7.4, an overview of the critical values for $\alpha = 0.10, 0.05$ and 0.01 according to the generated null distribution, the different bootstrap methods and the $\chi^2(1)$-distribution can be found for each of the three scenarios. For Hypothesis 2, Table 7.4 shows that both bootstrap

Table 7.4: Hypothesis 2: Critical points based on the parametric and semi-parametric bootstrap procedure (400 bootstrap runs) for two initial data sets. Lower lines show the critical points of the simulated null distribution based on 800 samples, together with those of the $\chi^2(1)$ distribution.

		quantiles			p-value
		0.10	0.05	0.01	
Scenario 1	Parametric	4.08	5.31	9.76	0.3092
		5.63	7.50	11.33	0.9302
	Semi-Parametric	5.79	7.48	11.99	0.3791
		5.24	8.30	15.11	0.9352
Scenario 2	Parametric	6.11	8.44	15.4	0.0025
		5.84	7.84	12.04	0.0075
	Semi-Parametric	9.21	11.02	15.68	0.0075
		5.93	7.72	13.30	0.0075
Scenario 3	Parametric	4.41	5.81	9.58	0.5362
		4.47	6.23	10.14	0.1920
	Semi-Parametric	5.31	7.69	12.95	0.6234
		14.16	17.35	22.07	0.5586
simulated H_0		3.18	4.75	8.74	
$\chi^2(1)$ distribution		2.71	3.84	6.63	

methods provide higher critical values compared to the generated null distribution. The difference is rather small for Scenario 1 but increases for Scenario 2 and 3. The semi-parametric bootstrap method gives higher critical values compared to the parametric bootstrap. For Scenario 3, there is a substantial difference between both of them. Thus the bootstrap approaches are not able to regenerate the null distribution in this particular case.

Hypothesis 3: MNAR vs MAR

Testing whether dropout occurs randomly or non-randomly is the most interesting situation. In this setting, the initial data set is generated according to the following model

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}\right]. \tag{7.12}$$

and the dropout process is given by

$$\text{logit}[P(R_i = 1|Y_{i1}, Y_{i2})] = -2 + Y_{i1} + \psi_2 Y_{i2}, \tag{7.13}$$

We generated 400 parametric bootstrap values and 400 semi-parametric bootstrap values in the three different scenarios. In Table 7.5, an overview of the critical values for $\alpha = 0.10, 0.05$ and 0.01 according to the generated null distribution, the different bootstrap methods and the $\chi^2(1)$-distribution in all three scenarios can be found for two initial data sets. The results in Table 7.5 globally show that for testing MAR versus MNAR (Hypothesis 3), also in this setting the bootstrap is not able to approximate the true null distribution. Especially the behaviour of the parametric bootstrap is very unstable and variable. The semi-parametric version seems to slightly perform better, especially for Scenario 1. As the bootstrap is also an asymptotic method, it suffers from the same slow convergence as the χ^2-type distributions.

In Hypotheses 2 and 3 the dropout model depends upon unobserved data, resulting in a strange behaviour of both bootstrap methods, while for Hypothesis 1 this is not the case. Therefore the irregular behaviour of the bootstrap methods seem to be caused by the dependence of the dropout on the unobserved outcome.

Table 7.5: Hypothesis 3: Critical points based on the parametric and semi-parametric bootstrap procedure (400 bootstrap runs) for two initial data sets. Lower lines show the critical points of the simulated null distribution based on 800 samples, together with those of the $\chi^2(1)$ distribution.

		quantiles			p-value
		0.10	0.05	0.01	
Scenario 1	Parametric	38.71	42.68	46.21	0.1870
		9.62	12.25	19.77	1.000
	Semi-Parametric	4.86	6.74	10.48	0.0998
		7.40	9.35	14.47	0.2743
Scenario 2	Parametric	22.07	24.84	30.32	0.0349
		9.36	11.39	14.04	0.0025
	Semi-Parametric	12.35	15.63	20.88	0.0050
		17.05	19.75	27.56	0.0224
Scenario 3	Parametric	8.17	10.09	15.02	0.0175
		15.46	17.85	24.38	0.9351
	Semi-Parametric	15.68	19.11	25.58	0.1397
		8.11	10.46	13.55	0.6085
	simulated H_0	6.44	9.17	12.10	
	$\chi^2(1)$ distribution	2.71	3.84	6.63	

7.6 Discussion

To asses the sensitivity of conclusions to model choices in the context of selection models for non-random dropout, one can contrast the different missing mechanisms with each other; e.g., by the likelihood ratio tests. A simulation study was performed to examine the asymptotic null distribution under a variety of missingness mechanisms. Additionally, the behaviour of a parametric and a semi-parametric bootstrap approach was also investigated.

From literature and the simulation results, it is clear that the likelihood ratio test for testing missing not at random does not fulfil the regular assumptions. The use of classical asymptotic results might clearly lead to false results. The dependence on unobserved data seems to be responsible for this behaviour. Using bootstrap methods to generate the null distribution showed irregularities when testing for missing not at random.

Together with influence measures as the kernel weighted global and local influence derived in Chapter 6 it becomes clear that care has to be taken when modelling longitudinal data with missing values. A sensitivity analysis should not merely be mentioned as a possible tool to assess the sensitivity of the model but should be considered to be a part of the model building process (Jansen *et al.*, 2005).

Chapter 8

Modelling the Force of Infection for Clustered Binary Data with Missing Values

8.1 Introduction

Veterinary epidemiology is a research area that deals with the investigation of diseases in animal populations. Modelling infectious diseases is often confronted with key features such as clustering and stratification. Moreover, it is not unlikely that such data have missing values.

In practice, one often analyzes the complete cases, while ignoring the missingness mechanism. If data are missing completely at random, these complete cases can be analyzed as they are, but even then complete case analysis is non-efficient. Moreover, if this assumption is not fulfilled, as is frequently the case in practice, analyses can be affected by merely using the complete cases. Several methods to handle missing data are known (see Section 1.2). None of them are without limitations. One of them is multiple imputation (Rubin, 1978), where each of the gaps in the data are imputed several times and the analyses of the augmented data sets are then combined. However, in data with a mix of continuous and discrete variables, the choice of imputation model is non-trivial. Another technique is to weight a subject by the inverse of the probability that it is observed (see e.g. Robins *et al.*, 1994;

Zhao *et al.*, 1996). In this way subjects unlikely to be observed gain more weight. This can be seen as an implicit imputation of missing values. Both techniques are valid under the missing at random assumption.

The seroprevalence survey of the Bovine Herpesvirus-1 (BoHV-1) in Belgian cattle, as introduced in Section 1.5.4, is a study of a transmissible disease in cattle, which is of economic importance and significance to international trade. A central characteristic of infectious disease dynamics is the transmission of the infection from infectious to susceptible subjects. The force of infection (FOI) is the rate of acquisition of the infection for a susceptible host and can be interpreted as the instant probability to get infected, given that no infection has occurred before. Empirical data show that, in general, the FOI is age-dependent.

Under the assumptions of life long immunity and that the disease is in a steady state, the prevalence and FOI can be estimated from such seroprevalence data (Grenfell and Anderson, 1985). Parametric models for the prevalence and FOI of childhood infections, estimated from seroprevalence data, were discussed by Grenfell and Anderson (1985) who modelled the FOI with a polynomial function of host age. Other parametric models fitted within the framework of generalized linear models (GLM) with binomial error (McCullagh and Nelder, 1989) were discussed by Becker (1989), Diamond and McDonald (1992) and Keiding *et al.* (1996). They used the complementary log-log link in order to parameterize the prevalence and the FOI as a Weibull model. Becker (1989) suggested to model a piecewise constant FOI by fitting a model with a log link. In the case that other covariates, in addition to exposure time, are included in the model, Jewell and Van Der Laan (1995) proposed, for current status data, a proportional hazards model with constant FOI which can be fitted as a GLM with a complementary log-log link. Grummer-Strawn (1993) discussed two parametric models for current status data, the first one being a Weibull proportional hazards model with complementary log-log link and the second being the log logistic model with logit link function. For the latter, the proportionality in the model is interpreted as proportional odds. Farrington (1990) and Farrington *et al.* (2001) proposed a non-linear model for which the FOI is restricted to be non-negative and applied the model for measles, mumps and rubella. Shkedy *et al.* (2003, 2005) proposed to use local and fractional polynomials for the estimation of the prevalence and FOI.

Like many other infectious diseases data, the BoHV-1 data are complicated and thus statistical modelling has to deal with these complications. In this chapter, we model the FOI, while dealing with clustering, missing values, informative cluster size and the constraint for the FOI to be non-negative or equivalently the seroprevalence to be monotonically increasing (Hens *et al.*, 2005c).

In a first section, the basic SIR model (Susceptible, Infected, Recovered) is introduced. In Section 8.3, the FOI is formally introduced. To account for the clustering and missing values, a weighted flexible population-averaged model is introduced in Section 8.4. In Section 8.5 the modelling of the age-specific seroprevalence and the derivation of the age-specific FOI thereof is illustrated. The influence of ignoring missing values and thus merely using complete cases as such is addressed throughout these analyses. Section 8.6 gives a discussion on model building for the BoHV-1 data and we end with a general discussion in Section 8.7.

8.2 The Basic SIR Model

Mathematical modelling of infectious diseases involves describing the flow of individuals from different infection states within the population. For simple infectious diseases that simulate long-lasting immunity following infection, the individuals can be classified into three different states as shown in Figure 8.1 (Anderson, 1982; Anderson and May, 1991). In a first stage individuals are *susceptible to infection*, meaning that they have not been exposed yet. The number of hosts at risk at time t and age a is denoted by $X(a,t)$. In a second stage, individuals are *infected and infectious to others*. $Y(a,t)$ is the number of infected hosts at time t and age a. The third and last stage consists of individuals who are *immune to reinfection*. $Z(a,t)$ is the number of immune hosts at age a and time t. The total population is given by

$$N(a,t) = X(a,t) + Y(a,t) + Z(a,t). \tag{8.1}$$

The model described here is called a SIR model (Susceptible, Infected, Recovered). The SIR model relies on the assumption that newborns are entered directly into the susceptible class and infection, the infectious period and the disease occur simultaneously. Furthermore the SIR model ignores the latent period in which the individual is infected but not infectious to others. Figure 8.1 illustrates the basic SIR model.

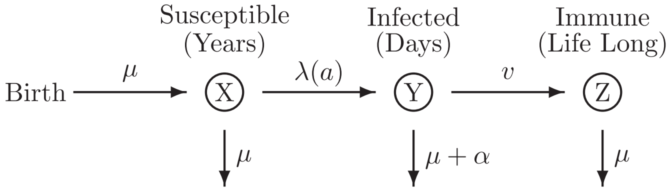


Figure 8.1: Illustration of the basic SIR model.

A central characteristic of the population dynamics of infection diseases is the transmission of the infection from the infected state to the susceptible state. Anderson (1982) and Anderson and May (1991) used a set of 3 partial differential equations to describe the flow of individuals within the population with respect to time and host age.

$$\begin{aligned} \tfrac{dX}{da} + \tfrac{dX}{dt} &= N\mu - [\lambda(a,t) + \mu]X(a,t), \\ \tfrac{dY}{da} + \tfrac{dY}{dt} &= \lambda X - (v + \alpha + \mu)Y(a,t), \\ \tfrac{dZ}{da} + \tfrac{dZ}{dt} &= vY - \mu Z(a,t). \end{aligned} \tag{8.2}$$

Here, N is the population size, μ is the natural rate of death ($1/\mu$ is the life expectancy), v is the recovery rate and α is the rate of death caused by the disease. $\lambda(a,t)$ is the FOI for age a at time t, i.e., the rate at which the host moves from the susceptible to the infected class. We refer to Shkedy (2003) for more details. It is often of interest to look upon the FOI as a function of age and time. Let us first derive the FOI in case of a generalized linear model assuming that the disease is in a steady state, i.e., time independent.

8.3 Force of Infection

Let $\pi(a) = \{Y(a) + Z(a)\}/N = 1 - X(a)/N$ be the probability to be infected before age a. In general, the seroprevalence $\pi(a)$ is modelled as

$$\pi(a) = g^{-1}(\eta(a)) = \delta(\eta(a)), \tag{8.3}$$

where $\eta(a)$ is the linear predictor and g is a link function. If it is assumed that the disease is in a steady state, then the age-dependent FOI, $\lambda(a)$, can be modelled according to equation (Anderson and May, 1991):

$$\frac{d}{da}q(a) = -\lambda(a)q(a), \tag{8.4}$$

with $q(a) = 1 - \pi(a)$. Indeed, in a steady state, the first equation in (8.2) simplifies to $\frac{dX}{da} = -\lambda(a)X(a)$. Using $q(a) = X(a)/N$ this becomes (8.4). The differential equation (8.4) describes the change in the fraction of susceptible individuals with the age of the host and so

$$\lambda(a) = \frac{\pi'(a)}{1 - \pi(a)}. \tag{8.5}$$

When a logit link is considered, the FOI can be expressed as:

$$\lambda(a) = \eta'(a)\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}. \tag{8.6}$$

In case that other covariates, except age, are included in the model, one can use the following structure for the linear predictor

$$g(\pi) = \eta(a) + \boldsymbol{\alpha} Z. \tag{8.7}$$

Here $\eta(a)$ is the fractional polynomial which is used to model the dependency of π and λ on age, Z is the design matrix for the additional covariates and $\boldsymbol{\alpha}$ is the parameter vector. Note that $\eta'(a)$ does not depend on $\boldsymbol{\alpha} Z$. Let us consider that Z is a binary predictor. For models with logit link we have

$$\frac{\lambda(a|z=1)}{\lambda(a|z=0)} = \frac{\eta'(a)}{\eta'(a)} \cdot \frac{\pi(a|z=1)}{\pi(a|z=0)} = \exp(\alpha) \cdot \frac{1-\pi(a|z=1)}{1-\pi(a|z=0)}. \tag{8.8}$$

The parameter α is in this case simply the log odds ratio. When Z is continuous, α is the log odds ratio for a unit change in Z.

For a model with complementary log-log link, i.e., a proportional hazard model,

$$\frac{\lambda(a|z=1)}{\lambda(a|z=0)} = \exp(\alpha), \tag{8.9}$$

i.e., α is the hazard ratio and can be seen as the relative FOI in our setting.

In the following section a weighted population-averaged model is introduced.

8.4 A Weighted Flexible Population-averaged Model

Once an infection is introduced in a herd, animals within the same herd have a high chance to get infected too. Thus, individual responses are more homogeneously distributed within herds than in the whole population. One cannot ignore the possibility of animals within herds to be more similar than between herds (Figure 8.2). There are several ways of dealing with such clustering (Aerts *et al.*, 2002c).

A first approach is to ignore the clustering. Let $Y_{ij}, j = 1, \ldots, n_i; i = 1, \ldots, K$, represent the binary response that equals 1 when the j-th animal of the i-th herd has antibodies to gB of BoHV-1, and 0 otherwise. Modelling the seroprevalence can be done by means of a logistic regression

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(\pi_{ij}), \\ \eta_{ij} &= \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = f(a_{ij}), \end{aligned} \tag{8.10}$$

where $f(a_{ij})$ is a functional form describing the dependency of the covariate of interest.

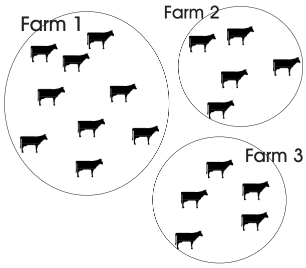


Figure 8.2: Clustering: animals within a farm (cluster) are more alike than between farms.

A logistic regression assumes that observations are independent and so the technique is not appropriate for clustered data (Figure 8.2). While logistic regression typically leaves the consistency of point estimation intact, the same is not true for measures of precision. In case of a 'positive' clustering effect (i.e., animals within a herd are more alike than between herds), then ignoring this aspect of the data will lead to overestimation of the precision and underestimation of standard errors and lengths of confidence intervals. Another strategy is to account for clustering, while the population mean is of major interest. This means that the existence of clustering is recognized but considered a nuisance characteristic. Generalized estimating equations (GEEs) can be used for this purpose. If one is interested in the clustering itself, one can use random-effects models. We restrict ourselves to GEEs and refer to Faes *et al.* (2005) for the random-effects approach.

8.4.1 Generalized Estimating Equations

Using GEEs, correlated binary data are modelled using the same link function and linear predictor setup (systematic component) as in the independence case (logistic regression). The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modelled.

Denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, the vector of measurements on the i-th cluster and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$, the corresponding vector of means. Let V_i denote the covariance matrix of $\boldsymbol{Y}_i$. Let the vector of explanatory variables for the j-th unit in the i-th cluster be denoted by $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$.

The GEE approach of Liang and Zeger (1986) for estimating the $p \times 1$ vector of regression parameters $\boldsymbol{\beta}$ is given by

$$S(\boldsymbol{\beta}, \boldsymbol{\phi}; R) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} {V_i}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}. \tag{8.11}$$

Since $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, where g is the link function, the $p \times n_i$ matrix of partial derivatives of the mean with respect to the regression parameters for the i-th cluster is given by

$$\frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{x_{i11}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_i1}}{g'(\mu_{in_i})} \\ \vdots & & \vdots \\ \frac{x_{i1p}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_ip}}{g'(\mu_{in_i})} \end{pmatrix}, \tag{8.12}$$

where $g'(\mu_{ij})$ denotes the derivative of g with respect to μ_{ij}. Let $R_i(\boldsymbol{\alpha})$ be an $n_i \times n_i$ 'working' correlation matrix that is fully specified by the vector of parameters $\boldsymbol{\alpha}$. The covariance matrix of $\boldsymbol{Y}_i$ is modelled as

$$V_i = \phi A_i^{\frac{1}{2}} R_i(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}, \tag{8.13}$$

where A_i is an $n_i \times n_i$ diagonal matrix with $v(\mu_{ij}) = \text{var}(\boldsymbol{Y}_{ij})$ as the j-th diagonal element. If $R_i(\boldsymbol{\alpha})$ is the true correlation matrix of $\boldsymbol{Y}_i$, then V_i is the true covariance matrix of $\boldsymbol{Y}_i$.

The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process using the current value of the parameter vector $\boldsymbol{\beta}$ to compute appropriate functions of the Pearson residual

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}.$$

If one specifies the working correlation by

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}, \tag{8.14}$$

then $R_i = I$, which is the identity matrix and the GEE reduces to the independence estimating equation. Several other correlation structures can be specified (Liang and Zeger, 1986). One interesting correlation structure is the exchangeable one, defined by

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}, \tag{8.15}$$

and this can be estimated by

$$\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} e_{ik}, \tag{8.16}$$

with $N^* = \sum_{i=1}^K n_i(n_i - 1)$.

The dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^2, \tag{8.17}$$

where $N = \sum_{i=1}^{K} n_i$ is the total number of measurements and p is the number of regression parameters. The square root of $\hat{\phi}$ is often called the scale parameter. The use of a dispersion parameter can be extremely useful in modelling residual overdispersion.

To estimate the covariance matrix of $\hat{\beta}$ one can use

$$\hat{V}_m = \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \, V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{-1}, \tag{8.18}$$

the so-called *model-based* covariance matrix. This matrix however is not a consistent estimator of the covariance matrix of $\hat{\beta}$ if the working correlation matrix is misspecified, that is, if $\text{Cov}(\boldsymbol{Y}_i) \neq V_i$. In that case one can use the robust or empirical estimator

$$\hat{V}_r = \hat{V}_m \cdot \left(\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \, V_i^{-1} \text{Cov}(\boldsymbol{Y}_i) \, V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right) \cdot \hat{V}_m, \tag{8.19}$$

which is a consistent estimator of $\text{Cov}(\hat{\boldsymbol{\beta}})$. An attractive point of the GEE approach is that it yields a consistent estimator of $\boldsymbol{\beta}$ even when the working correlation matrix is misspecified (Liang and Zeger, 1986). Zeger *et al.* (1988) and McDonald (1993) have shown that in the case of a working independence model, $R = I$, which is often convenient, $\hat{\boldsymbol{\beta}}$ is relatively efficient at least when the correlation between responses is not large. In the next section, weighted GEEs are introduced to deal with missing data.

8.4.2 Inverse Probability Weighted GEE

One of the techniques to deal with data which are missing at random, that gained a lot of attention, is the 'weighted estimating equation' (Robins *et al.*, 1994; Zhao *et al.*, 1996), where each contribution of a case is weighted with the inverse of the probability that this case is observed as introduced in Section 1.2.1. In this way cases with a low probability to be observed gain more influence, resulting in an implicit imputation of missing values. The generalization towards GEEs is straightforward:

$$S_w(\boldsymbol{\beta}, \boldsymbol{\phi}, R) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \, V_i^{-1} \, W_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}, \tag{8.20}$$

where W_i is a $n_i \times n_i$ diagonal matrix with elements w_{ij} equal to the inverse probability for the j-th unit in the i-th cluster to be observed, $i = 1, \ldots, K$; $j = 1, \ldots, n_i$. This probability is preferably estimated non-parametrically (Wang *et al.*, 1998), by, e.g., a generalized additive model (Hastie and Tibshirani 1990, Section 1.3.1). Let us denote $\hat{\boldsymbol{\beta}}_w$ as the solution to (8.20).

8.4.3 Fitting a Flexible Model

We will use fractional polynomials, as introduced in Section 1.3.1, to model the relationship between the seroprevalence and age. In this way a flexible parametric model is provided (Royston and Altman, 1994). Fractional polynomials were used before by Shkedy *et al.* (2005) and Faes *et al.* (2005) in modelling infectious diseases and correlated animal data.

The use of splines could offer a fully non-parametric alternative to the use of fractional polynomials. However, a fractional polynomial offers a simpler derivation of the FOI and permits constrained optimization. An appealing feature of fractional polynomials is that they, as a parametric tool, offer a wide range of flexible functional forms and that they include the conventional polynomials, often used in practice.

Let us now formulate the AIC and AIC$_W$-criterion to select the appropriate model and the appropriate powers of the fractional polynomial in a unweighted and weighted logistic regression setting.

Model Selection

In a logistic regression setting, the Akaike Information Criterion is given by

$$\text{AIC} = -2\sum_{i=1}^{n} y_i \log(\hat{\pi}_i) + (1 - y_i)\log(1 - \hat{\pi}_i) + 2p, \tag{8.21}$$

where p is the number of regression parameters and $\pi = \pi(\boldsymbol{\beta})$ is the probability for Y to be 1.

Recalling the weighted model selection criteria presented in Chapter 3, the weighted version of (8.21) is given by

$$\text{AIC}_W = -2\sum_{i=1}^{n} w_i \left(y_i \log(\hat{\pi}_{w,i}) + (1 - y_i)\log(1 - \hat{\pi}_{w,i})\right) + 2p, \tag{8.22}$$

where $\hat{\pi}_{w,i}$ denotes the estimated probability for the i-th unit based on a weighted logistic regression with weights w_i, $i = 1, \ldots, n$. When $w_i = 1$ for all $i = 1, \ldots, n$, this criterion reduces to its unweighted version.

8.5 Modelling the FOI for the BoHV-1 Study

In the Bohv-1 data not only clustering and missing values complicate analyses but also the informative cluster size and monotonicity constraints have to be dealt with. Let us give an overview of the different complications and the way we will handle these to model the FOI for the BoHV-1 study.

8.5.1 Overview of the Methods

To deal with clustering and missingness, we propose to use constrained weighted GEEs (8.20), where the weights are the inverse probability for the animal to be observed. We will use a fractional polynomial to model the dependency of the test result, the presence of antibodies, with age. Let us point out how the informative cluster size and monotonicity constraints are dealt with.

Faes *et al.* (2005) showed that the herd size in the BoHV-1 data is informative, i.e., the herd size is related with the outcome of interest. When dealing with an informative cluster size, one can be interested in the probability of a randomly sampled unit from all units or in the probability of a randomly sampled unit from a randomly selected cluster. In the GEE approach, the correlation between cluster members is modelled in order to determine the weight that should be assigned to the data from each cluster. If interest goes out to a randomly selected unit from all units, one can use the working independence correlation. If interest goes out to a randomly sampled unit from a randomly selected herd and the cluster size is not related to the outcome, the same analysis will be valid and the same asymptotic parameter estimates will be obtained (Williamson *et al.*, 2003). However, when the cluster size is related to the outcome, the latter analysis is not valid anymore. Williamson *et al.* (2003) proposed to use weighted GEEs where the weights equal the inverse of the cluster size. In this way subject-specific weights turn into cluster-specific weights. The motivation of this method is the same as when dealing with design-based samples (Section 3.3.3). Faes *et al.* (2005) proposes an alternative method where the cluster size is incorporated as a categorized covariate. This method facilitates to look upon the FOI from a herd-specific point of view (Section 8.5.3). Following these strategies, we obtain two approaches to deal with an informative cluster size in an inverse probability weighted GEE. A first approach is to use weights that provide a correction for both the informative cluster size and the occurrence of missing values by multiplying the inverse cluster size and the inverse probability for an observation to be observed. A second approach, which we will use, includes herd size as a covariate in the model, correcting for the informative cluster size, and uses weights equal to the inverse probability for an observation to be observed to correct for missing values.

The final complication to be dealt with is that the FOI as a function of age cannot be negative and thus the age-specific prevalence has to be monotone increasing. Determining $\hat{\boldsymbol{\beta}}_w$ is therefore subject to constraints depending on the functional relationship with age. Selecting a model when dealing with constrained parameters is not straightforward either. A modified version of the AIC-criterion when dealing with a parameter that is restricted to be in a range $[a, b]$ has been proposed by Hossain (2002). This MAIC-criterion uses a penalization which is 1/2 instead of 1 for a parameter on the border of the range. When the parameter lays in the range $[a, b]$ the MAIC-criterion equals the AIC-criterion. Equivalently one can think of a modification of the AIC-criterion when dealing with more general constraints of the form

$$C(\boldsymbol{\beta}) \geq 0, \tag{8.23}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$. This however is a topic of further research and will not be pursued in this thesis.

Recently a lot of attention has gone out to model selection for GEEs. The work of Pan (Pan, 2001a,b) is key in this context. However, using these criteria when dealing with constraints requires even more caution because of the additional selection of an appropriate correlation structure. It is not clear what the effect is of constraints on the estimation of the variance.

In the following sections the FOI for the BoHV-1 data is derived. We will select the appropriate model by first selecting the appropriate constrained (weighted) logistic regression model using the (weighted) AIC-criterion. The selected model will then be fitted using a constrained (weighted) GEE in order to obtain a more honest estimate of the variability. Let us first study the missingness in the Bohv-1 data.

8.5.2 Missing Data in the BoHV-1 Data

From the 11284 records, 2148 records have at least one missing value in response and covariates. In Table 8.1, the specific amount of missingness for each variable is given. From this table, it is clear that the only variable with a substantial amount of missingness is ‘purchase’. Therefore, the remainder of this chapter observations with one or more missing values for ‘age’, ‘sex’, and ‘gB’ are ignored. The purchase-missing values were caused by a technical problem while conducting the survey; for animal-level identification, the animals’ working eartag numbers were noted, not their official ones. The advantage of the former ones is higher readability. Unfortunately, these working eartag numbers were not indexed. To asses the influence of

Table 8.1: BoHV-1 data: Overview of the amount of missingness for each variable. Variables without any missing values were omitted.

Variable	Miss. (#)	Miss. (%)
Age	26	0.23 %
Sex	14	0.12 %
Purchase	2091	19.00 %
gB	36	0.32 %

the different variables on the missingness of 'purchase', we use a generalized additive model as proposed by Wood and Augustin (2002) to estimate the probability, π^o, for an observation to be observed. Starting from the generalized additive model (8.24), we apply the 3-step ad hoc method proposed by Wood and Augustin (2002) to drop terms (see Section 4.4).

$$\begin{aligned} \text{logit}(\pi^o) &= \beta_0 + f_{c_1}(\text{herdtype}) + f_{c_2}(\text{gB}) + f_{c_3}(\text{sex}) + f_{c_4}(\text{province}) \\ &+ f_{s_1}(\text{age}) + f_{s_2}(\text{herd size}) + f_{s_3}(\text{densanim}) + f_{s_4}(\text{densherd}), \end{aligned} \tag{8.24}$$

where $f_{c_i}(\cdot)$ denotes a main effect of a categorical variable and $f_{s_i}(\cdot)$ denotes a smooth function. In Figure 8.3, an overview of the smooth terms together with 95% confidence intervals is shown (R package mgcv 1.1-8). Based on Figure 8.3, Table 8.2 and the fact that all categorical variables contributed significantly to the model, no term could be dropped from the model. In practice, one could think of using surface smoothers, tensor product smoothers and category-specific smoothing to include interactions. For a large dataset as the BoHV-1 this was computationally not feasible.

Table 8.2: BoHV-1 data: Overview of the missing data modelling result.

Variable	Estimated df.
herd size	8.01
age	8.15
densanim	8.54
densherd	8.65

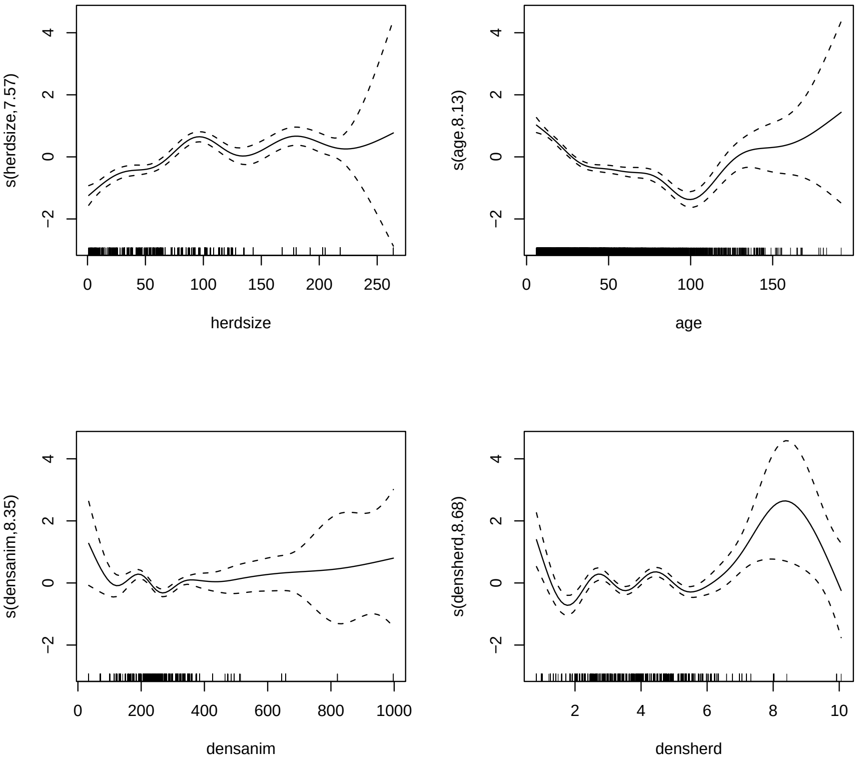


Figure 8.3: Missingness in the BoHV-1: GAM-plots for the continuous variables.

To illustrate the effect of ignoring missing data, we compare the analysis based on the complete cases, i.e., cases for which the ‘purchase’-variable is observed with the analysis based on the available cases, i.e., cases for which ‘purchase’ can be observed or unobserved and show that a weighted analysis on the complete cases can be used to correct for the missing values. The animal-specific weight is the inverse of the estimated probability that the animal is observed, i.e., all characteristics for that animal are observed. The latter probability is derived from model (8.24).

In Figure 8.4, the fraction of positive tests for the antibodies as a function of age based on the available and on the complete cases is shown. To distinguish between animals coming from herds with different sizes, each of the plots is an overlay of seroprevalence plots for animals with herd size lower or equal to 30 (circles), between 30 and 60 (stars) and higher or equal to 60 (triangles), respectively. Figure 8.4 shows that the seroprevalence for animals from larger herds is higher. The seroprevalence

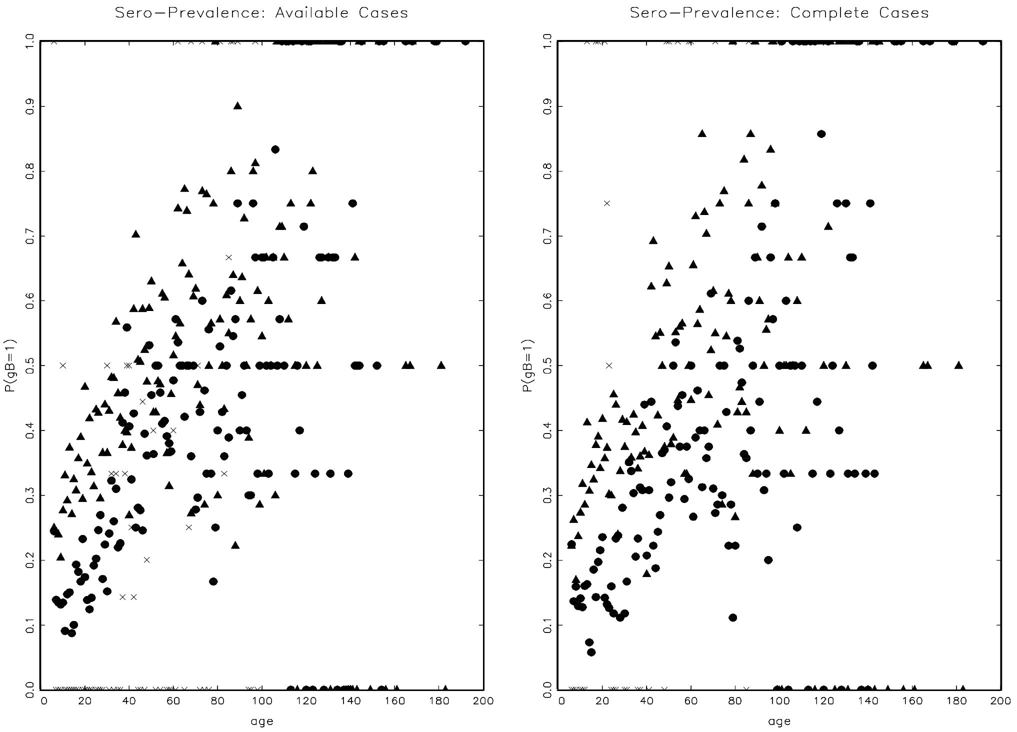


Figure 8.4: Seroprevalence scatterplot as a function of age based on the available cases (left) and on the complete cases (right) for small (circles), medium (stars) and large herds (triangles).

plot for the complete cases differs slightly from the one for the available cases, e.g., the seroprevalence, based on the available cases, shows less variability over the different herd sizes. We will show that merely using the complete cases can result in a wrong model and thus has its effect on the FOI.

8.5.3 Constrained Logistic Regression

Let us start from a logistic regression with a fractional polynomial of age to model the age-specific seroprevalence, subject to the constraint of monotonicity and based on three different methods. The first method is based on the available cases (AC), the second on the complete cases (CC) and the third method uses a weighted logistic regression based on the complete cases (WCC) where the subject-specific weight equals the inverse probability for that subject to be observed as estimated by the generalized additive model (8.24). Contreras and Ryan (2000) give an overview of optimization software to fit non-linear and constrained GEEs. We used the 'Constrained

Optimization'-module in Gauss 6.0. The procedure uses a sequential quadratic programming method in combination with the Newton-Raphson procedure. In an initial stage, the Broyden-Fletcher-Goldfarb-Shanno procedure (Shanno, 1985) was used to obtain starting values for the Newton-Raphson procedure.

As pointed out in Section 8.4.3, we will use fractional polynomials of degree 2 to describe the dependency of seroprevalence on age. Because of the informative herd size, we included herd size as a main effect in the constrained logistic regression model

$$\text{logit}(P(gB = 1)) = \beta_0 + \beta_1 \text{age}^{p_1} + \beta_2 \text{age}^{p_2} + \beta_3 \text{herd size}. \tag{8.25}$$

The appropriate powers of the fractional polynomial, $p_1, p_2 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, $p_1 \leq p_2$, were determined by minimizing the AIC (available cases, complete cases) and AIC_W-criterion (weighted complete cases). Since the number of parameters in these models stays the same, this corresponds with the deviance criterion used by Royston and Altman (1994). In Table 8.3, an overview of the different models is given together with the AIC- and AIC_W-values. The powers, parameters and standard errors for the three different methods can be found in Table 8.4. The results in this table are difficult to compare since the different methods selected different fractional polynomials.

As an illustration, Figure 8.5 shows the resulting seroprevalence curves together with the FOI for herd sizes 15, 45, 80, and 120, representing small, medium, large and very large herds, respectively.

Figure 8.5 indicates an improved seroprevalence fit when using weighted complete cases instead of complete cases only, especially for larger herd sizes. The FOI-curves show that using the weighted complete cases give a substantial correction compared to using the complete cases only. The latter finding can be translated in that the curvature of the seroprevalence curves for both the available and weighted complete cases are quite similar.

In practice, interest often goes out to the age at which the maximal FOI is reached, age_{max}. In Table 8.5, the $\widehat{\text{age}}_{\text{max}}$ is shown for four herd sizes 15, 45, 80 and 120, representing small, to large-sized farms. Using the complete cases only, age_{max} is severely overestimated with 10 to 17 months compared to $\widehat{\text{age}}_{\text{max}}$ based on the available cases, while the use of weights gives a slight underestimation of age_{max} with about 2.5 months.

The results, using all methods (CC,AC,WCC), show that the age at which the FOI reaches its maximum value decreases with herd size, i.e. the cluster size (see Table 8.5). In Figure 8.6, the age-specific FOI for the available cases is shown for herd sizes 15, 45, 80 and 120. An increasing herd size corresponds to an increasing

Table 8.3: BoHV-1 data: Overview of the AIC- and AIC$_W$-values for different choices of powers for the logistic regression model described by (8.25).

Powers			Selection		Powers			Selection	
p_1	p_2	AIC(CC)	AIC(AC)	AIC$_W$	p_1	p_2	AIC(CC)	AIC(AC)	AIC$_W$
-2.0	-2.0	10025.63	12756.67	12902.12	-0.5	1.0	9961.00	12710.29	12855.05
-2.0	-1.0	9966.01	*12704.41*	12853.35	-0.5	2.0	9965.11	12711.64	12856.13
-2.0	-0.5	9958.61	12705.29	*12852.77*	-0.5	3.0	9969.04	12712.56	12857.08
-2.0	0.0	10073.98	12815.65	12958.77	0.0	0.0	9958.36	12708.29	12853.86
-2.0	0.5	9959.91	12714.68	12857.97	0.0	0.5	9958.90	12708.03	12853.75
-2.0	1.0	9965.82	12722.31	12863.75	0.0	1.0	9959.44	12707.80	12853.66
-2.0	2.0	9984.75	12741.13	12880.05	0.0	2.0	9960.36	12707.57	12853.63
-2.0	3.0	10007.03	12760.48	12898.59	0.0	3.0	9960.98	12707.70	12853.75
-1.0	-1.0	9980.96	12715.20	12861.18	0.5	0.5	9958.53	12714.01	12859.39
-1.0	-0.5	*9956.62*	12706.74	12853.18	0.5	1.0	9958.17	12707.08	12853.65
-1.0	0.0	10007.60	12738.87	12883.96	0.5	2.0	9957.66	12712.22	12858.32
-1.0	0.5	9959.53	12711.15	12855.54	0.5	3.0	9957.81	12717.54	12863.18
-1.0	1.0	9962.67	12713.83	12857.42	1.0	1.0	9961.79	12731.86	12876.85
-1.0	2.0	9971.22	12719.41	12862.06	1.0	2.0	9959.54	12728.03	12873.76
-1.0	3.0	9980.40	12724.45	12866.98	1.0	3.0	9964.93	12742.16	12887.26
-0.5	-0.5	9967.59	12708.49	12854.24	2.0	2.0	10003.66	12809.93	12953.45
-0.5	0.0	9979.40	12714.73	12860.28	2.0	3.0	10003.10	12811.59	12955.43
-0.5	0.5	9959.24	12710.07	12854.50	3.0	3.0	10079.35	12916.32	13058.69

Table 8.4: Logistic Regression: Maximum likelihood parameter estimates using fractional polynomials with powers (p_1, p_2) for the three different methods: complete cases (CC), available cases (AC) and weighted complete cases (WCC).

Parameter	CC $(-1, -0.5)$	AC $(-2, -1)$	WCC $(-2, -0.5)$
Intercept	2.640(0.410)	0.410(0.113)	1.726(0.205)
age^{p_1}	6.638(1.383)	5.321(1.039)	1.813(0.658)
age^{p_2}	-10.969(1.537)	-7.095(0.726)	-5.215(0.506)
herd size	0.008(5.0e-4)	0.007(4.0e-4)	0.004(5.1e-4)

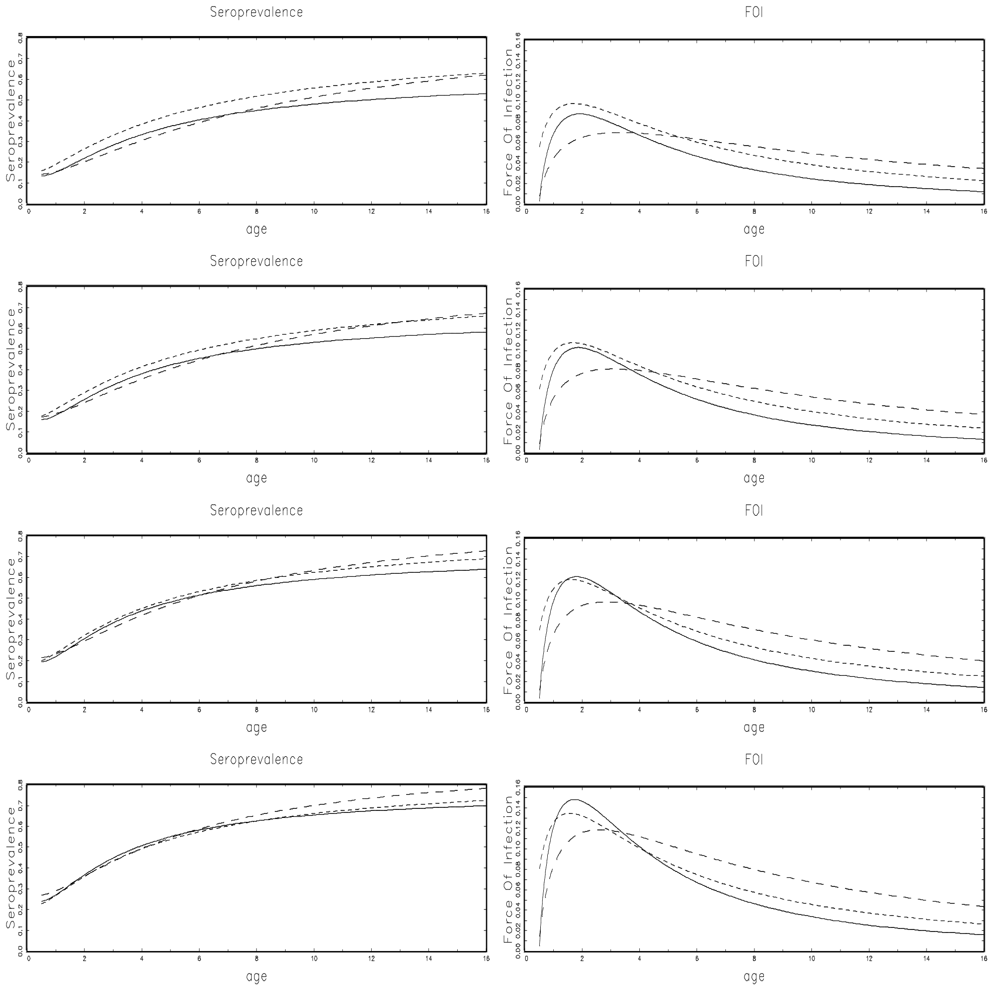


Figure 8.5: Age-specific seroprevalence curves together with the age-specific FOI for the available cases (solid curve), the complete cases (long dashed curve) and weighted complete cases (short dashed curve) for herd sizes 15, 45, 80 and 120.

FOI. That was expected from a veterinary and epidemiological point of view since animals in a larger herd have a higher probability to get infected at younger age.

Looking at the seroprevalence and FOI from a different angle, Figure 8.7 shows the 'herd size'-specific curves for animals at the age of 30, 90 and 180 months. Both the seroprevalence- and FOI-curves show a positively-related herd size. For animals with age larger than $\widehat{\text{age}}_{\text{max}}$ (see Figure 8.7), one observes a positive effect of age on the seroprevalence and a negative effect of age on the FOI. This was observed before in Figure 8.5, where the seroprevalence increases and the FOI decreases with

Table 8.5: Age (in years) where the maximal FOI is reached for herd size 15, 45, 80 and 120 for the three different methods.

Herd size	CC	AC	WCC
15	3.32	1.91	1.72
45	3.09	1.86	1.67
80	2.85	1.80	1.62
120	2.60	1.73	1.56

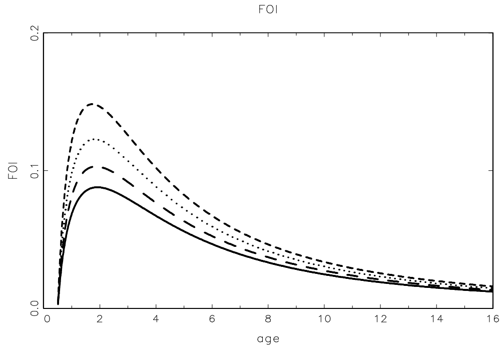


Figure 8.6: The age-specific FOI for herd sizes 15 (solid curve), 45 (long dashed curve), 80 (dotted curve) and 120 (short dashed curve) using the available cases.

age larger than $\widehat{\text{age}}_{\text{max}}$. Similarly, for animals at an age lower than $\widehat{\text{age}}_{\text{max}}$, the seroprevalence- and FOI-curve would both show a positive effect of age.

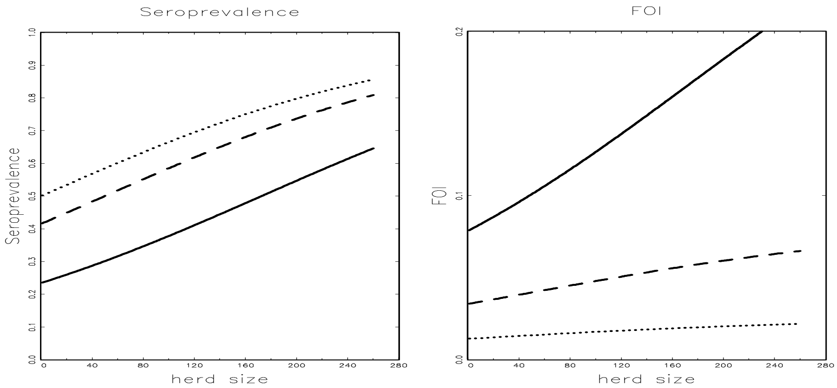


Figure 8.7: The herd size-specific seroprevalence (left panel) and FOI (right panel) at the age of 30 months (solid curve), 90 months (dashed curve) and 180 months (dotted curve) using the available cases.

Since the use of a logistic regression analysis does not take into account the clustering, the use of GEEs to model the seroprevalence is provided in the next section.

8.5.4 Constrained Generalized Estimating Equations

While the use of a logistic regression to model clustered binary data typically leaves the consistency of point estimation intact, precision is overestimated in case of a "positive"clustering effect (i.e., animals within a herd are more similar than between herds). The use of GEEs accounts for the correlations in the data in a manner that clustering is considered to be a nuisance parameter.

Selecting the powers of the fractional polynomials for the GEE (WGEE) with independence working correlation matrix is done using AIC (AIC_W) and the constrained (weighted) logistic regression for reasons pointed out before (Section 8.5.1).

Table 8.6: GEE parameter estimates, standard errors and corresponding p-values for the three different methods.

Parameter	Estimate	Emp.S.E.(p-value)	Mod. S.E.(p-value)
	Complete Cases		
Intercept	2.640	0.888(0.003)	0.410(<0.001)
age^{-1}	6.638	2.935(0.024)	1.383(<0.001)
age$^{-0.5}$	-10.969	3.312(0.001)	1.537(<0.001)
herd size	0.008	0.004(0.046)	5.0e-4(<0.001)
	Available Cases		
Intercept	0.410	0.304(0.177)	0.113(<0.001)
age^{-2}	5.321	2.238(0.017)	1.039(<0.001)
age^{-1}	-7.095	1.661(<0.001)	0.726(<0.001)
herd size	0.007	0.003(0.020)	4.0e-4(<0.001)
	Weighted Complete Cases		
Intercept	1.726	0.563(0.002)	0.205(<0.001)
age^{-2}	1.813	1.967(0.357)	0.658(0.006)
age$^{-0.5}$	-5.215	1.571(0.001)	0.506(<0.001)
herd size	0.004	0.005(0.424)	5.1e-4(<0.001)

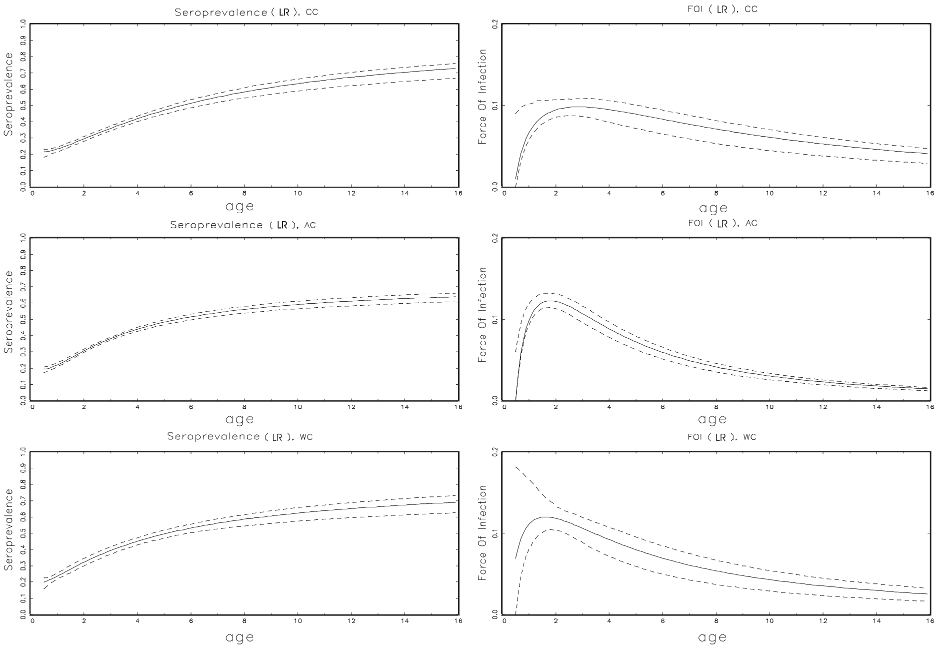


Figure 8.8: Herd size 80: Age-specific seroprevalence curves and corresponding FOI with 95% bootstrap confidence intervals using a logistic regression.

In Table 8.6, the parameter estimates and standard errors are tabulated for the three different methods. There is a positive effect of herd size and the components of the fractional polynomial counteract. Taking into account the clustering effect has a substantial impact as can be seen from the difference between the empirical standard errors, i.e., taking into account clustering and the model-based standard errors, i.e., ignoring clustering.

Calculating 95% confidence bounds for a constrained (weighted) logistic regression or constrained (weighted) GEE is not straightforward, since these bounds are typically not symmetric due to the constraint(s). We will use two bootstrap techniques to produce these confidence intervals.

A first technique was used to generate bootstrap confidence intervals for the constrained logistic regression. We refer to Davison and Hinkley (1997) for more details about bootstrap based confidence intervals. It consists of three successive steps:

(1) Resample animals with equal probabilities,

(2) fit model (8.25) using constrained (weighted) logistic regression to the resampled data, while keeping (p_1, p_2) fixed and recalculating the herd size,

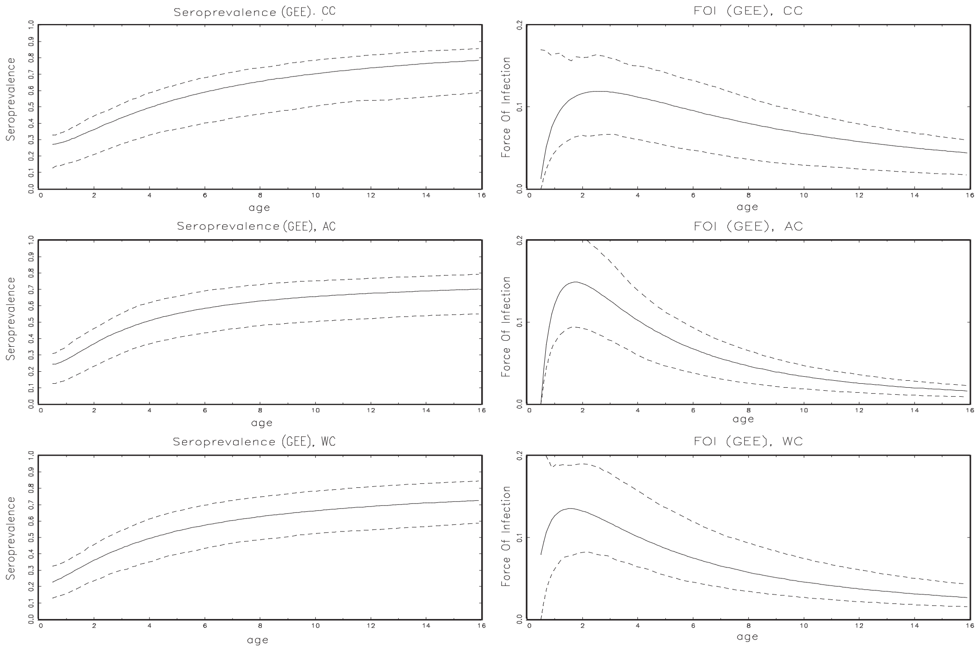


Figure 8.9: Herd size 80: Age-specific seroprevalence curves and corresponding FOI with 95% bootstrap confidence intervals using GEEs.

(3) calculate age-specific fitted values.

These steps were repeated 400 times and bootstrap confidence intervals were calculated using the age-specific 2.5% and 97.5% percentiles points. In Figure 8.8, seroprevalence and FOI curves are shown together with 95% pointwise bootstrap confidence intervals for herd size 80 based on the complete cases, available cases and weighted complete cases, respectively. Similar plots were obtained for other herd sizes.

A second technique was used to generate bootstrap confidence intervals for the constrained GEE. It consists of the following three successive steps:

(1) Resample herds with equal probabilities,

(2) fit model (8.25) using a constrained (weighted) GEE with independence working correlation to the resampled data, while keeping (p_1, p_2) and herd sizes fixed,

(3) calculate age-specific fitted values.

Again these steps were repeated 400 times and bootstrap confidence intervals were calculated using the age-specific 2.5% and 97.5% percentiles points. By resampling herds instead of animals clustering is taken into account. This non-parametric bootstrap procedure is based on Moulton and Zeger (1989); Sherman and Le Cessie (1997), who applied the bootstrap in a repeated measures context. In Figure 8.9, seroprevalence and FOI curves are shown together with 95% pointwise bootstrap confidence intervals for herd size 80 based on the complete cases, available cases and weighted complete cases, respectively.

From these confidence intervals it is clear that clustering has an impact on the variability and thus should be taken into account. Bootstrap procedures for correlated binary data were applied before by Gemechis and Aerts (2004).

8.6 Extending the Model

To extend model (8.25) and maintain flexible modelling for age, one could incorporate all other variables in the model as main effects. Additionally, one could add all two-way interactions, quadratic effects, three-way interactions and so on. For continuous variables, taking an interaction with the fractional polynomial of age would imply taking interactions with each of the components (when $m = 2$). For categorical variables, one could fit a fractional polynomial of age for each of the categories. Selecting the appropriate powers by means of a selection criterion applied on a two-dimensional grid for the powers (p_1, p_2) would result in an enormous number of candidate models to be considered, e.g., for m=2, including variables age, sex, purchase, herdtype and herd size and using a grid $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}^2$, $244,903$ models have to be fit. The monotonicity constraint makes the model building process very computer intensive and time consuming. There is currently no software which allows these features to be combined, while keeping the computing time acceptable. On average one such analysis, using Gauss 6.0 on an PIV (512 MB Ram, 2.6 Ghz), runs 181 seconds.

In what follows we first consider a classification tree analysis, to gain some insight in the relation between the seroprevalence and the explanatory variables (Table 1.2).

8.6.1 A Classification Tree Analysis

Although a classification tree is a fully non-parametric technique to model binary data (Section 1.3.1), it is not suitable to derive the FOI, but can provide insights in the relations between the different variables in the data.

A classification tree analysis obtains an optimal tree of size 189 when applying the 1 SE-error rule (Figure 8.10). All variables were used in the tree-construction. A closer look at the cross-validation relative error in Figure 8.10 shows minor de-

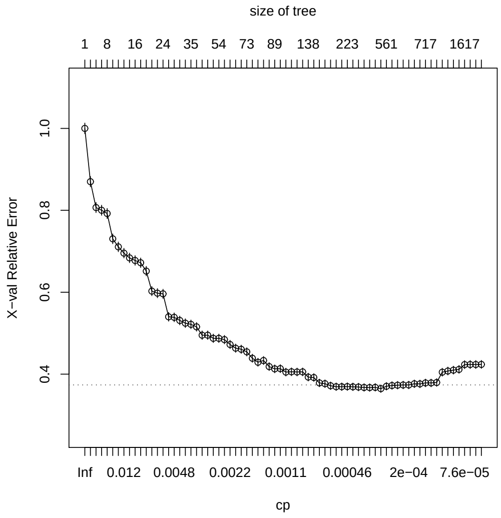


Figure 8.10: Classification tree analysis of the BoHV-1 data: X-validation relative error.

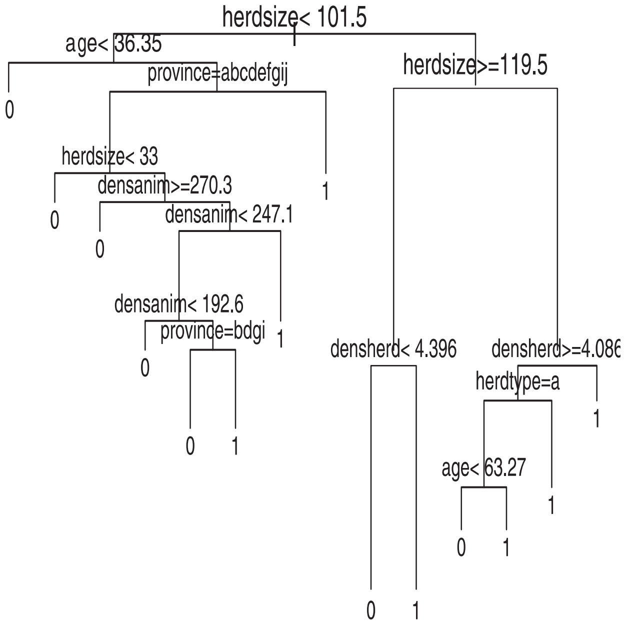


Figure 8.11: Classification tree analysis of the BoHV-1 data: A subtree of size 27.

creases in the cross-validation relative error based on trees of size larger than 27. We therefore restrict ourself to a tree of size 27. As an illustration, Figure 8.11 shows this subtree of size 27. All variables except sex and herdtype are shown in this tree.

This classification tree analysis confirms the difficulties of model building in this specific situation.

Although model building is hard, we will consider two models. The first model is an additive model consisting of a fractional polynomial of degree 2 for age and all other variables as main effects. Model 2 illustrates the use of interactions by considering an additive model consisting of a fractional polynomial of degree 2 for age, a main effect for herd size and a main effect of purchase together with the interaction of purchase with the fractional polynomial and herd size. The analyses presented are based on both complete and weighted complete cases, to illustrate the impact of ignoring missing observations.

8.6.2 Model 1: An Additive Model

Let us start from an additive logistic regression model

$$\begin{aligned}\text{logit}(P(\text{gB}=1)) &= \beta_0+\beta_1\text{age}+\beta_2\text{herd size}+\beta_3\text{herdtype}+\beta_4\text{densanim}\\ &\quad +\beta_5\text{densherd}+\beta_6\text{province}+\beta_7\text{sex}+\beta_8\text{purchase}, \qquad (8.26)\end{aligned}$$

and perform a stepwise deletion. Looking at the correlations among the different explanatory variables, the animal density and herd density have the highest correlation of about 0.48, however after some investigation no multicollinearity was established. Selecting the appropriate model is done by the use of the (weighted) AIC-criterion. In Table 8.7, the submodels of (8.26) are given together with their (weighted) AIC-value. Deletion stops when the (weighted) AIC-values of the submodels are all larger than the model under consideration. For both the complete and weighted complete cases, the model

$$\begin{aligned}\text{logit}(P(\text{gB}=1)) &= \beta_0+\beta_1\text{age}+\beta_2\text{herd size}+\beta_3\text{densanim}\\ &\quad +\beta_4\text{densherd}+\beta_5\text{province}+\beta_6\text{sex}+\beta_7\text{purchase}, \qquad (8.27)\end{aligned}$$

has the minimal AIC-, AIC$_W$-value.

This final constrained logistic regression model can be altered to include a fractional polynomial of degree 2 for age. The powers (p_1, p_2) from the grid $\{-2,-1,-0.5,0,0.5,1,2,3\}^2$ are determined by selecting the model with minimal AIC-, AIC$_W$-value under constraints. Recall the equivalence with the deviance criterion used by Royston and Altman (1994). A constrained logistic regression with powers $(p_1, p_2) = (-1, -0.5)$ on the complete cases resulted in a minimal AIC of 9509.47 while on the weighted complete cases a minimal (weighted) deviance of

Table 8.7: Constrained additive logistic regression models for the BoHV-1 data based on complete cases and weighted complete cases together with their AIC and AIC$_W$-values.

	CC	WCC
Model	AIC-value	AIC$_W$-value
Initial Model (8.26)	9554.28	12332.37
- province	9916.76	12794.52
- herdtype	9554.16*	12330.36*
- densherd	9564.91	12352.31
- densanim	9558.57	12356.38
- sex	9572.57	12347.73
- purchase	9573.78	12405.02
- herd size	9736.46	12440.69
- herdtype,province	9914.84	12794.46
- herdtype,densherd	9564.49	12355.46
- herdtype,densanim	9560.61	12350.29
- herdtype,sex	9571.47	12345.94
- herdtype,purchase	9572.25	12405.72
- herdtype,herd size	9748.89	12443.73

12248.84 was obtained for the powers $(p_1, p_2) = (-2, -0.5)$. These powers are the same as those found for model (8.25). The summary of the final models using (weighted) GEEs; i.e., powers, estimates, empirical and model-based standard errors with corresponding p-values, is given in Tables 8.8 and 8.9. The estimates and model-based standard errors correspond to using a (weighted) logistic regression while the empirical standard errors reflect the effect of clustering.

While there is a clear difference between the empirical and model-based standard errors, reflecting the clustering in the data, this has little impact on the significance (α-level 0.05) of the different covariates. Comparing the weighted complete case analysis with the unweighted complete case analysis there is some difference between the different estimates but little difference between the effects of the different variables. From these analyses, one can conclude that purchased animals have a

Table 8.8: Final additive models for the BoHV-1 data: complete cases.

	GEE (Independence)		
Parameter	Estimate	Emp.S.E.(*p*-value)	Mod.S.E.(*p*-value)
Complete Cases			
Intercept	1.714	0.477(<0.001)	0.001(<0.001)
age^{-1}	7.219	1.438(<0.001)	0.003(<0.001)
$\text{age}^{-0.5}$	-12.405	1.593(<0.001)	0.003(<0.001)
herd size	0.009	4.3e-6(<0.001)	6.7e-4(<0.001)
purchase	0.259	0.058(<0.001)	1.7e-4(<0.001)
sex	0.486	0.088(<0.001)	1.8e-4(<0.001)
densanim	0.001	3.5e-4(0.004)	2.0e-6(<0.001)
densherd	-0.090	0.029(0.002)	1.5e-4(<0.001)
province (ref.cat. Namur)			
- Antwerp	1.747	0.240(<0.001)	0.001(<0.001)
- Brabant	0.178	0.273(0.514)	0.001(<0.001)
- West Flanders	1.476	0.236(<0.001)	0.001(<0.001)
- East Flanders	1.745	0.238(<0.001)	0.001(<0.001)
- Hainaut	1.454	0.233(<0.001)	0.001(<0.001)
- Liège	0.818	0.234(<0.001)	0.001(<0.001)
- Limburg	1.983	0.244(<0.001)	0.001(<0.001)
- Luxembourg	0.370	0.255(0.1468)	0.001(<0.001)

Table 8.9: Final additive models for the BoHV-1 data: weighted complete cases.

	GEE (Independence)		
Parameter	Estimate	Emp.S.E.(p-value)	Mod.S.E.(p-value)
Weighted Complete Cases			
Intercept	0.881	0.270(0.001)	0.001(<0.001)
age^{-2}	2.544	0.682(<0.001)	0.002(<0.001)
age$^{-0.5}$	-6.571	0.527(<0.001)	0.001(<0.001)
herd size	0.006	5.7e-4(<0.001)	4.2e-6(<0.001)
purchase	0.422	0.050(<0.001)	2.1e-4(<0.001)
sex	0.469	0.077(<0.001)	1.7e-4(<0.001)
densanim	0.002	3.2e-4(<0.001)	1.9e-6(<0.001)
densherd	-0.107	0.026(<0.001)	1.4e-4(<0.001)
province (ref.cat. Namur)			
- Antwerp	1.371	0.183(<0.001)	0.001(<0.001)
- Brabant	0.004	0.211(0.985)	0.001(<0.001)
- West Flanders	1.485	0.176(<0.001)	0.001(<0.001)
- East Flanders	1.515	0.181(<0.001)	0.001(<0.001)
- Hainaut	1.246	0.174(<0.001)	0.001(<0.001)
- Liège	0.635	0.176(<0.001)	0.001(<0.001)
- Limburg	1.710	0.187(<0.001)	0.001(<0.001)
- Luxembourg	0.110	0.194(0.571)	0.001(<0.001)

higher seroprevalence than homebred animals. An increasing herd size, increasing animal density and decreasing herd density give an increase in the seroprevalence. The apparent contradictive effect of animal density and herd density on the seroprevalence has been observed before in veterinary epidemiology.

One can think of possible explanations as that low herd density points at regions where family and amateur farms are located, while a high density refers to regions of professional farms. The latter being more aware of the potential danger of infectious diseases like the BoHV-1. This however should be investigated further.

8.6.3 Model 2: Including Purchase as an Interaction

In this section we focus on

$$\begin{aligned}\text{logit}(P(\text{gB}=1)) \quad = \quad & \beta_0 + (\beta_1 \text{age}^{p1} + \beta_2 \text{age}^{p2}) * I_0 + (\beta_3 \text{age}^{p3} + \beta_4 \text{age}^{p4}) * I_1 \\ & + \beta_5 \text{herd size} + \beta_6 \text{purchase} + \beta_7 \text{purchase} * \text{herd size},\end{aligned} \tag{8.28}$$

where I_i denotes an indicator variable which takes the value 1 if purchase $= i$ and 0 otherwise, $i = 0, 1$ (homebred and purchased, respectively).

Whether the animals were purchased or homebred has a substantial influence on the powers chosen for both fractional polynomials in the model. While there is a rather small difference between the use of complete cases and weighted complete cases for homebred animals, there is a considerable difference between the two methods for purchased animals. The interaction between herd size and purchase is not significant (empirical S.E.) based on the complete cases, but it is, based on the weighted complete cases. From this model and the additive model in the previous section the FOI can be derived. A graphical representation of the FOI for Model 1 is not feasible due to the high dimensional covariate space. In Figure 8.12, the age-specific seroprevalence and FOI for Model 2 show that purchase is an important discriminator. From a veterinary point of view, purchased animals are expected to have a higher seroprevalence compared to homebred animals (Boelaert *et al.*, 2005). The interaction model shows that young purchased animals have a higher seroprevalence than young homebred animals, while the seroprevalence for older purchased animals is smaller compared to older homebred animals. Indeed, animals are purchased at a young age and are likely to either be infected or to have recovered from an infection. After introduction into the herd, they can spread the infection to the other animals in the herd, which are mostly homebred (Once recovered from infection animals can turn infectious again due to numerous reasons like, e.g., stress).

Table 8.10: BoHV-1 data: Interaction model of purchase and age.

	GEE (Independence)		
Parameter	Estimate	Emp.S.E.(p-value)	Mod.S.E.(p-value)
Complete Cases			
Intercept	3.038	0.939(0.001)	0.488(<0.001)
age_0^{-1}	7.711	3.134(0.014)	1.657(<0.001)
$\text{age}_0^{-0.5}$	-12.592	3.554(<0.001)	1.838(<0.001)
$\text{age}_1^{0.5}$	0.835	0.487(0.086)	0.201(<0.001)
age_1^{2}	-0.003	0.007(0.668)	0.004(0.453)
herd size	0.009	0.005(0.072)	7.2e-4(<0.001)
purchase	-5.593	1.231(<0.001)	0.593(<0.001)
herd size*purchase	-0.004	0.003(0.182)	0.001(<0.001)
Weighted Complete Cases			
Intercept	3.151	0.946(0.001)	0.397(<0.001)
age_0^{-1}	7.336	2.868(0.011)	1.331(<0.001)
$\text{age}_0^{-0.5}$	-12.341	3.385(<0.001)	1.486(<0.001)
age_1^{0}	1.758	3.357(0.600)	1.185(0.138)
$\text{age}_1^{0.5}$	-1.1e-034	1.394(1.000)	0.505(1.000)
herd size	0.008	0.004(0.046)	0.001(<0.001)
purchase	-4.509	1.546(0.004)	0.530(<0.001)
herd size*purchase	-0.010	0.004(0.012)	0.001(<0.001)

Purchased animals are thus more likely to be infected at a young age in contrast to homebred animals. Secondly, animals in beef herds are slaughtered at young age (18-20 months) and therefore a decline for older ages is caused by the absence of these animals compared to homebred animals.

For the weighted complete cases there is a significant effect of purchase on the influence of herd size. For purchased animals, an increasing herd size gives a decrease in seroprevalence, while for homebred animals there is an increasing effect.

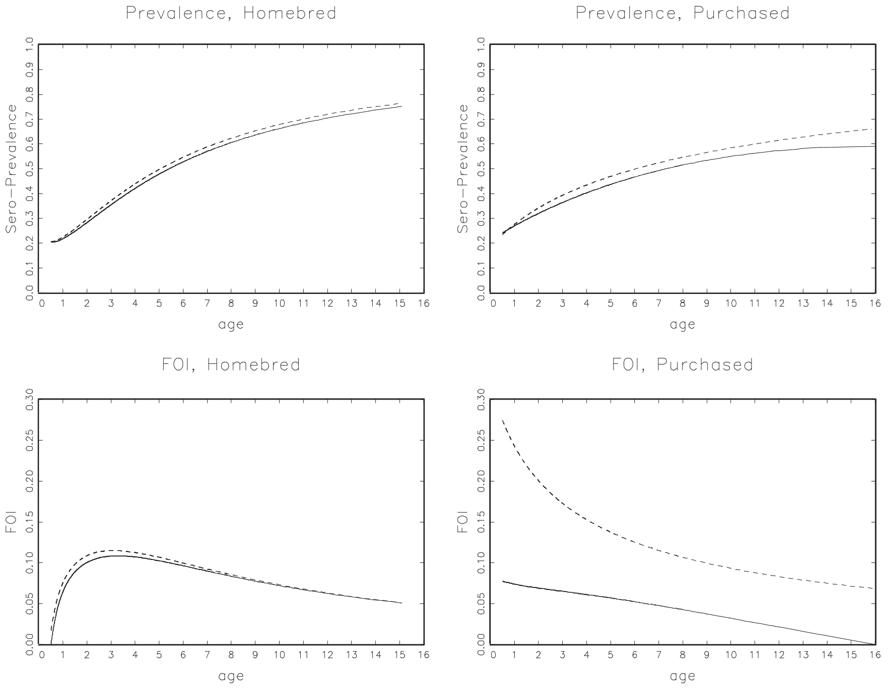


Figure 8.12: Plot of the seroprevalence and FOI for homebred (left) and purchased (right) as a function of age using complete cases (solid curve) and weighted complete cases (dashed curve).

From the FOI-curve, it can be seen that animals which are homebred have the typical tendency to have a maximal FOI around 3 years while for the purchased animals, a monotone decrease in the FOI can be observed. Especially for the weighted complete cases, this is a substantial decrease.

8.7 Discussion

In this chapter, the BoHV-1 data were analyzed to determine the FOI. It is clear from the results that the dataset has several complications. To overcome the complication of missing covariates an inverse probability weighted analysis is proposed (Robins *et al.*, 1994; Zhao *et al.*, 1996). Data are assumed to be missing at random throughout this chapter. Clustering, if regarded merely as a nuisance parameter, can be taken into account by using GEEs. Since the FOI can be seen as a hazard rate, i.e., the instant probability for an animal to get infected given that infection has not occurred yet, it has to be positive and thus the seroprevalence monotone increasing. To handle this latter complication a constrained analysis was performed. To correct the influence of an informative cluster size on the analysis, herd size was added to the model as a main effect. The combination of all these techniques to model such complex veterinary data can be termed as “Constrained, Flexible, Weighted Generalized Estimating Equations”(CFWGEE), where flexibility was achieved using a fractional polynomial for age.

A multiple imputation analysis can be seen as an alternative to the inverse probability weighted analysis. An alternative to GEE is the use of random-effects models where interest goes out to the clustering itself. Alternatives to the use of fractional polynomials are smoothing splines. The derivation of the model in that case will be even more computer intensive due to the constraint of the FOI to be positive. Selecting an appropriate flexible model when dealing with constraints together with other complications is an interesting topic of further research.

References

Aerts, M. and Claeskens, G. (1997) Local polynomial estimation in multiparameter likelihood models. *Journal of the American Statistical Association*, **92**, 1536–1545.

Aerts, M., Claeskens, G., Hens, N. and Molenberghs, G. (2002a) Local multiple imputation. *Biometrika*, **89**, 375–388.

Aerts, M., Claeskens, G. and Wand, M. P. (2002b) Some theory for penalized spline additive models. *Journal of Statistical Planning and Inference*, **103**, 455–470.

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M. (2002c) *Topics in Modeling of Clustered Data.* London: Chapmann and Hall.

Aerts, M., Janssen, P. and Veraverbeke, N. (1994) Bootstrapping regression quantiles. *Journal of Nonparametric Statistics*, **4**, 1–20.

Afifi, A. A. and Elashoff, R. M. (1969a) Missing observations in multivariate statistics: III: Large sample analysis of simple linear regression. *Journal of the American Statistical Association*, **64**, 337–358.

Afifi, A. A. and Elashoff, R. M. (1969b) Missing observations in multivariate statistics: IV: A note on simple linear regression. *Journal of the American Statistical Association*, **64**, 358–365.

Afifi, A. A. and Elsahoff, R. M. (1966) Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, **61**, 595–604.

Agostinelli, C. (2002) Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*, **56**, 289–300.

Ahmad, I. A. (1995) On multivariate kernel estimation for samples from weighted distributions. *Statistics and Probability Letters*, **22**, 121–129.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. Petrov and F. Csaki), 267–281. Budapest: Akademia Kiado.

Anderson, R. M. (1982) *Population dynamics of infectious diseases, theory and applications.* London: Chapman and Hall.

Anderson, R. M. and May, R. M. (1991) *Infectious diseases of humans: dynamic and control.* Oxford: Oxford University Press.

Baker, S. G., Ko, C.-W. and Graubard, B. I. (2003) A sensitivity analysis for non-randomly missing categorical data arising from a national health disability survey. *Biostatistics*, **4**, 41–56.

Becker, N. G. (1989) *Analysis of infectious diseases data.* London, Chapman and Hall.

Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour.* Princeton: Princeton University Press.

Boelaert, F., Biront, P., Soumare, B., Dispas, M., Vanopdenbosch, E., Vermeersch, J., Raskin, A., Dufey, J., Berkvens, D. and Kerkhofs, P. (2000) Prevalence of bovine herpesvirus-1 in the Belgian cattle population. *Preventive Veterinary Medicine*, **45**, 285–295.

Boelaert, F., Speybroeck, N., de Kruif, A., Aerts, M., Burzykowski, T., Molenberghs, G. and Berkvens, D. L. (2005) Risk factors for bovine herpesvirus-1 seropositivity. *Preventive Veterinary Medicine*, In Press.

Bottai, M. (2003) Confidence regions when the Fisher information is zero. *Biometrika*, **90**, 73–84.

Box, G. E. P. (1980) Samplin and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383–430.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Belmont, California: Wadsworth International Group.

Buck, S. F. (1960) A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B*, **22**.

Bühlmann, P. and Yu, B. (2002) Analyzing bagging. *The Annals of Statistics*, **30**, 927–961.

Burnham, K. P. and Anderson, D. R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach.* New York: Springer-Verlag.

Burzykowski, T., Molenberghs, G., Tafforeau, J., Van Oyen, H., Demarest, S. and Bellamammer, L. (1999) Missing data in the health interview survey 1997 in Belgium. *Archives of Public Health*, **57**, 107–129.

Carpenter, J. R. and Kenward, M. G. (2005) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A, Submitted.*

Cavanaugh, J. E. and Oleson, J. J. (2001) A diagnostic for assessing the influence of cases on the prediction of missing data. *The Statistician*, **50**, 427–440.

Cavanaugh, J. E. and Shumway, R. H. (1998) An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference.*, **67**, 45–65.

Cheng, P. E. (1994) Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81–87.

Christóbal Christóbal, J. A., Faraldo Roca, P. and González Manteiga, W. (1987) A class of linear regression parameter estimators constructed by nonparametric estimation. *Annals of Statistics*, **15**, 603–609.

Chu, C. K. and Cheng, P. E. (1995) Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, **48**, 85–99.

Contreras, M. and Ryan, L. M. (2000) Fitting nonlinear and constrained generalized estimating equations with optimization software. *Biometrics*, **56**, 1268–1271.

Cook, R. D. (1986) Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Cook, R. D. and Weisberg, S. (1982) *Residuals and influence in regression.* New York: Chapman and Hall.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics.* London: Chapman and Hall.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Crouchley, R. and Ganjali, M. (2002) The common structure of several models for non-ignorable dropout. *Statistical Modeling*, **2**, 39–62.

Daniels, M. J. and Hogan, J. W. (2000) Reparametrizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, **56**, 1241–1248.

Davidian, M. and Giltinan, D. M. (1995) *Nonlinear Models for Repeated Measurement Data.* London: Chapman and Hall.

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

Diamond, L. D. and McDonald, J. M. (1992) *Demographic Application of Event History Analysis.*, chap. Analysis of current-status data. Oxford University Press.

Diggle, P. and Kenward, M. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1–25.

Efron, B. (1994) Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, **89**, 463–75.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with b-splines and penalties (with discussion). *Statistical Science*, **89**, 89–121.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Faes, C., Hens, N., Aerts, M., Shkedy, Z., Geys, H., Mintiens, K., Laevens, H. and Boelaert, F. (2005) Population-averaged versus herd-specific force of infection. *Applied Statistics, Submitted.*

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

Faraldo, R. P. and Gonzalez Manteiga, W. (1987) *New perspectives in theoretical and applied statistics.*, chap. Efficiency of a new class of linear regression estimates obtained by preliminary nonparametric estimation., 229–242. New York: John Wiley.

Farrington, C. P. (1990) Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, **9**, 953–967.

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2001) Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Applied Statistics*, **50**, 251–292.

Feelders, A. (2000) Handling missing data in trees: surrogate splits or statistical imputation? *Tech. rep.*, Tilburg University.

Feldesman, M. R. (2002) Classification trees as an alternative to linear discriminant analysis. *American Journal of Physical Anthropology*, **119**, 257–275.

Fitzmaurice, G. and Laird, N. (2000) Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, **1**, 141–156.

Flanders, W. and Greenland, S. (1991) Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, **10**, 739–747.

Freund, Y. and Shapire, R. (1997) A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Gemechis, D. and Aerts, M. (2004) A comparative study of models for correlated binary data with applications to health services research. *Ethiopian Journal of Science*, **27**, 000–000.

Goss, P., Winer, E., Tannock, I. and Schwartz, L. (1999) Randomized phase iii trial comparing the new potent and selective third-generation aromatase inhibitor vorozole with megestrol acetate in postmenopausal advanced breast cancer patients. *Journal of Clinical Oncology*, **17**, 52–63.

Gray, H. L. and Schucany, W. R. (1972) *The Generalized Jackknife Statistic*, vol. 1 of *Statistics Textbooks and Monographs.* New York: Marcel Dekker.

Green, P. and Silverman, B. (1994) *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

Greenlees, J., Reece, W. and Zieschang, K. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251–261.

Grenfell, B. T. and Anderson, R. M. (1985) The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, **95**, 419–36.

Grummer-Strawn, L. M. (1993) Regression analysis of current status data: an application to breast feeding. *Biometrika*, **72**, 527–537.

Hall, P. and Presnell, B. (1999) Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society B*, **61**, 143–58.

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

Hart, J. D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer.

Hastie, T. and Tibshirani, R. (1987) Generalized additive models: some applications. *Journal of the American Statistical Association*, **82**, 371–386.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning.* New York: Springer.

Heitjan, D. and Little, R. (1991) Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13–29.

Hens, N., Aerts, M. and Molenberghs, G. (2005a) Model selection for incomplete and design-based samples. *Statistics in Medicine, Submitted.*

Hens, N., Aerts, M., Molenberghs, G., Thijs, H. and Verbeke, G. (2005b) Kernel weighted influence measures. *Computational Statistics and Data Analysis*, **48**, 467–487.

Hens, N., Bruckers, L., Arbyn, M., Aerts, M. and Molenberghs, G. (2002) Classification tree analysis of cervix cancer screening in the belgian health interview survey. *Archives of Public Health*, **60**, 275–294.

Hens, N., Faes, C., Aerts, M., Shkedy, Z., Mintiens, K., Laevens, H. and Boelaert, F. (2005c) The influence of missing data on the force of infection for the BoHV-1 data. *Applied Statistics, Submitted.*

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004) Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine*, **23**, 1455–1497.

Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Hossain, Z. (2002) Modified akaike information criterion (maic) for statistical model selection. *Pakistan Journal of Statistics*, **18**, 383–393.

Hurvich, C., Simonoff, J. and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271–293.

Hurvich, C. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Ibrahim, J. G. (1990) Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**, 765–769.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G. and Kenward, M. (2005) The nature of sensitivity in monotone missing not at random models. *Computational Statistics and Data Analysis, To Appear.*

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H. and Van Steen, K. (2003) A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 410–419.

Janssen, P. and Mikosch, T. (1997) An elementary proof of bootstrap consistency for the sample mean. *Tech. rep.*, Diepenbeek and Groningen.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2001) Efficiency of linear regression estimators based on presmoothing. *Communications in Statistics - Theory and Methods*, **30**, 2079–2097.

Jewell, N. P. and Van Der Laan, M. (1995) Generalizations of current status data with applications. *Lifetime data analysis*, **1**, 101–109.

Jones, M. (1991) Kernel density estimation for length biased data. *Biometrika*, **78**, 511–519.

Keiding, N., Begtrup, K., Scheike, T. H. and Hasibeder, G. (1996) Estimation from current status data in continuous time. *Lifetime Data Analysis*, **2**, 119–129.

Kenward, M. G. (1998) Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Kish, L. (1995) *Survey Sampling.* New York: Wiley.

Lawrance, A. J. (1995) Deletion influence and masking in regression. *Journal of the Royal Statistical Society, Series B*, **57**, 181–189.

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Liang, K. and Zeger, S. (1986) Longitudinal dat analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, D. and Ying, Z. (2003) Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics*, **4**, 385–398.

Lipsitz, S., Zhao, L. and Molenberghs, G. (1998) A semiparametric method of multiple imputation. *Journal of the Royal Statistical Society, Series B*, **60**, 127–144.

Little, R. (1992) Regression with missing x's: A review. *Journal of the American Statistical Association*, **87**, 1227–1237.

Little, R. and An, H. (2004) Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, **14**, 949–968.

Little, R. and Rubin, D. (1987) *Statistical Analysis with Missing Data.* New York.: Wiley.

Marx, B. D. and Eilers, P. H. C. (1998) Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models.* London: Chapman and Hall.

McDonald, B. W. (1993) Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society, Series B*, **55**, 391–397.

Michiels, B., Molenberghs, G. and Lipsitz, S. (1999) A pattern-mixture odds ratio model for incomplete categorical data. *Communications in Statistics - Theory and Methods*, **28**, 2843–2870.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M., Mallinckrodt, C. and Carroll, R. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445–464.

Molenberghs, G., Thijs, H., Kenward, M. G. and Verbeke, G. (2003) Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistica Neerlandica*, **57**, 112–135.

Molenberghs, G., Verbeke, G., Thijs, T., Lesaffre, E. and Kenward, M. (2001) Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Moons, E., Aerts, M. and Wets, G. (2004) Tree based lack-of-fit test. *Statistics in Medicine*, **23**, 1425–1438.

Moulton, L. H. and Zeger, S. L. (1989) Analysing repeated measures on generalized linear models via the bootstrap. *Biometrics*, **45**, 381–394.

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

Nielsen, S. F. (2001) Nonparametric conditional mean imputation. *Journal of Statistical Planning and Inference*, **99**, 129–150.

Pan, W. (2001a) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–125.

Pan, W. (2001b) Model selection in estimating equations. *Biometrics*, **57**, 529–534.

Quataert, P., Van Oyen, H. and Tafforeau, J. (1998) Health interview survey 1997. protocol for selection of the households and the respondents. *S.P.H.*, **12**.

Quenouille, M. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Quinlan, J. (1986) Induction of decision trees. *Machine Learning*, **1**, 81–106.

R Development Core Team (2004) *R: A language and environment for statistical computing.* Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.: R Foundation for Statistical Computing.

Rao, J. and Shao, J. (1992) Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79**, 811–822.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Robins, J., Rotnitzky, A. and Zhao, L. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

Robins, J., Rotnitzky, A. and Zhao, L. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rosenbaum, P. and Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **60**, 211–3.

Rotnitzky, A., Cox, D., Bottai, M. and Robins, J. (2000) Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243–284.

Rotnitzky, A. and Robins, J. (1995) Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, **82**, 805–820.

Rotnitzky, A., Robins, J. and Scharfstein, D. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**, 1321–1339.

Roy, J. and Lin, X. (2002) The analysis of multivariate longitudinal outcomes with nongignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association*, **97**, 40–52.

Royston, P. and Altman, D. (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.

Rubin, D. (1978) *Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse.*, 1–23. U.S. Department of Commerce.

Rubin, D. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, R. (1976) Inference about means from incomplete multivariate data. *Biometrika*, **63**, 593–604.

Ruppert, D. and Carroll, R. J. (2000) Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **45**, 204–225.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression.* Cambridge: Cambridge University Press.

Ryan, B. and Joiner, B. (1994) *Minitab Handbook.* Belmont, California: Wadsworth Publishing, 3rd edition edn.

Ryan, T. (1997) *Modern Regression Methods.* New York: Wiley.

Schafer, J. (1997) *Analysis of Incomple Multivariate Data.* London: Chapman and Hall.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse model (with discussion). *Journal of the American Statistical Association*, **94**, 1096–1146.

Schenker, N. and Welsh, A. (1988) Asymptotic results for multiple imputation. *Annals of Statistics*, **16**, 1550–1566.

Schipper, H., Clinch, J. and McMurray, A. (1984) Measuring the quality of life of breast cancer patients: the functional-living-index-cancer: Development and validation. *Journal of Clincial Oncology*, **2**, 472–483.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Segal, M. R. (1992) Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, **87**, 407–418.

Segal, M. R. (1995) Extending the elements of tree-structured regression. *Statistical Methods in Medical Research*, **4**, 219–236.

Shanno, D. F. (1985) On Broyden-Fletcher-Goldfarb-Shanno method. *Journal of Optimization Theory and Applications*, **46**, 87–94.

Shannon, W. D. and Banks, D. (1999) Combining classification trees using mle. *Statistics in Medicine*, **18**, 727–740.

Sherman, M. and Le Cessie, S. (1997) A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics, Part B - Simulation and Computation*, **26**, 901–925.

Shimodaira, H. (1994) A new criterion for selecting models from partially observed data. In *Selecting Models from Data: Artificial Intelligence and Statistics IV.* (eds. P. Cheeseman and R. W. Oldford), vol. 89, 21–29.

Shkedy, Z. (2003) *Flexible Statistical Modelling: Application to Infectious Diseases and Astronomical Data.* Ph.D. thesis, Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2003) Modelling forces of infection by using monotone local polynomials. *Applied Statistics*, **52**, 469–485.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2005) Modeling age dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine: In Press.*

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, **47**, 1–21.

Simonoff, J. (1996) *Smoothing Methods in Statistics.* New York: Springer.

Speybroeck, N., Berkvens, D., Mfoukou-Ntsakala, A., Aerts, M., Hens, N., Van Huylenbroeck, G. and Thys, E. (2004) Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa. *Agricultural Systems*, **80**, 133–149.

Speybroeck, N., Boelaert, F., Renard, D., Burzykowski, T., Mintiens, K., Molenberghs, G. and Berkvens, D. L. (2003) Design-based analysis of surveys: a bovine herpesvirus 1 case study. *Epidemiology and Infection*, **131**, 991–1002.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.

Takeuchi, K. (1976) Discussion of informational statistics and a criterion for model fitting. *Suri-Kagaku*, **153**, 12–18.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Thijs, H., Molenberghs, G. and Verbeke, G. (2000) The milk protein trial: Influence analysis of the dropout process. *Biometrical Journal*, **42**, 1–30.

Titterington, D. and Sedransk, J. (1989) Imputation of missing values using density estimation. *Statistics and Probability Letters*, **8**, 411–418.

Tukey, J. (1958) Bias and confidence in not quite large samples. *Annals of Mathematical Statistics.*, **29**, 614.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* New York: Springer Verlag.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M. (2001) Sensitivity analysis for non-random dropout: A local influence approach. *Biometrics*, **57**, 7–14.

Wahba, G. (1980) *Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data.*, chap. Approximation Theory III, 905–912. New York: Academic Press.

Wahba, G. (1990) *Spline Models for Observational Data.* CBMS-NSF series. SIAM, Philadelphia.

Wand, M. and Jones, M. (1995) *Kernel Smoothing.* London: Chapman and Hall.

Wang, C., Wang, S., Gutierrez, R. and Carroll, R. (1998) Local linear regression for generalized linear models with missing data. *Annals of Statistics*, **26**, 1028–50.

Wang, Q., Linton, O. and Härdle, W. (2004) Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, **99**, 334–345.

Watson, G. (1964) Smooth regression analysis. *Sankyā A*, **26**, 359–72.

Williamson, J. M., Datta, S. and Satten, G. A. (2003) Marginal analyses of clusterd data when cluster size is informative . *Biometrics*, **59**, 36–42.

Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, **62**, 413–428.

Wood, S. N. (2001) mgcv: Gams and generalized ridge regression for R. *R News*, **1**, 20–25.

Wood, S. N. (2005) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association.* In Press.

Wood, S. N. and Augustin, N. H. (2002) Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157**, 157–177.

Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988) Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zhang, H. and Singer, B. (1999) *Recursive Partitioning in the Health Sciences.* New York: Springer-Verlag.

Zhao, L. P. and Lipsitz, S. (1992) Design and analysis of two-stage studies. *Statistics in Medicine*, **11**, 769–782.

Zhao, L. P., Lipsitz, S. and Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.

Zhu, H., Lee, S.-Y., Wei, B.-C. and Zhou, J. (2001) Case-deletion measures for models with incomplete data. *Biometrika*, **88**, 727–737.

Samenvatting

Het Gebruik van Niet- en Semi-parametrische Technieken bij het Modelleren van Ontbrekende Gegevens

Het doel van een statistische analyse is om, aan de hand van een steekproef, geldige en efficiënte gevolgtrekkingen te maken omtrent de beschouwde populatie. In verschillende studies zoals bijvoorbeeld klinische en epidemiologische studies stoot men echter vaak op het probleem van ontbrekende gegevens die dit proces bemoeilijken. In het verleden werd er gebruik gemaakt van parametrische modellen om onvolledige datasets te modelleren (zie bv Afifi and Elsahoff, 1966; Ibrahim, 1990). Recent is er een algemene trend naar het gebruik van niet- en semi-parametrische technieken die de typische veronderstellingen, waarop de parametrische methoden steunen, versoepelen. Parametrische technieken steunen op verschillende veronderstellingen zoals bijvoorbeeld in een regressie-context op de verdeling van de response variabele en op de functionele relatie die het verband tussen de response en de verklarende variabelen weergeeft. Niet-parametrische technieken zijn vrij van veronderstellingen, terwijl semi-parametrische technieken enkel gedeeltelijk veronderstellingen gebruiken.

Omtrent het gebruik van deze niet- en semi-parametrische technieken bestaan er twee verschillende standpunten. Een eerste standpunt werd beschreven door Silverman (1985): “An initial non-parametric estimate may well suggest a suitable parametric model (such as linear regression), but nevertheless will give the data more of a chance to speak for themselves in choosing the model to be fitted.”Daarmee duidt Silverman (1985) op de motivatie die niet-parametrische schattingen met zich kunnen meebrengen om een gepast parametrisch model te kiezen. Deze niet-parametrische schattingen geven de data de gelegenheid om voor zichzelf te “spreken”. Een tweede standpunt komt voort uit een standpunt geformuleerd door Box (1980): “Known facts (data) suggest a tentative model, implicit or explicit, which in turn suggests a particular examination and analysis of data and/or the need to acquire further data;

analysis may then suggest a modified model that may require further practical illumination and so on."Niet-parametrische technieken zijn het aangewezen hulpmiddel om bestaande parametrisch modellen te optimaliseren (zie Hastie and Tibshirani, 1987; Simonoff, 1996; Hart, 1997).

Niet- en semi-parametrische methoden zijn in het algemeen niet zo efficiënt als parametrische methoden indien het veronderstelde model geschikt is. Indien het model echter niet het geschikte model is, kunnen de gevolgtrekkingen hieruit misleidend zijn.

Onder deze niet- en semi-parametrische technieken vind men kernschatters (zie bv Watson, 1964; Nadaraya, 1964), splines (Eubank, 1988), veralgemeende additieve modellen (Hastie and Tibshirani, 1987; Wood, 2001; Wood and Augustin, 2002; Wood, 2005) en classificatie- en regressiebomen (Breiman *et al.*, 1984), die in deze thesis aan bod komen.

Bij het modelleren van onvolledige data, maakt men meermaals gebruik van de terminologie die geïntroduceerd werd door Little and Rubin (1987) en Rubin (1987). Vooreerst, zegt men dat data volledig willekeurig ontbreken (*missing completely at random*, MCAR) indien de kans om te ontbreken onafhankelijk is van zowel de geobserveerde als ontbrekende gegevens. Indien deze kans mogelijk afhangt van de geobserveerde gegevens, maar niet van de ontbrekende gegevens, dan zegt men dat de data willekeurig ontbreken (*missing at random*, MAR). Tenslotte noemt men het ontbreken van gegevens niet-willekeurig indien deze kans afhangt van ontbrekende en mogelijk ook van de geobserveerde gegevens (*missing not at random*, MNAR). In de praktijk is het meestal niet aannemelijk dat ontbrekende gegevens aan de MCAR-veronderstelling voldoen. De MAR-veronderstelling wordt doorgaans veel gebruikt en is in vele situaties te verdedigen. Indien men echter data met niet-willekeurig ontbrekende gegevens wilt analyseren, moet men doorgaans verdere ontestbare veronderstellingen maken. Een sensitiviteitsanalyse is in deze laatste situatie onontbeerlijk.

In de literatuur zijn verschillende technieken voorgesteld om met ontbrekende gegevens om te gaan. Ze kunnen ruwweg onderverdeeld worden in vier groepen: (1) *'complete case analysis'*; (2) *'multiple imputation'*; (3) *'inverse probability weighting'* en (4) *'fully model-based procedures'*. Bij de 'complete case analysis' gebruikt de analyse enkel de eenheden die volledig geobserveerd zijn. Deze methode is gemakkelijk toe te passen maar kan tot inefficiëntie en vertekening leiden (Little and Rubin, 1987). Bij 'multiple imputation' worden er meerdere malen gegevens geïmputeerd. Vervolgens worden de vervolledigde datasets geanalyseerd en de respectievelijke resultaten gecombineerd. Ook deze methode heeft beperkingen (zie bv Rubin, 1978;

Rubin and Schenker, 1986; Little and Rubin, 1987; Tanner and Wong, 1987; Schafer, 1997). 'Inverse probability weighting' geeft de volledig geobserveerde eenheden een gewicht, gelijk aan de inverse van de kans dat de eenheid volledig geobserveerd is. Zo vertegenwoordigen ze op een impliciete manier de ontbrekende gegevens (Flanders and Greenland, 1991; Zhao and Lipsitz, 1992; Robins *et al.*, 1994; Zhao *et al.*, 1996). In een recente publicatie van Carpenter and Kenward (2005), werd deze laatste methode vergeleken met 'multiple imputation'. 'Fully model-based procedures' modelleren naast het meetmodel ook het mechanisme achter het ontbreken van gegevens. Deze procedures steunen op ontestbare veronderstellingen waardoor een sensitiviteits-analyse aangewezen is. Van deze laatste groep zijn er vele methoden ontwikkeld voor herhaalde metingen zoals bijvoorbeeld voor klinische studies waar het uitvallen van patiënten meestal de oorzaak is van het ontbreken van gegevens. Dit laatste fenomeen wordt *'dropout'* (uitvallen) genoemd. Voorbeelden van zulke modellen zijn 'selection'-modellen, 'pattern-mixture'-modellen en 'shared-parameter'-modellen. Voor een vollediger overzicht verwijzen we naar Hogan *et al.* (2004) en Molenberghs *et al.* (2004).

In deze thesis worden verschillende niet- en semi-parametrische technieken gebruikt voor het modelleren van onvolledige gegevens. Het gepresenteerde materiaal geeft duidelijk het voordeel weer van het versoepelen van de veronderstellingen. Vele auteurs zoals Lipsitz *et al.* (1998), Rubin and Schenker (1986) en Heitjan and Little (1991) hebben reeds gebruik gemaakt van niet- en semi-parametrische technieken in het modelleren van ontbrekende gegevens.

Het inleidende hoofdstuk van deze thesis geeft een kort overzicht van niet- en semi-parametrische technieken en van verschillende technieken om onvolledige data te analyseren. In een tweede hoofdstuk stellen we voor om een lokale meervoudige imputatie methode te gebruiken in een regressie-context met ontbrekende response gegevens (Aerts *et al.*, 2002a). De interesse gaat uit naar een marginale parameter van de response-verdeling. Hiervoor gebruikt men een driedelige imputatie-methode: (1) in eerste fase worden de gegevens geresampled d.m.v. een lokale bootstrap, (2) gebaseerd op deze bootstrap sample worden gegevens geïmputeerd: hetzij gebaseerd op een nieuwe lokale bootstrap; hetzij gebaseerd op een normale likelihood. Deze twee stappen worden telkens m keer herhaald, wat resulteert in een meervoudige imputatie. De laatste stap, (3), is het construeren van de schatter voor de marginale parameter van de response verdeling. De lokale bootstrap die hier toegepast wordt, maakt gebruik van omgevingen van de response variabele om uit te resamplen. De motivatie voor deze methode werd gegeven door Cheng (1994) die gebruik maakt van een enkelvoudige imputatie en aldus de variabiliteit onderschatte. Door de bijkomende eerste stap, wordt de variabiliteit juist ingeschat. Dit wordt zowel theoretisch als met simulaties aangetoond.

Het derde en vierde hoofdstuk geven aan dat er in een regressie-analyse met ontbrekende gegevens nood is aan een aangepast model selectie criterium. Het toepassen van het Akaike Informatie Criterium (AIC, Akaike, 1973) enkel op de volledige geobserveerde eenheden kan leiden tot foutieve modelkeuzes. Een eerste oplossing hiervoor is het gebruik van een gewogen AIC-criterium waar volledig geobserveerde eenheden een gewicht krijgen zoals bij 'inverse probability weighting'. De methode is toepasbaar zowel op onvolledige data als op design-gebaseerde steekproeven (Hens *et al.*, 2005a). Indien de gewichten ongekend zijn, kan men semi-parametrische methoden zoals veralgemeende additieve modellen gebruiken om deze te schatten. Theoretische argumenten geven samen met simulaties aan dat de methode vaker leidt tot een correcte modelkeuze. Indien er slechts enkele volledige eenheden zijn, is het gebruik van gewichten niet geschikt. In dit geval kan men opteren om eerste te imputeren en gebaseerd op de geïmputeerde dataset een model te kiezen. De imputatie gebeurt hier semi-parametrisch door gebruik te maken veralgemeende additieve modellen. Een simulatie studie toont aan dat het selecteren na imputatie heel wat potentieel heeft. Het nadeel van deze laatste methode is echter dat deze grotendeels op het onderliggende imputatiemodel steunt. Indien dit imputatiemodel niet goed gedefinieerd is, zal dit tot gevolg hebben dat de methode onderuit gaat.

In een vijfde hoofdstuk wordt een gevallenstudie bekeken omtrent baarmoederhalskanker waar, voor vele vrouwen, gegevens ontbreken. Naast het probleem van onvolledige data, is er een tweede moeilijkheid, namelijk het specifieke design waarmee deze studie is opgebouwd als deel van de nationale gezondheidsenquête (HIS) van België uitgevoerd in 1997. Met behulp van de technieken geïntroduceerd in Hoofdstukken 4 en 5 wordt de invloed van het negeren van deze complicaties aangetoond. We vergelijken het parametrisch logistisch regressiemodel met classificatiebomen die volledig niet-parametrisch zijn (Hens *et al.*, 2002). Verschillende manieren om met het design en de ontbrekende gegevens in classificatiebomen om te gaan worden hier geïllustreerd en besproken.

Een tweede deel van deze thesis betreft gecorreleerde gegevens. Hoofdstukken 6 en 7 geven aan dat een sensitiviteitsanalyse onmisbaar is bij het analyseren van herhaalde metingen met dropout. Het model dat hier beschouwd wordt, is het selectiemodel van Diggle and Kenward (1994). In Hoofdstuk 6 wordt een niet-parametrische variant van *'global'* en *'local influence'* ontwikkeld. Deze tool is in staat om de invloed van verschillende types van observaties op een selectie-model te detecteren door gebruik te maken van kerngewichten die een omgeving van een