Nittner:

Missing at Random (MAR) in Nonparametric Regression - A Simulation Experiment

Sonderforschungsbereich 386, Paper 284 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Missing at Random (MAR) in Nonparametric Regression

—

A Simulation Experiment

T. Nittner

July 26, 2002

Abstract

This paper considers an additive model $y = f(x) + \epsilon$ when some observations on x are missing at random but corresponding observations on y are available. Especially for this model missing at random is an interesting case because of the fact that the complete case analysis is not expected to be suitable. A simulation study is reported and methods are compared based on superiority measures as the sample mean squared error, sample variance and estimated sample bias. In detail, complete case analysis, zero order regression plus random noise, single imputation and nearest neighbor imputation are discussed.

KEY WORDS: missing at random, additive models, nonparametric, nearest neighbor imputation, simulation.

1 Introduction

This report is the proceeding of Nittner (2002) which adresses the problem and consequences of missing data completely at random (MCAR) for an additive model. A simulation experiment is conducted and four imputation procedures are compared with the standard complete case analysis under the criterion of sample mean squared error. Based on the results, some modifications concerning the settings of the simulation experiment were done besides the fact that the missingness here is supposed to depend on the response y. It is an interesting model because it is expected to show the asymptotic deficits of the complete case analysis.

The plan of the paper is as follows. Sections 1.1, 1.2 and 1.3 introduce the data model, the model for the missing mechanism with well known definitions and terms and the approach through which the model is estimated. The imputation methods for the missing values are described within Section 2. In Section 3 the details on the simulation experiment are given along with the discussion on the results obtained.

1.1 The Model for the Data

Let an additive model connecting y_i and x_i be

$$y_i = f(x_i) + \epsilon_i\,(x_i, y_i), i = 1, \ldots, n. \tag{1.1}$$

We assume $\mathrm{E}(\epsilon \mid X) = \mathrm{E}(\epsilon)$ and $\mathrm{V}(\epsilon \mid X) = \mathrm{V}(\epsilon) = \sigma^2 I_n$. In order to prevent the free constant, we assume $\mathrm{E}(f(x)) = 0$.

As already mentioned in the beginning, the independent covariate X is assumed to be affected by missing values according to MAR. What exactly is meant by missing at random will be described in the next section after a short introduction to the missing data pattern.

1.2 The Model for the Missing Mechanism

Here, the response vector y is assumed to be completely observed whereas the covariate X is only affected by missing values. Based on this assumption the data matrix can be partitioned as

$$(y, X) = \left(\begin{pmatrix} y_{\mathrm{obs}} \\ y_{\mathrm{mis}} \end{pmatrix}, \begin{pmatrix} X_{\mathrm{obs}} \\ X_{\mathrm{mis}} \end{pmatrix} \right). \tag{1.2}$$

The index 'obs' indicates rows without missing data for X and y, the index 'mis' contains rows where y is observed and X is completely missing.

1.2.1 Missing Data Pattern

The well known missing data pattern is the first step to characterize the situation of an incomplete data set by visualizing observed and missing parts of the data and the variables respectively. It simply represents the data set variable–by–variable. Each bar represents a variable whose length indicates if there are missing cases for this variable or not. Situation (1.2) leads to Figure 1.1 called univariate missing data pattern (Little and Rubin (1987)). Univariate missing data are a special case of the monotone pattern as shown in Figure 1.2 where the variables can be ordered in such a way that a variable is observed for at least the cases of the previous one. These pattern may first give an impression

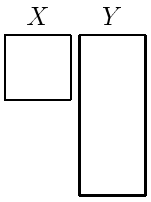

Figure 1.1: Univariate missing data pattern.

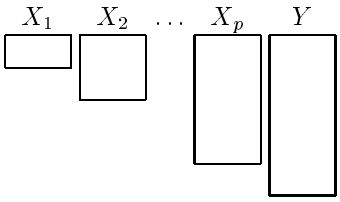

Figure 1.2: Monotone missing data pattern.

to what extent the data are missing. If X is assumed to be missing for large

values of y, the values can be ordered and a missing data pattern may describe this behavior too. But obviously, this technique may be swamped with a higher level of dependencies. A way to overcome this defect consists of defining the so called missing data mechanisms.

1.2.2 Missing Data Mechanism

The main question within this context is whether the missing data mechanism can be ignored or not. One possibility is to make an assumption that the mechanism is ignorable in the sense that the other possibility consists of including the missing data mechanism (which still is to be defined) in the statistical model by including the distribution of an indicator variable indicating whether a component is observed or missing. Little and Rubin (1987), for example, define the data matrix $Z = (Z_{\text{obs}}, Z_{\text{mis}})$ representing the data that would occur without missing values. Further, define a random variable R indicating the missingness within the data matrix Z according to

$$r_{ij} = \begin{cases} 1 & \text{if} \quad z_{ij} \quad \text{observed} \\ 0 & \text{if} \quad z_{ij} \quad \text{missing} \end{cases} \quad \forall\, i = 1, \ldots, n,\ j = 1, \ldots, p+1. \tag{1.3}$$

The question whether the missing mechanism can be ignored for the estimation of θ is the same as the question whether statistical inferences are based on $f(Z_{\text{obs}}, R \mid \theta, \Phi)$ with Φ being an unknown parameter of the missing mechanism and θ being the parameter of the density of $(Z_{\text{obs}}, Z_{\text{mis}})$—or just on the density $f(Z_{\text{obs}}, \theta)$ which ignores the missing mechanism. Considering the density $f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)$ allows to classify the missingness into

1. MCAR (missing completely at random) if

$$f(R \mid Z, \Phi) = f(R \mid \Phi) \quad \forall Z\,, \tag{1.4}$$

2. MAR (missing at random) if

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, \Phi) \quad \forall Z_{\text{mis}}\,, \text{ and} \tag{1.5}$$

3. MNAR (missing not at random)

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)\,. \tag{1.6}$$

Following Little and Rubin (1987), the missing data mechanism is said to be ignorable in the context of likelihood inference when the distribution of the missing mechanism is independent of the missing values [(1.5)] themselves. This becomes more clear by computing the density of the actual observed data obtained by integrating Z_{mis} out of the density

$$f(Z_{\text{obs}}, R \mid \theta, \Phi) = \int f(Z_{\text{obs}}, Z_{\text{mis}} \mid \theta) f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi) dZ_{\text{mis}} \tag{1.7}$$

which by using (1.5) leads to

$$\begin{aligned} f(Z_{\text{obs}}, R, \theta, \Phi) &= f(R \mid Z_{\text{obs}}, \Phi) \int f(Z_{\text{obs}}, Z_{\text{mis}}, \theta) dZ_{\text{mis}} \\ &= f(R \mid Z_{\text{obs}}, \Phi) f(Z_{\text{obs}} \mid \theta)\,. \end{aligned} \tag{1.8}$$

If the parameters θ and Φ concerning the density of the data Z and the missing mechanism, respectively, are distinct then the likelihood–based inferences for θ based on $f(Z_{\text{obs}}, R \mid \theta, \Phi)$ and for θ based on $f(Z_{\text{obs}} \mid \theta)$ are the same. Following Schafer (1997), θ and Φ are distinct if each parameter contains no information about the other.

The next section describes the methods of inference which are applied to the different complete data sets. Readers being familiar with the method of iteratively reweighted least squares may just take a look at the way of estimating the smoothing parameters; this section is equivalent to the corresponding section in Nittner (2002).

1.3 The Inference

The well known trade–off between wiggliness of the estimated curve and closeness to the data motivates the minimization of the target function

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int f''(t)^2 dt\,. \tag{1.9}$$

The value of λ controls the trade–off. We assume that f' and f'' are continuous, f'' is twice integrable. For $\lambda \to \infty$ the estimated curve is a straight line whereas for $\lambda \to 0$ the estimated curve is an interpolating spline.

Following Wood (2000) the problem of estimating the parameters $\beta^{(k+1)}$ of the nonlinear function f with $\mathrm{E}(f(Y_i)) = f(\beta)$ by Fisher–Scoring is equivalent to solving the weighted penalized least squares problem iteratively. We obtain

$$\min \lambda \mid\mid W^{\frac{1}{2}}(z^{(k)} - X\beta) \mid\mid^2 + \sum_i \theta_i \beta' S_i \beta \tag{1.10}$$

with the iteratively least squares (IRLS) where the least squares problem at each iteration is replaced by a penalized one. The S_i is a non–negative definite matrix of coefficients defining the ith penalty associated with the smoothing parameter θ_i, W is a diagonal matrix of weights and λ is the overall smoothing parameter. With practical point of view, in the case of an estimable parameter θ_i it is more interesting to consider generalized cross validation (GCV), one method to select smoothing parameters *that has proven effective and has good theoretical properties*, Wood (2000), p. 413. The problem here is to minimize the GCV–Scores

$$V = \frac{\mid\mid W^{\frac{1}{2}}(y - A(\lambda, \theta_i)y) \mid\mid^2 / n}{[1 - \mathrm{tr}(A(\lambda, \theta_i))/n]^2} \tag{1.11}$$

with respect to θ_i/λ. Combining the two procedures solves problem (1.9) and could be written in two steps.

1. Estimate μ and the variances V_i for each y_i with the help of $\beta^{(k)}$; compute

   (i) the diagonal weight matrix W with $W_{ii} = (g'(\mu_i)^2 V_i)^{-1}$

   (ii) the vector $z = X\beta + \Gamma(y - \mu)$, of pseudo–data with the diagonal matrix $\Gamma_{ii} = (g'(\mu_i))^{-1}$

2. Compute θ_i by minimizing $\frac{||W^{\frac{1}{2}}(z - X\beta)||^2}{(\mathrm{tr}(I-A))^2}$ where β is the solution obtained by minimizing $|| W^{\frac{1}{2}}(z - X\beta) ||^2 + \sum \theta_j \beta' S_j \beta$ with respect to β and A denotes the hat matrix $A = X(X'WX + \sum \theta_j \beta' S_j \beta)^{-1} X'W$.

Table 1.1: Iteratively Reweighted Least Squares with GCV.

For a more detailed description of additive models and estimation concepts,see Hastie and Tibshirani (1990), Fahrmeir and Tutz (2001) and Wood (2000). The settings of the simulation experiment concerning parameters for estimating the model are described in Section 3.1.

2 The Imputation Methods

First we describe the two nonparametric methods. According to the results of the first simulation experiment assuming MCAR in Nittner (2002) the nearest neighbor imputations showed better results than the alternative imputation procedures, namely the zero order regression plus random noise and the single imputation .

2.1 Nearest Neighbor Imputation

After an introduction to the usual nearest neighbor imputation, hereafter called as classical version or NN1 and a modified version (NN2), investigated in the earlier work are presented.

2.1.1 Nearest Neighbor Imputation—Version I (NN1)

Despite its long history within theoretical and practical work, the nearest neighbor imputation according to Chen and Shao (2001) is still not fully investigated.

Referring to the data structure (1.2) with m missing values for the row indices $i = n-m+1, \ldots, n$ visualized as

$$\underbrace{x_1, \ldots, x_{n-m}}_{\text{observed}}, \underbrace{x_{n-m+1}, \ldots, x_n}_{\text{missing}} \quad \text{and} \tag{2.1}$$

$$\underbrace{y_1, \ldots, y_{n-m}, y_{n-m+1}, \ldots, y_n}_{\text{observed}} \quad , \tag{2.2}$$

a missing value $x_j, j = n-m+1, \ldots, n$, is imputed by choosing that value x_i, $1 \leq i \leq n-m$, which is the nearest neighbor of j. The nearest neighborhood is measured in y–values such that i satisfies

$$\mid y_i - y_j \mid \quad = \quad \min_{1 \leq l \leq n-m} \mid y_l - y_j \mid . \tag{2.3}$$

In case the solution is not unique the mean of the corresponding x–values is imputed. One may also employ other solutions, especially for categorical covariates for this same problem.

The nearest neighbor imputation is a hot deck imputation procedure which yields values unlikely to be nonsensical. Chen and Shao (2000)) show *that under some conditions, the nearest neighbor imputation method provides asymptotically unbiased and consistent estimators or functions of population means (or totals), population distributions, and populations quantiles*, Chen and Shao (2000)), p. 1. Since it is a nonparametric method, so it is expected to be somewhat more robust against model violations. Chen and Shao (2001)) give a detailed overview on several possibilities for adjusting the procedure in order to get asymptotically unbiased and consistent variance estimates.

2.1.2 Nearest Neighbor Imputation—Version II (NN2)

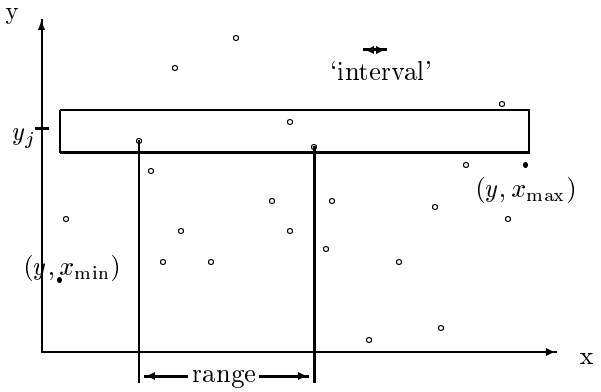

Figure 2.1: Fixed neighborhood based on $k = 3$.

As the nearest neighbor may also lead to substitutes which are far away from the 'true' value, a modified version of the NN1 is implemented which is based on some plausibility. Let x_j again denote a missing value and y_j be the corresponding response value. Consider a neighborhood of y_j based on a fixed

number of neighbors k. The main idea is to control the range of this fixed neighborhood in comparison to a percentage of the length of the data interval ('interval'). Figure 2.1 illustrates a data situation with three nearest neighbors, an artificial range between the three nearest neighbors and the reference value 'interval' with a 5%–data rate.

The neighborhood is defined according to (2.3) whose solution for $k = 3$ is a (3×1)–vector containing the ordered values $x_{[s]}$ for $s = 1, 2, 3$ satisfying (2.3). The range for $k = 3$ defined by $x_{[3]} - x_{[1]}$ is the first essential value for the procedure of the NN2; 'interval' is defined as a fixed percentage, which is 5% in our case, of $x_{\max} - x_{\min}$ and should be a reference value for the range of the neighborhood. A more detailed introduction to the NN2, especially to its algorithm is given in Nittner (2002).

The complete case analysis could also be seen as an imputation procedure—which, for example, is the case when the missing values themselves are estimated like unknown parameters and is equivalent to Bartlett's analysis of covariance, see Bartlett (1937) or its presentation in Rao and Toutenburg (1999), pp. 247–248.

2.2 Complete Case Analysis

The complete case analysis (CCA), also known as listwise deletion, simply discards all cases containing at least one missing value. Based on the partitioning according to (1.2) the analysis reduces to the estimation of

$$y_{\text{obs}} = f(X_{\text{obs}}) + \epsilon_{\text{obs}} \,. \tag{2.4}$$

An apparent problem is wastage of information which reaches its maximum when the number of deleted cases equals the number of missing values. Estimates may also be biased if there is stratified data. According to Schafer (1997) the CCA is thought to be suitable up to a percentage of 5% of missing values.

The estimates of the CCA are unbiased if the missingness does not depend on y, i.e., if $f(R \mid y, X) = f(R \mid X)$ holds. Then, we have

$$f(y \mid R, X) = \frac{f(y, R \mid X)}{f(R \mid X)} = \frac{f(R \mid y, X) f(y \mid X)}{f(R \mid X)} = f(y \mid X) \,. \tag{2.5}$$

Equation (2.5) means that the conditional density of the response vector y given R and $X = (X_{\text{obs}}, X_{\text{mis}})$ is independent of the value of R, i.e., the conditional expectation of $f(y \mid x)$ is the same for $R = 0$ and $R = 1$. This yields unbiased estimates for an analysis based on the complete cases if the missingness does not depend on y. In the simulation experiment based on the model (1.1) missing at random means that the missingness depends on the response y. Therefore, (2.5) is not fulfilled and the CCA is expected to show worse results in the sense of biased estimates.

2.3 Stochastic Mean Imputation

The stochastic mean imputation is an extension to the classical mean imputation, also known as *zero order regression* (ZOR) first described in Wilks (1932). The ZOR also interesting for users doing analysis with popular software where this method often is implemented. A missing value x_{ij} is imputed by

$$\hat{x}_{ij} = \bar{x}_j = \frac{1}{n - m_j} \sum_{i \notin \Phi_j} x_{ij} \,, \tag{2.6}$$

where $\Phi_j = \{i : x_{ij} \text{ missing}\}$ denotes the indices of the missing values and m_j denotes the number of missing values for X_j. If X_j is discontinuous then mode and median are the other well suited alternatives.

The most important disadvantage of the ZOR is that it underestimates the variance which in turn distorts the corresponding tests because of resulting in small confidence intervals. That is why the imputed value is modified in terms of an additive random error; this procedure is denoted by ZOR+, the zero order regression plus random noise which is a kind of stochastic mean imputation.

A procedure reflecting more reference to the distribution of the population may be the single imputation which is described in the next section.

2.4 Single Imputation

In comparison to the mean imputation the single imputation (see for example Little and Rubin (1987)) should provide substitutes representing more variation in terms of the postulated distribution than the ZOR+. As already discussed, the single imputation (sI) could be based on the distribution of the complete cases, i.e., impute a random number out of the distribution characterized by its estimated parameters. This distribution sometimes is known. Otherwise one may consider conditional distributions based on the complete cases and on an auxiliary model. This can be used for predicting the missing values which, however, here is not of interest because of having just the simple model $y = f(x) + \epsilon$ with one covariate.

Example 1 Assume a linear model $y = X\beta + \epsilon$ where $X = (\mathbf{1}, X_2, X_3)$ where X_3 is supposed to be binary and partially incomplete. Compute an auxiliary model, for example, a logistic regression for the complete cases with X_3 being the response vector, X_1, X_2 and may be y representing the independent variables. The resulting estimates are used to compute conditional probabilities π_j via the logit link using the values of the observed variables X_1, X_2, y for the missing indices. The π_j could be considered as parameters of row–wise binomial distributions which motivate the following imputation steps for $j = n - m + 1, \ldots, n$

1. Draw a random number z_i from a continuous uniform distribution over the interval [0;1]

2. Impute

$$x_j = \begin{cases} \text{'1'} & \text{if} \quad z_j \leq \pi_j \\ \text{'0'} & \text{if} \quad z_j > \pi_j \end{cases} . \tag{2.7}$$

See Toutenburg and Nittner (2002) for some simulation results for model.

3 A Simulation Experiment

After a short introduction to the simulation experiment concerning technical details and parameter settings, the results are analyzed and reported. Based on the sample mean squared error (SMSE) and its components, the sample bias and the sample variance as well as differences among the methods are illustrated.

3.1 A Short Introduction

The simulation experiment was conducted using R programming language (see Venables and Smith (2001)). The time which took an experiment to run depended on the missing percentage and was between 11 hours for 10% and 30 hours for 50% missingness.

We assumed X to be truncated Gaussian with mean $\nu = 1.0$ which is the middle of the data interval $[0.0; 2.0]$ and fixed standard deviation $\delta = 0.5$. The errors were assumed to be distributed Gaussian according as $\epsilon_i \sim N(0; \sigma^2)$ with σ being 0.5, 1.0 and 2.0 for each setting. With three values of σ and 10%, 30% and 50% missing percentage m_p, nine models were computed, see Table 3.1. The sample size was chosen to be $N = 1000$, the number of replications also

model	1	2	3	4	5	6	7	8	9
m_p	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
σ	0.5	0.5	0.5	1.0	1.0	1.0	2.0	2.0	2.0

Table 3.1: Settings of the simulation experiment.

were 1000. With X_i and ϵ_i being distributed as described above, the response vector y was computed according to

$$y_i = x_i - 4x_i^2 + 2x_i^3 + \epsilon_i \, .$$

See Figure 3.1 to get an impression what $f(x)$ looks like in the simulation experiment.

The usual algorithms generate data missing at random according to (see for example Little (1992),

$$r_{ij} = \begin{cases} 0 & \text{für} \quad U_i > 0 \\ 1 & \text{für} \quad U_i \leq 0 \end{cases} \tag{3.1}$$

with $U_i = \alpha y_i + \tau_i$ which is a random variable depending on the value of the standardized y_i, and a standard normal error τ_i with a disadvantage of not

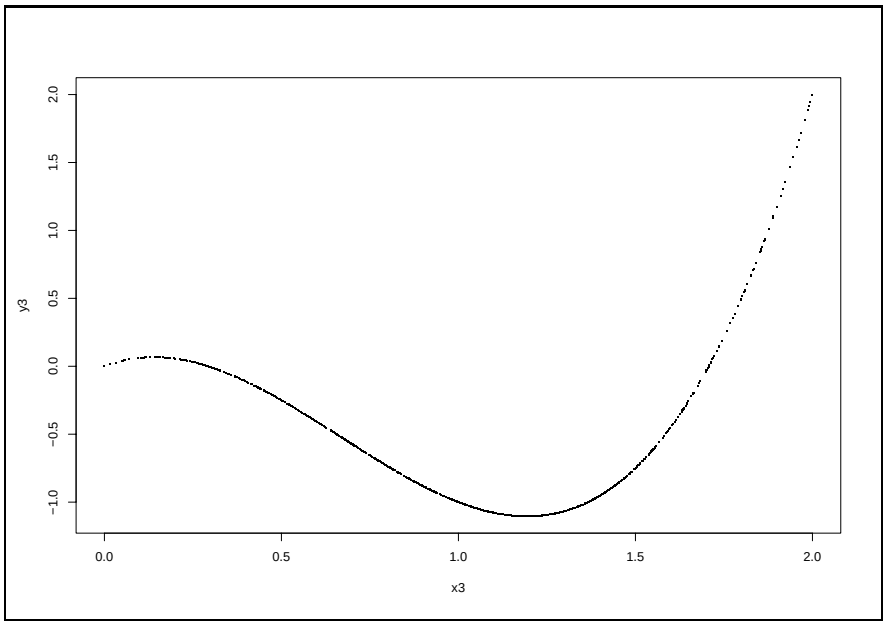

Figure 3.1: $f(x)$ in the simulation experiment.

having a fixed missing percentage a priori. This is the reason why a slight modification was applied within this experiment: If the fixed missing percentage m_p had reached for $i < N$, the algorithm stopped; if $i = N$ and the missing percentage was smaller than it should be the algorithm started again for the cases still observed. Fixed percentages of missing values and fixed parameter values in general enabled us to compare results of different simulation experiments in a more steady way.

3.2 Results

In this section, the sample mean squared error (SMSE), the sample variance and the estimated bias of the different predicted values $\hat{y}$ are compared at the ten fixed knots, i.e., each curve is estimated based on the imputation procedure and the filled up matrix X, respectively. The ten fixed knots of the 'true' model are used to compute the predicted values $\hat{y}_j, j = 1, \ldots, 10$, in order to be able to compare these values.

Three settings (models 6,8 and 9) especially lead for the single imputation and the modified nearest neighbor imputation to large values of the global smoothing parameter λ. For some replications the confidence band for the estimated curve included the zero everywhere so that $f(X)$ could be substituted by a parametric term, i.e., for these settings single and nearest neighbor imputation NN2 should be left out for comparison with the alternatives.

3.2.1 The Sample Mean Squared Error (SMSE)

The sample mean squared error of $\hat{y}$ is given by

$$\widehat{\mathrm{SMSE}}(\hat{y}, y) = \sum_{j=1}^{\text{knots}} \hat{\mathrm{V}}(\hat{y}_j) + [\hat{\mathrm{B}}(\hat{y}_j, y_j)]^2 \,. \tag{3.2}$$

Analyzing the sum of the ranks of all procedures for all nine experiments showed worse results for the complete case analysis and the stochastic mean imputation whereas good values for the classical version of the nearest neighbor imputation. The exact values are summarized in Table 3.2. It can be seen from Table 3.2

	σ								
	0.5			1.0			2.0		
	m_p			m_p			m_p		
model	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
TRUE	0.099	0.095	0.094	0.251	0.228	0.307	0.902	0.862	0.747
CCA	0.106	0.250	2.004	0.336	1.062	7.641	1.298	5.667	25.604
ZOR+	0.195	0.733	2.495	0.429	1.362	5.837	1.317	5.159	17.754
sI	0.197	0.876	3.609	0.384	0.896	4.562	1.215	2.600	6.441
NN1	0.113	0.145	0.884	0.283	0.352	2.484	0.993	1.537	10.348
NN2	0.111	0.238	2.335	0.302	0.530	4.845	1.239	3.899	10.962

Table 3.2: SMSE for all nine models depending on m_p and σ.

that the SMSE of the imputation increases for rising values of the error variance as well as for an increasing percentage of missing values. These trends could be seen in Figure 3.2 for $\sigma = 0.5$ and $\sigma = 1.0$ and for $\sigma = 2.0$ in Figure 3.3. White bars correspond to a missing percentage of 10%, grey bars to $m_p = 0.3$ and dark grey bars to $m_p = 0.5$. With respect to the problems in the models

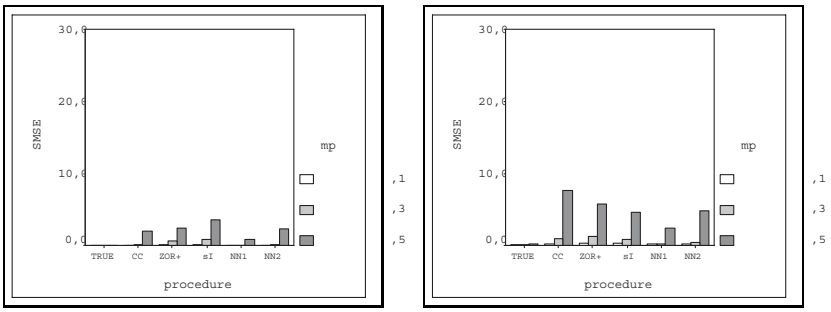

Figure 3.2: SMSE for $\sigma = 0.5$ (left) and $\sigma = 1.0$ (right).

6, 8 and 9, the classical **nearest neighbor imputation** obviously seems to be the best procedure concerning the SMSE–criterion. Despite its additive random error, the **stochastic mean imputation** is not adequate because of the large values of the SMSE which is apparent more for a larger missing percentage. The **complete case analysis** especially shows for $m_p > 10\%$ the behavior expected because of MAR. Its SMSEs are even larger than those of the ZOR+. The relative position of the **single imputation** gets better with an increasing error. Comparing the two versions of the nearest neighbor imputations it could be said

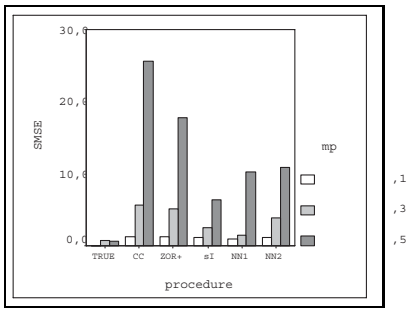

Figure 3.3: SMSE for $\sigma = 2.0$.

that the **modified nearest neighbor imputation** tends to have similar but slightly worse values than the classical version.

To summarize, we have seen an apparent trend of the different procedures concerning their behavior depending on σ and m_p and their SMSE–superiority among themselves. In the following section, the two components of the sample mean squared error are analyzed with an aim to detect possible reasons in general or trade–offs between variance and bias.

3.2.2 The Sample Variance

First of all we are interested in the behavior of the estimated variances which depend on the two simulation parameters σ and m_p. Like the SMSE, the estimated variance also increases with an increase in the error variance. See Figure 3.4, where the sample variances for the complete case analysis and the stochastic mean imputation are shown—for ease of presentation the ordinate was changed to logarithmic scale.

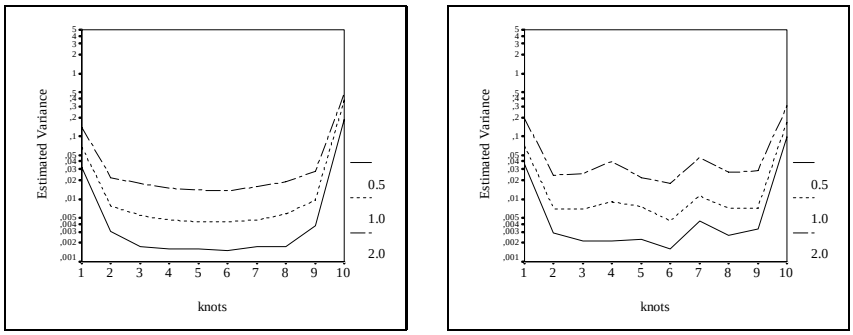

Figure 3.4: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for $m_p = 0.5$; continuous line for $\sigma = 0.5$, dashed line for $\sigma = 1.0$, semi–dashed line for $\sigma = 2.0$; CCA (left) and ZOR+ (right).

The behavior of the variance depending on the missing percentage is not as clear as the one observed for σ. Whereas both nearest neighbor imputations have increasing variances with a raising percentage of missing values the alter-

natives seem to be independent of the value of m_p concerning their estimated variances. Figure 3.5 illustrates the situation for the single imputation and the classical nearest neighbor imputation for $\sigma = 2.0$.

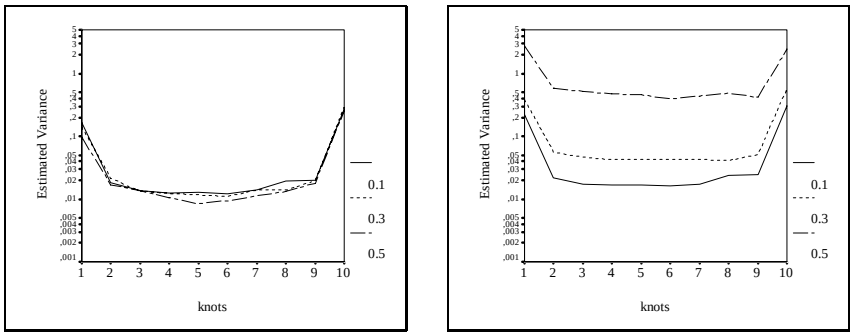

Figure 3.5: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for $\sigma = 2.0$; continuous line for $m_p = 0.1$, dashed line for $m_p = 0.3$, semi–dashed line for $m_p = 0.5$; sI (left) and NN1 (right).

Now we discuss the variances of the different estimates arising from the imputation methods. Two groups are built up to compare the procedures. The first group compares the stochastic mean imputation with the single imputation whereas the second group compares the complete case analysis with the classical nearest neighbor imputation. The modified nearest neighbor imputation is analyzed with respect to the classical version at the end of this paragraph. Each graphic represents the situation for a fixed missing percentage with a varying error variance.

The differences between the ZOR+ and the single imputation are small even if trends could be observed. Figure 3.6 shows that the single imputation tends to have smaller variances than the ZOR+ and shows a more continuous behavior over the knots which becomes more apparent with an increasing percentage of missing values.

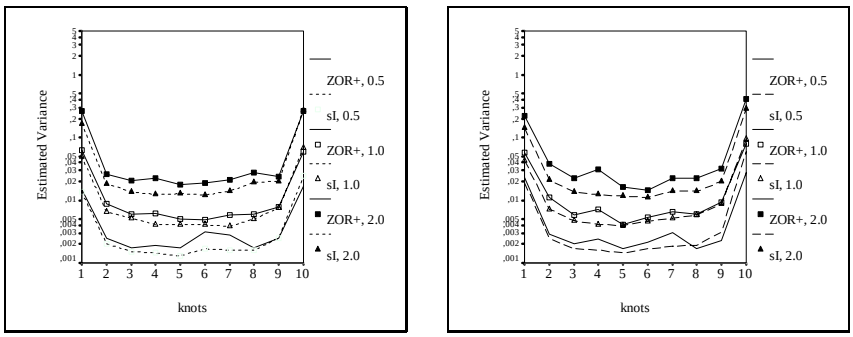

Figure 3.6: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for the ZOR+ (continuous line) and the single imputation (dashed line); $m_p = 0.1$ (left–hand side), $m_p = 0.3$ (right–hand side); different σ marked by squares and triangles.

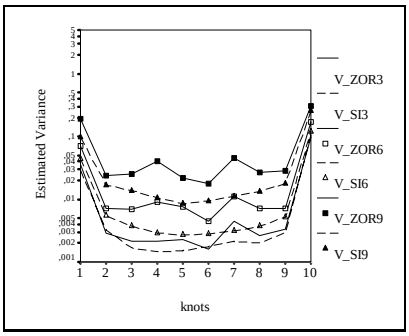

Figure 3.7: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for the ZOR+ (continuous line) and the single imputation (dashed line); $m_p = 0.5$; different σ marked by squares and triangles.

Figure 3.7 confirms this result, especially the more wiggly curve of the stochastic mean imputation. The differences between these two methods seem to rise with an increase in the missing percentage.

The second group which compares the complete case analysis and the nearest neighbor imputation is illustrated in Figures 3.8 and 3.9.

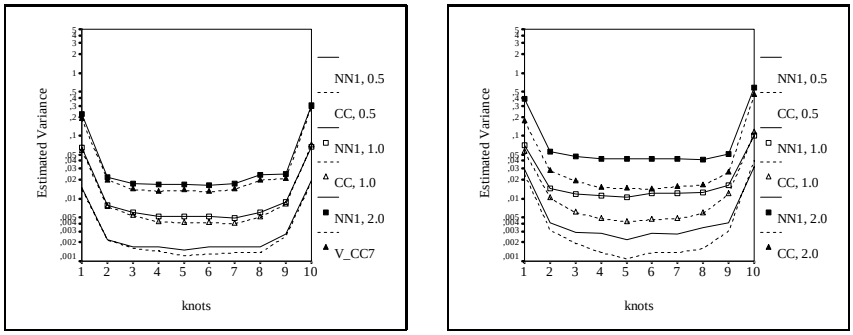

Figure 3.8: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for the NN1 (continuous line) and the complete case analysis (dashed line); $m_p = 0.1$ (left–hand side), $m_p = 0.3$ (right–hand side); different σ marked by squares and triangles.

Similar differences in this group could also be observed. We see that the nearest neighbor imputation for every setting tends to have larger variances than the complete case analysis. This cohesion becomes more clear with an increase in the missing percentage and leads to the extreme situation as in Figure 3.9 where for $\sigma = 2.0$, the CCA has smaller variances than the NN1 for $\sigma = 0.5$. The fact that the NN1 has superiorities concerning the SMSE therefore has to result from noticeable smaller biases of the NN1 in comparison to the CCA.

Finally we want to consider the estimated variances of the two nearest neighbor imputations to get an idea to what extent they differ. Figure 3.10 shows that the estimated variances of both nearest neighbor methods depend on the error variance. An illustration of the situation for $m_p = 0.5$ was ignored because of

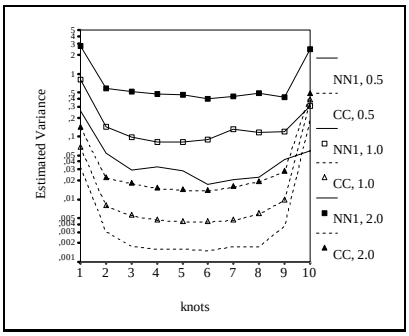

Figure 3.9: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for the NN1 (continuous line) and the complete case analysis (dashed line); $m_p = 0.5$; different σ marked by squares and triangles.

the presumed linear relation.

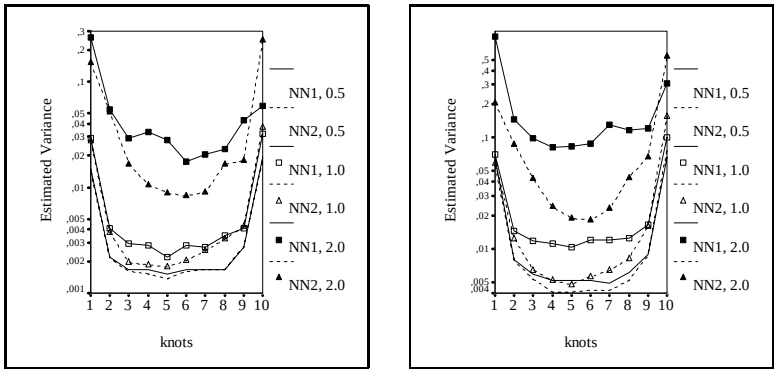

Figure 3.10: $\widehat{\mathrm{Var}}_{\hat{\beta}}$ for the NN1 (continuous line) and the NN2 (dashed line); $m_p = 0.1$ (left–hand side), $m_p = 0.3$ (right–hand side); different σ marked by squares and triangles.

As seen earlier, the difference between the two methods get larger with an increase in the missing percentage. The modified nearest neighbor imputation tends to show smaller variance than the classical version, except in some cases at the outer knots. Again, the superiority of the NN1 versus NN2 in terms of the SMSE is presumed to result because of magnitude of bias.

So far, some noticeable differences between the methods have been observed—however, the analysis of the bias may explain the synthesis of the SMSE.

3.2.3 The Sample Bias

In this section we first take a look at the bias depending on σ and m_p and then at the differences between the methods.

The bias tends to increase with an increase in the error variance and missing percentage both. This context just *tends* to hold because of local differences in the sense of knot location and type of procedure. The complete case analysis shows an obvious increase of the bias. The imputation methods, especially the ZOR+, tend to follow this behavior in the center of the interval.

Next, we want to take a look at the groupwise consideration between the methods as in the previous section. Figures 3.11 and 3.12 show the bias for fixed missing percentages with varying error variances.

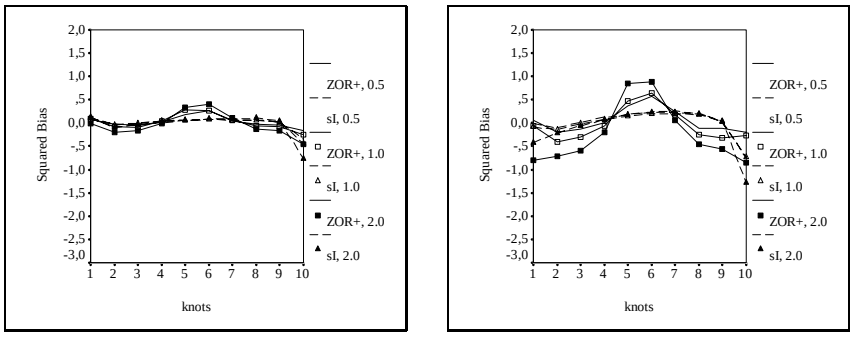

Figure 3.11: $\widehat{\text{Bias}}(\hat{\beta}, \beta)$ for the ZOR+ (continuous line) and the single imputation (dashed line); $m_p = 0.1$ (left–hand side), $m_p = 0.3$ (right–hand side); different σ marked by squares and triangles.

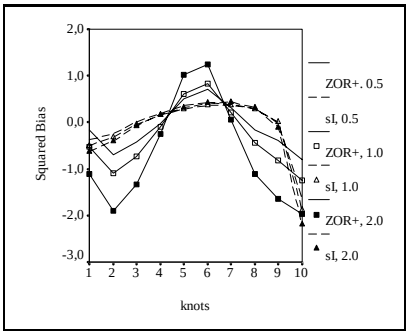

Figure 3.12: $\widehat{\text{Bias}}(\hat{\beta}, \beta)$ for the ZOR+ (continuous line) and the single imputation (dashed line); $m_p = 0.5$; different σ marked by squares and triangles.

The stochastic mean imputation tends to be more biased than the single imputation except to the values at the right outer knot. As expected, the ZOR+ overestimates in the center of the interval and indicates underestimation at the margins. The differences between the methods become larger with an increase in the missing percentage.

The nearest neighbor imputation NN1 is less biased than the complete case analysis with respect to the analysis of the SMSE and the sample variance. Figure 3.13 illustrates the bias for $m_p = 0.1$ and $m_p = 0.3$ for different values of σ,

Figure 3.14 contains the values for a missing percentage of 50%.

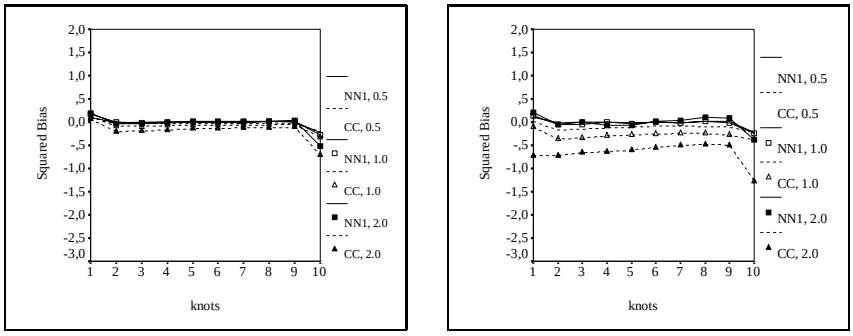

Figure 3.13: $\widehat{\text{Bias}}(\hat{\beta}, \beta)$ for the Nn1 (continuous line) and the complete case analysis (dashed line); $m_p = 0.1$ (left–hand side), $m_p = 0.3$ (right–hand side); different σ marked by squares and triangles.

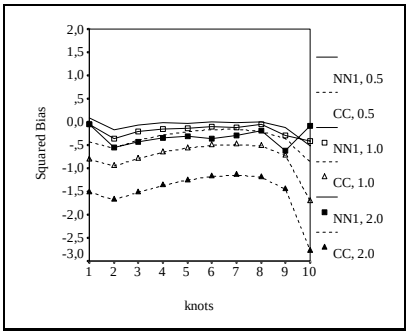

Figure 3.14: $\widehat{\text{Bias}}(\hat{\beta}, \beta)$ for the NN1 (continuous line) and the complete case analysis (dashed line); $m_p = 0.5$; different σ marked by squares and triangles.

For both methods, the bias increases with an increase in the missing percentage. The difference between the methods themselves becomes more clear with an increase in m_p also. Unbiased estimates of the NN1 and an underestimation of the complete case estimators are obvious.

A comparison of the two nearest neighbor imputations yielded advantages for the classical version. The modified version shows a similar behavior as the ZOR+ which indicates an overestimation in the center and an underestimation on the margins of the interval.

4 Conclusion

The additive model $y = f(x) + \epsilon$ with missing values in the independent variable x depending on the response vector y, meaning missing at random, was considered. For nine different settings resulting from three values of m_p and σ, four imputation procedures were compared with the complete case analysis.

The **complete case analysis** is not suitable when more than 10% observations are missing. Its large SMSEs result from a large sample bias. The **zero order regression plus random noise** as a popular but shady standard procedure shows its well known properties. An ad hoc alternative, the **single imputation** seems to be somewhat more suited because of its advantages versus the ZOR+. However, in some simulations its term changed to be parametric whereas the **classical nearest neighbor imputation** showed best properties for each setting and seems to be adequate. Its modified version is more or less strongly biased and yields good values for the SMSE because of small variances.

Some more work has to be done within this context, for example, including further nonparametric or parametric terms and interactions. This would enable a kind of first order regression which also could be modeled nonparametrically. One could also think of alternatives based on higher–order distance measures or including cluster analysis.

References

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied botany, *Journal of the Royal Statistical Society, Series B* **4**: 137–170.

Chen, J. and Shao, J. (2000). Biases and variances of survey estimators based on nearest neighbor imputation, *Technical report*, University of Wisconsin-Madison.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest–neighbor imputation, *Journal of the American Statistical Association* **96**(453): 260–269.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2 edn, Springer–Verlag, New York.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

Little, R. J. A. (1992). Regression with missing X's: A review, *Journal of the American Statistical Association* **87**(420): 1227–1237.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Nittner, T. (2002). The additive model with missing values in the independent variable - theory and simulation., *SFB386–Discussion Paper 272*, Ludwig-Maximilians-Universität München.

Rao, C. R. and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*, 2 edn, Springer–Verlag, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Toutenburg, H. and Nittner, T. (2002). Linear regression models with incomplete categorical covariates, *Computational Statistic* **17**(2): 215–232.

Venables, W. and Smith, D. (2001). *An Introduction to R*.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples, *Annals of Mathematical Statistics* **3**: 163–195.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society, Series B* **62**(2): 413–428.