JIRSS (2002)
Vol. 1, Nos. 1-2, pp 79-109

Parametric and Nonparametric Regression with Missing X's—A Review

Helge Toutenburg, Christian Heumann, Thomas Nittner, Sandro Scheid

Department for Statistics, Ludwig–Maximilians–University Munich, Ludwigstr. 33, 80539 Munich, Germany.
(toutenb@stat.uni-muenchen.de, heumann@stat.uni-muenchen.de, nittner@stat.uni-muenchen.de, scheid@stat.uni-muenchen.de)

Abstract. This paper gives a detailed overview of the problem of missing data in parametric and nonparametric regression. Theoretical basics, properties as well as simulation results may help the reader to get familiar with the common problem of incomplete data sets. Of course, not all occurences can be discussed so this paper could be seen as an introduction to missing data within regression analysis and as an extension to the early paper of [19].

Received: September 2002
Key words and phrases: Generalized additive models, imputation, missing data mechanism, MSE–superiority, regression analysis.

1 Introduction

Statistical analysis with missing data is a common problem in practice. Nonresponse in sample surveys or drop–out in clinical trials may be two of many examples one could imagine. Apart from the estimation of sample statistics regression analysis as a main tool within statistical analyses of dependencies therefore often is affected by missing values, too. Whereas parametric regression has been investigated extensively, nonparametric methods haven't been considered within this context so far. Apart from the standard literature concerning missing data, i.e., [20] and [26], linear regression (e.g. [19]), logistic regression (e.g. [40]) and generalized linear models (e.g. [18]) were considered within the scope of parametric methods. Little emphasis has been put on nonparametric regression analysis—e.g., [7] or some simulation experiments by [22].

Before defining the basic terms and relations within the context of missing data, the parametric as well as the nonparametric regression model is introduced.

1.1 Linear Regression Models

Assume the linear regression model

$$Y = X_1\beta_1 + \ldots + X_p\beta_p + \epsilon \tag{1}$$

and its sample version

$$y = X\beta + \epsilon \tag{2}$$

where y is the $(n\times 1)$–vector of observations of the dependent variable, X is the $(n \times p)$–matrix of the independent regressiors and ϵ is the $(n \times 1)$–vector of disturbances. We confine ourselves to nonstochastic X and assume X to be of full column rank. Further let

$$\epsilon \sim (\sigma^2, I_n) \tag{3}$$

or

$$\epsilon \sim N(\sigma^2, I_n) \tag{4}$$

for testing hypothesis. If X is complete, the BLUE of β is given by

$$b = (X'X)^{-1}X'y. \tag{5}$$

In statistical practice, however, we often have incomplete data— marked by '$*$'— in the response y as well as in the data matrix, i. e.

$$(yX) = \begin{pmatrix} y_1 & x_{11} & \cdots & \cdots & x_{1p} \\ y_2 & \vdots & * & & \vdots \\ * & & & & * \\ \vdots & \vdots & & * & \vdots \\ y_n & x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix} \tag{6}$$

In general, we may assume the following structure of the data which is discussed in full detail in [24], Chapter 6,

$$\begin{pmatrix} y_{\text{obs}} \\ y_{\text{mis}} \\ y^*_{\text{obs}} \end{pmatrix} = \begin{pmatrix} X_{\text{obs}} \\ X^*_{\text{obs}} \\ X_{\text{mis}} \end{pmatrix} \beta + \epsilon\,. \tag{7}$$

Estimation of y_{mis} corresponds to the prediction problem. Based on these results, we may confine ourselves to the substructure

$$\begin{pmatrix} y_{\text{obs}} \\ y^*_{\text{obs}} \end{pmatrix} = \begin{pmatrix} X_{\text{obs}} \\ X_{\text{mis}} \end{pmatrix} \beta + \epsilon \tag{8}$$

of (7) and change the notation as follows:

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \quad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim (0, \sigma^2 \mathrm{I}). \tag{9}$$

The submodel

$$y_c = X_c\beta + \epsilon_c \tag{10}$$

stands for the completely observed data (c : complete), and we have $y_c : \ m \ \times \ 1, X_c : m \ \times \ p$, and $\text{rank}(X_c) = p$.

The other submodel

$$y_* = X_*\beta + \epsilon_* \tag{11}$$

is of dimension $(n-m) = J$. The vector y_* is observed completely. In the matrix X_* some observations are missing. The notation X_* will underline that X_* is partially incomplete, in contrast to the matrix X_{mis}, which is completely missing. Combining both of the submodels in model (9) corresponds to the so–called mixed model ([32]). Therefore, it seems to be natural to use the method of mixed estimation.

1.2 Generalized Additive Models

Generalized Additive Models (GAM) became more and more popular since the work of [16]. One could consider GAMs as the generalization of linear as well as generalized linear models and, of course, additive models. Its flexibility concerning the modelling of the functional relation build the main advantage over linear models and generalized linear models (GLM) based on the a priori unknown function $f(X)$ which has, for example, to be specified within polynomial regression when the purpose is a non–linear relation between y and X.

Before introducing the GAM a short summary is given to generalized linear models to get familiar with the necessary terms. Following the notation within the previous section the observations y_i are assumed to be independent identically distributed with $\mu_i = \mathrm{E}(y_i)$, the mean given by $\mu_i = x_i'\beta$. GLMs may then be described by

1. the *distribution assumption* which postulates the y_i to be conditionally independent of the x_i with the conditional distribution of y_i belonging to a simple exponential family with $\mu_i = \mathrm{E}(y_i \mid x_i)$ and a scaling parameter ϕ.
2. the *structural assumption* which relates μ_i with the linear predictor $\eta_i = x_i'\beta$ according to

$$\mu_i = h(\eta_i) = h(x_i'\beta), \quad \text{resp.} \quad \eta_i = g(\mu_i), \tag{12}$$

   with the one–to–one known function h and g being the inverse function of h called link function.

Following [12] a generalized linear model is characterized by the type of the exponential family, the link function, and the design vector x_i.

Generalized additive models differ from generalized linear models by assuming an *additive* predictor instead of a linear predictor and are defined by

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j)\,, \tag{13}$$

with an appropriate link function. Partition (8) and, especially, the distribution of the errors noted in (9) are assumed to hold here, too. The mixed estimator within this context can not be written in such

a way which forces us to introduce the inference within GAMs just in general.

Similar to the minimization of the target function within the linear model we formulate a target function with respect to the smoothness of $f(x)$ according to

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx\,, \tag{14}$$

which is to be minimized with respect to the parameters of $f(x)$. f' and f'' have to be continuous, f'' has to be quadratically integrable. λ controls the trade–off between variance and bias well known from a simple scatterplot smoother. $\lambda \to \infty$ equals a straight line, $\lambda = 0$ leads to an unsmoothed estimate $\hat{f}(x_i) = y_i$ meaning a reproduction of the data, see [39].

The estimation is done by *iteratively re–weighted least squares* (IRLS) where each least–squares step is replaced by a penalized one. The criterion is the maximization of the penalized loglikelihood

$$l_\beta = \lambda \sum_{i=1}^{n} l_i(y_i, X_i\beta) - \frac{1}{2}\sum_{i=1}^{m} \theta_i \beta' S_i \beta\,, \tag{15}$$

where m quadratic penalties are to be applied to the parameter vector β. The matrix S_i contains the penalties which imply each smoothing parameter θ_i. $\beta^{(k)}$ is estimated by Fisher–Scoring Algorithmus, see Table 1.

$\beta^{(k+1)}$	$=$	$\beta^{(k)} + \mathrm{E}[-\frac{\delta^2 l_\beta}{\delta\beta^{(k)}\delta\beta^{('k)}}]^{-1}\frac{\delta l_\beta}{\delta\beta^{(k)}}$	respectively,
$\beta^{(k+1)}$	$=$	$\beta^{(k)} + [X'W^{(k)}X\lambda + \sum\theta_i S_i]^{-1}\{X'W^{(k)}\Gamma^{(k)}(y-\mu^{(k)}) - \sum\theta_i S_i\beta^{(k)}\}$	

Table 1: Fisher–Scoring.

$W_{ii}^{(k)} = g'(\mu_i^{(k)})V_i^{-1}$ is a weighting matrix, V_i is the variance of y with respect to $\mu_i^{(k)}$; $\Gamma_{ii} = g'(\mu_i^{(k)})$. $g(\mu_i)$ is a monotone link function.

Following [43], the determination of $\beta^{(k+1)}$ within f with $\mathrm{E}(f(Y_i)) =$

1. By the help of $\beta^{(k)}$ one gets estimates for μ and the variances V_i for each y_i; compute

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

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

2'. Compute λ_i by minimizing $\frac{||W^{\frac{1}{2}}(z - X\beta)||^2}{(\mathrm{sp}(I-A))^2}$ with β being the solution of minimizing $|| W^{\frac{1}{2}}(z - X\beta) ||^2 + \sum \lambda_j \beta' S_j \beta$ with respect to β and A being the hat–matrix with $A = X(X'WX + \sum \lambda_j \beta' S_j \beta)^{-1} X'W$.

Table 2: IRLS with GCV.

$f(\beta)$ is equivalent to solving the weighted penalized least squares problem

$$\min\lambda \; || W^{\frac{1}{2}}(z^{(k)} - X\beta) ||^2 + \sum \theta_i \beta' S_i \beta \tag{16}$$

with a pseudo data vector $z^{(k)} = X\beta^{(k)} + \Gamma^{(k)}(y - \mu^{(k)})$, the global smoothing parameter λ, the diagonal matrix of weights W and the nonnegative definite matrix S of coefficients containing the penalty terms θ_i for the smoothing parameters. (16) in practice is solved by minimizing the *generalized cross validation* (GCV) scores

$$V = \frac{|| W^{\frac{1}{2}}(y - A(\lambda, \theta)y) ||^2 / n}{[1 - \mathrm{tr}(A(\lambda, \theta))/n]^2} \tag{17}$$

with respect to $\frac{\theta_i}{\lambda}$. $\hat{\mu} = Ay$ holds for the hat–matrix A. Combining IRLS and GCV represents the algorithm of interest, illustrated in Table 2 (see [14] or [43]).

Note that the degrees of freedom is an integrative part of the estimation. An extensive description of estimating GAMs can be found in [39].

2 Missing Data Pattern and Missing Data Mechanism

So far, we introduced just methods and assumptions analyzing the data set as it is. Because of the effects of the missingness, i.e. the amount of missing values, on the data structure and, therefore, the amount of information the analyst has to deal with these problems. The missing data pattern and the missing data mechanism are two important terms visualizing and characterizing the situation of the data.

2.1 Missing Data Pattern

As already mentioned above, visualizing the structure of the data set with respect to the missing values may be a first way to get an impression of the situation. Also implemented in statistical software, the missing data pattern do a good job; the observed cases of a variable correspond to one bar—the more missing values, the shorter the bar of the variable, see Figures 1–4. Figure 1 shows the situation when one variable is incomplete and all other variables are completely observed—a special case of the monotone pattern in Figure 2 where each variable X_j is observed for at least the cases of X_{j-1}. An example for a special pattern is shown in Figure 3 where X_2 is observed for the cases where X_1 is missing and vice–versa. This is a common problem known as double sampling, see [26]. Figure 4 illustrates a situation with no special structure.

Although the missing data pattern represent an easy way to get a first impression, more complex dependencies between observed and incomplete variables or incomplete variables themselves will reduce this ability strongly. This is one reason why the missing data mechanism has to be considered.

2.2 Missing Data Mechanisms (MDM)

The main question within the context of analyzing incomplete data sets is whether the missing data mechanism can be 'ignored'—a term

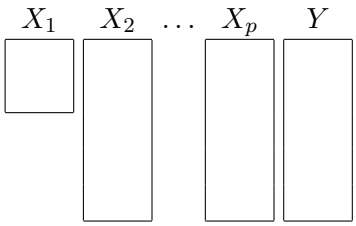

Figure 1: Univariate Missing Data Pattern

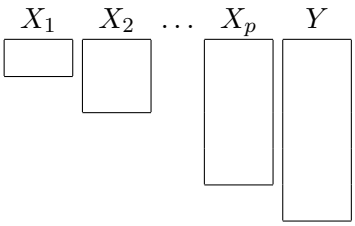

Figure 2: Monotone Missing Data Pattern

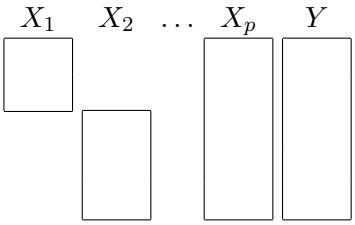

Figure 3: Special Missing Data Pattern

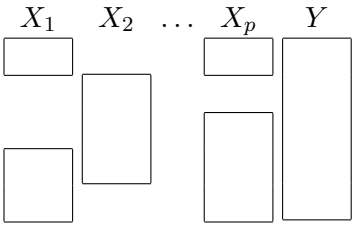

Figure 4: General Missing Data Pattern

which is to be specified—or not. One could make the assumption that the mechanism is ignorable in the sense described below; the other possibility consists of including the missing data mechanism—which still is to define—in the statistical model. Including the MDM means including the distribution of an indicator variable R indicating if a component of the data matrix Z is observed or missing. [20] define the data matrix $Z = (Z_{\text{obs}}, Z_{\text{mis}})$ representing the data that would occur without missing values. The random variable R indicating the missingness within the data matrix Z is defined according to

$$r_{ij} = \begin{cases} 1 & \text{if} \quad z_{ij} \quad \text{observed} \\ 0 & \text{if} \quad z_{ij} \quad \text{missing} \end{cases} \quad \forall\, i = 1, \ldots, n,\ j = 1, \ldots, p+1. \quad (18)$$

The question whether the missing mechanism can be 'ignored' for the estimation of θ equals the question whether statistical inference is based on the density $f(Z_{\text{obs}}, R \mid \theta, \Phi)$—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 on the simpler density $f(Z_{\text{obs}}, \theta)$ which is 'ignoring' the missing mechanism. The classification of the missing data mechanism is based on the density $f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)$ and leads

to the definition of

1. MCAR (missing completely at random) if

$$f(R \mid Z, \Phi) = f(R \mid \Phi) \quad \forall Z\,, \tag{19}$$

2. MAR (missing at random) if

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, \Phi) \quad \forall Z_{\text{mis}}\,, \text{ and} \tag{20}$$

3. MNAR (missing not at random)

$$f(R \mid Z, \Phi) = f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi)\,. \tag{21}$$

Following [20], 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 [(20)] themselves. This may be more apparent by computing the density of the actual observed data obtained by integrating Z_{mis} out of the density

$$f(Z_{\text{obs}}, R \mid \theta, \Phi) = \int f(Z_{\text{obs}}, Z_{\text{mis}} \mid \theta) f(R \mid Z_{\text{obs}}, Z_{\text{mis}}, \Phi) dZ_{\text{mis}} \tag{22}$$

which by help of (20) leads to

$$\begin{aligned} f(Z_{\text{obs}}, R, \theta, \Phi) &= f(R \mid Z_{\text{obs}}, \Phi) \int f(Z_{\text{obs}}, Z_{\text{mis}}, \theta) dZ_{\text{mis}} \\ &= f(R \mid Z_{\text{obs}}, \Phi) f(Z_{\text{obs}} \mid \theta)\,. \end{aligned} \tag{23}$$

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 if the parameters θ and Φ concerning the density of Z and the missing mechanism, respectively, are distinct in the sense of each parameter containing no information about the other (see for example [26]).

3 Inference and Missing Data

3.1 The Mixed Model with Missing Regressor Values

Following the concept of missing values introduced by [20] leads to a partition of the sample $(y_1, \ldots, y_n)$ in two samples, the first contains

all those y_i with completely given x_i–vectors (say $m < n$ elements of the sample).

$$y_c = X_c\beta + \epsilon_c \qquad \text{(subscript } c: \text{ complete)}. \tag{24}$$

In general we assume X_c to be of full column rank p.

The second sub–sample contains all those y_i where the associated x–rows are partially or fully unknown (say J elements of the sample, $m + J = n$).

$$y_* = X_*\beta + \epsilon_*\,, \qquad \epsilon_* \sim (0, \sigma^2 I_J). \tag{25}$$

That is, y_c, y_* and X_c are known, but X_* is partially or fully unknown. Combining (24) and (25) gives the mixed model

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix}\beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}. \qquad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim (0, \sigma^2 I_n)\,. \tag{26}$$

The optimal but due to the unknown elements of X_* not operational estimator (BLUE) of β is given by the mixed estimator (cf. [24], Ch. 5)

$$\begin{aligned} \hat{\beta}(X_*) &= (X_c^{'}X_c + X_*^{'}X_*)^{-1}(X_c^{'}y_c + X_*^{'}y_*) \\ &= b_c + S_c^{-1}X_*'(I_J + X_*S_c^{-1}X_*')^{-1}(y_* - X_*b_c) \end{aligned} \tag{27}$$

having the covariance matrix

$$\mathrm{V}(\hat{\beta}(X_*)) = \sigma^2(X_c^{'}X_c + X_*^{'}X_*)^{-1} = \sigma^2(S_c + S_*)^{-1} \tag{28}$$

where $S_* = X_*^{'}X_*$ and $S_c = X_c^{'}X_c$ and $b_c = (X_c'X_c)^{-1}X_c'y_c$ is the OLSE in the complete case submodel (24).

3.2 Common Missing Values Procedures

3.2.1 Complete Case Analysis

The first (and in many situations most obvious) method to obviate the problem of an incompletely observed design matrix results in resigning the incomplete model (25). This so–called classical LSE of β makes use of the completely observed design matrix X_c, only.

That is, the classical LSE (CLSE) estimates β from the model (1.2) according to

$$b_c = (X_c^{'} X_c)^{-1} X_c^{'} y_c \tag{29}$$

having

$$\mathrm{V}(b_c) = \sigma^2 (X_c^{'} X_c)^{-1} = \sigma^2 S_c^{-1} \,. \tag{30}$$

The CLSE discards the partial information contained in y_* and the observed elements of X_* of the incomplete model (25). This may lead to a loss in efficiency compared with estimators using a "repaired" version of model (25) whereas repairing means to fill the gaps in X_* by some substitution method, for example.

3.2.2 Available Case Analysis

These methods estimate β from normal equations (see [15]) according to

$$\mathrm{cov}(x_i x_j)\hat{\beta} = \mathrm{cov}(x_i y) \qquad (i, j = 1, \ldots, p) \,, \tag{31}$$

where $\mathrm{cov}(x_i x_j)$ is the $p \times p$ covariance matrix with the (i, j)th element $(i, j = 1, \ldots, p)$ computed from the observations common to both x_i and $x_j (i \neq j)$ as well as from all existing measurements on x_i for $i = j$. Similarly, $\mathrm{cov}(x_i y)$ is computed from all measurements common to both x_i and $y \, (i = 1, \ldots, p)$.
¿From (31) we arrive at

$$\mathrm{cov}(x_i x_j)\mathrm{E}(\hat{\beta}) = \mathrm{E}\{(n_{ij}/n_{iy})\mathrm{cov}(x_i x_j)\beta + \mathrm{cov}(x_i \epsilon)\} \tag{32}$$

and hence at

$$\mathrm{E}(\hat{\beta}) = (\mathrm{cov}(x_i x_j))^{-1}\{(n_{ij}/n_{iy})\mathrm{cov}(x_i x_j)\}\beta \qquad (i, j = 1, \ldots, p). \tag{33}$$

n_{ij} and n_{iy} are the numbers of measurements common to both x_i and x_j, and x_i and y, respectively, minus unity. So β is unbiased only when all n_{iy}'s are equal. Similarly, $\mathrm{V}(\hat{\beta})$ corresponds to the CLSE form when $n_{ij} = n_{iy} = n_{jy} (i, j = 1, \ldots, n)$. In a Monte–Carlo experiment for various patterns of missing observations [15] came to the conclusion that in most cases the complete case estimator is superior to the available case estimator.

3.2.3 Imputation by Zero–Order–Regression (ZOR)

By this method ([42]), which is also called unconditional mean imputation, missing values x_{ij} of the jth regressor X_j are replaced by the sample (column) mean $\hat{x}_{ij}$ of the observed values of X_j. This method is expected to be convenient if the span or the range of the X_j–realizations is moderate. It may fail in cases where the sample mean is not a satisfactory representative of the missing sample elements of X_j. This happens for example if trending time series or growth curves are the laws generating the X_j–values.

The replacement of the missing x_{ij}–values in X_* by $\hat{x}_{ij}$ transforms the (partially or fully unknown) matrix X_* into a known matrix $X_{(1)}$. Thus we are led to the operational model of mixed regression type

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_{(1)} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_{(1)} \end{pmatrix}, \tag{34}$$

where the error term

$$\epsilon_{(1)} = (X_* - X_{(1)})\beta + \epsilon_* \tag{35}$$

has

$$\epsilon_{(1)} \sim \{(X_* - X_{(1)})\beta, \sigma^2 I_J\}. \tag{36}$$

[17] describe a version of this method, the so–called modified zero–order–regression.

3.2.4 Imputation by First–Order–Regression (FOR)

By this notion there is understood a complexity of methods to estimate missing elements of X_*. In principle one constructs an auxiliary regression

$$x_{ij} = \theta_{0j} + \sum_{\substack{\mu=1 \\ \mu \neq j}}^{p} x_{i\mu}\theta_{\mu j} + u_{ij}\,, \qquad i \notin \Phi = \bigcup_{j=1}^{p} \Phi_j\,, \tag{37}$$

(u_{ij} : error term), Φ_j being the index set of missing values in x_j, to estimate the dependence between X_j ($j = 1, \ldots, p$, j fixed) and the

other regressors $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$. The missing value x_{ij} then is estimated by

$$\hat{x}_{ij} = \theta_{0j} + \sum_{\substack{\mu=1 \\ \mu \neq j}}^{p} x_{i\mu} \hat{\theta}_{\mu j} \ , \qquad (i \in \Phi_j). \tag{38}$$

(For examples compare [35].)

To overcome the difficulty caused by overlapping index sets, there are proposed certain methods depending on the pattern of missing values and on perceptible laws in the design matrix. [1] mentioned some averaging procedures (see also [3]).

Another possibility is to use the auxiliary regression with the highest measure of determination whereas remaining missing values are replaced by other estimates, e.g., the corresponding sample means.

[9] proposed a generalized LSE procedure, where the matrix X_c is completed by first–order regression approximations.

[38] investigated some different procedures based on the FOR within a linear regression model with an incomplete binary covariate. The so–called pi imputation—simply imputing the probabilities based on the estimates of the logistic regression model—in terms of the empirical mean squared error showed good results.

An extension to the FOR in the context of generalized additive models could lead to an auxiliary regression which is also of the GAM type.

3.2.5 Imputation by Modified First–Order–Regression (MFOR)

The first order regression doesn't use the response y for imputing the missing data which is the idea of the modified first order regression (MFOR). Within the auxiliary regression model additionally the completely observed response vector is used to predict the missing values. [3] considered this situation within the context of estimating μ and Σ in the normal model of $(y, X_1, \ldots, X_p)$. [36] did some work on the asymptotic properties of the MFOR estimates.

Regression parameters are biased when imputing for missing data using the MFOR. The early work of [1] provides some bias adjustment for the univariate case. The reason for the biased estimates lies in the inversion of the regression direction.

In [38] also a MFOR was compared to some alternatives concerning an incomplete binary variable; the estimates showed the expected larger variances resulting from the use of y.

3.2.6 Multiple Imputation

In this paragraph we will describe the main ideas of the multiple imputation. For details see for example [26] or [25].

So far, we imputed values that are a 'good' fitting to the data, taking into account either only the variable where missingness occurs, or additionally other variables. A problem that will occur is that the empirical variances of the variables will reduce and in the case of first order or modified first order regression association will be higher that they actually are.

Instead of imputing a 'best' value, draw values for the missing values, from conditional distributions that consider the uncertainty due to the prediction. The uncertainty arises from a distribution assumption and from the fact that parameter estimates themselves have variances. In contrast to the former methods, we will not only produce one complete dataset, but draw several ones, that vary in the imputed values. The variation over the datasets may now reflect the uncertainty due to imputation.

First, distribution assumptions for the covariates where missing values may occur have to be made. Together with the error distribution and the model equation we obtain a common distribution

$$P_{\Theta}(y, X_{\text{mis}} \mid X_{\text{obs}}), \tag{39}$$

where X_{mis} denotes all covariates where missingness may occur, X_{obs} denotes completely observed covariates that depend on the parameters Θ, including the variance of the errors ϵ.

The method we will use to do multiple imputations is called data augmentation. As data augmentation is a Bayesian procedure one has to choose an appropriate prior distribution for the parameters Θ. For choices see e.g. [2].

Altogether we obtain

$$P(y, X_{\text{mis}}, \Theta \mid X_{\text{obs}}) = P_{\Theta}(y, X_{\text{mis}} \mid X_{\text{obs}})P(\Theta). \tag{40}$$

Dependent on the distribution assumptions, we are able to draw data from the conditional distributions either directly or using sampling methods like MCMC.

Having chosen the first imputations for the missing values data augmentation consists of two steps that are applied consecutively many times until we can assume that the joint distributuion of the missing values and of the parameters Θ converge.

The steps are:

1. Imputation Step: For every row of the data set where missing values occur, draw from

$$P(X_{\text{mis}} \mid y, X_{\text{obs}}, \Theta), \tag{41}$$

   and impute the drawn values as new values. X_{mis} denotes the covariates where missing values occur and X_{obs} denotes the observed covariates.

2. Propability Step: Using the completed data set draw from

$$P(\Theta \mid y, X_1, ..., X_k) \tag{42}$$

   and take the drawn values as new parameter values.

Applying this procedure leads to several completed data sets. Assume that we have drawn M data sets we now obtain the following estimators using estimates of the single data sets.

Let $\widehat{q_t}$, $t = 1, ..., M$ be point estimates for the M completed datasets, and $\widehat{U_t}$ be the variance estimate of the estimator $\widehat{q_t}$. As new estimates we obtain:

$$\widehat{q} = \frac{1}{M} \sum_{t=1}^{M} \widehat{q_t}, \tag{43}$$

see [26].

$$\widehat{\mathrm{V}\widehat{q}} = \frac{1}{M}\sum_{t=1}^{M}\widehat{U_t} + \sum_{t=1}^{M}(\widehat{q_t} - \widehat{q})(\widehat{q_t} - \widehat{q})' \tag{44}$$

The formulas above can be applied to scalar as well to multivariate estimators and will in our context in general be the estimates of our parameter vector β.

3.2.7 Nearest Neighbor Imputation

The nearest neighbor imputation has a long history but according to [6] is still not fully investigated although it is used in many surveys. Assuming the data structure with J missing values for the row indices $i = n - J + 1, \ldots, n$ visualized by

$$\underbrace{x_1, \ldots, x_{n-J}}_{\text{observed}}, \underbrace{x_{n-J+1}, \ldots, x_n}_{\text{missing}} \quad \text{and} \tag{45}$$

$$\underbrace{y_1, \ldots, y_{n-J}, y_{n-J+1}, \ldots, y_n}_{\text{observed}} \quad , \tag{46}$$

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

$$\mid y_i - y_j \mid \;\; = \;\; \min_{1 \leq l \leq n-J} \mid y_l - y_j \mid \; . \tag{47}$$

If the solution is not unique the mean of the corresponding x–values may be imputed.

The nearest neighbor imputation is a hot deck imputation procedure which yields values unlikely to be nonsensical. Population means and quantiles are asymptotically unbiased and consistent (see [5]). Since it is a nonparamteric method it is expected to be somewhat more robust against model violations. [6] give a detailed overview over several possibilities for adjusting the procedure in order to get asymptotically unbiased and consistent variance estimates.

[22] investigated a simple additive model with missing completely

at random in the covariate and came to the result that the nearest neighbor imputation showed results similar to the complete case analysis—a procedure with best asymptotic properties when the missingness is independent of y. A forthcoming work considers the situation when the missingness depends on y; first results showed that the CCA became worse and the nearest neighbor imputation still shows good results.

3.2.8 ML Estimation of the Missing Values

Let us now assume that the disturbances are normally distributed,

$$\epsilon_c \sim \mathrm{N}(0, \sigma^2 I_m),\ \epsilon_* \sim \mathrm{N}(0, \sigma^2 I_J). \tag{48}$$

Handling the nonobseved regressor values like unknown parameters which have to be estimated common with β and σ^2 leads to the following considerations. For reasons of simpler mathematical presentation we confine ourselves to models without a constant and to the case of a fully nonobserved regressor matrix X_* which has to be estimated from the model

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_* \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix}, \qquad \begin{pmatrix} \epsilon_c \\ \epsilon_* \end{pmatrix} \sim \mathrm{N}(0, \sigma^2 I_n) \tag{49}$$

The logarithm of the likelihood is

$$\begin{aligned} \ln \mathrm{L}(\beta, \sigma^2, X_*) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}(\sigma^2) \\ &- \frac{1}{2\sigma^2}(y_c - X_c\beta, y_* - X_*\beta)' \begin{pmatrix} y_c - X_c\beta \\ y_* - X_*\beta \end{pmatrix}. \end{aligned} \tag{50}$$

Differentiating (50) with respect to $\beta, \sigma^2,$ andX_* and equating to zero results in the normal equations and their solutions

$$\hat{\beta} = b_c = S_c^{-1} X_c' y_c\,, \tag{51}$$

$$\hat{\sigma}^2 = \frac{1}{n}(y_c - X_c b_c)'(y_c - X_c b_c) \tag{52}$$

which are based on the complete observations, only. The maximum likelihood estimator (MLE) $\hat{X}_*$ is solution of the relation

$$y_* = \hat{X}_* b_c \tag{53}$$

which is uniquely determined in the case of $p = 1$, only. In general we have a $(J \times (p-1))$–dimensional manifold of admissible solutions $\hat{X}_*$. To find a unique solution one may pose an additional criterion which chooses an $\hat{X}_*$ such that it fulfills relation (53) and that it is optimal with respect to the specified criterion.

In the missing value regression the mixed estimation framework may be understood as a two–step procedure: first, replace X_* by some $\hat{X}_*$ and second, estimate β by

$$\hat{\beta}(\hat{X}_*) = (S_c + \hat{X}_*'\hat{X}_*)^{-1}(X_c'y_c + \hat{X}_*'x_*). \tag{54}$$

Choosing $\hat{X}_*$ according to the ML–normal equation (53) gives the result

$$\begin{aligned} \hat{\beta}(\hat{X}_*) &= (S_c + \hat{X}_*'\hat{X}_*)^{-1}(S_c\beta + X_c'\epsilon_c + \hat{X}_*'\hat{X}_*\beta + \hat{X}_*'\hat{X}_*S_c^{-1}X_c'\epsilon_c) \\ &= \beta + (S_c + \hat{X}_*'\hat{X}_*)^{-1}(S_c + \hat{X}_*'\hat{X}_*)S_c^{-1}X_c'\epsilon_c \\ &= \beta + S_c^{-1}X_c'\epsilon_c \\ &= b_c. \end{aligned} \tag{55}$$

That is, whatever the solution $\hat{X}_*$ of (53), the corresponding mixed estimator $\hat{\beta}(\hat{X}_*)$ coincides with the CLSE b_c.

Note. The algorithms of [23] and [10] for solving ML–equations may be used for patterns of missing values which are different from a fully unknown matrix $\hat{X}_*$.

For a further discussion of MLE in missing values regression see [41] and [37].

3.3 The Mixed Regression Framework

3.3.1 Imputation and Biased Mixed Estimation

Let us go back to the completely observed model (24) and to the model (25) with the incomplete X_*–matrix. Model (25) may be interpreted as J additional observations on the independent variable y but some of the independent variables are missing.

Certain methods of this section are such that missing observations

in X_* are replaced by approximations transforming X_* into a known matrix, say X_R. Substituting X_* in (25) by the nonstochastic $(J\times p)$–matrix X_R leads to

$$y_* = X_R\beta + (X_* - X_R)\beta + \epsilon_* = X_R\beta + v_*\,, \qquad \text{say,} \tag{56}$$

where the disturbance term v_* has

$$v_* = (X_* - X_R)\beta + \epsilon_* \sim (\delta, \sigma^2 I_J) \tag{57}$$

with

$$\delta = (X_* - X_R)\beta\,. \tag{58}$$

Combining the completely observed sample (24) with the additional sample and by the substitution of X_* by X_R now operational information leads to the mixed model (see [17])

$$\begin{pmatrix} y_c \\ y_* \end{pmatrix} = \begin{pmatrix} X_c \\ X_R \end{pmatrix}\beta + \begin{pmatrix} \epsilon_c \\ v_* \end{pmatrix} \tag{59}$$

with

$$\begin{pmatrix} \epsilon_c \\ v_* \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ \delta \end{pmatrix}, \sigma^2 I_n \right). \tag{60}$$

In the mixed regression framework due to [31] the relation (56) may be interpreted as J additional linear stochastic restrictions $r = R\beta + v_*$. The mixed estimator due to Theil was developed for the case $\delta = 0$. Investigations on biased stochastic restrictions on β are given in [30]. [34] came to this problem in considering misspecified linear restrictions.

The mixed estimator of β in the model (59) is just the OLSE, i.e.,

$$b_R = (S_c + S_R)^{-1}(X_c' y_c + X_R' y_*)\,, \tag{61}$$

where $S_c = X_c' X_c$ and $S_R = X_R' X_R$. This estimator is biased

$$\operatorname{bias} b_R = (S_c + S_R)^{-1} X_R' \delta \tag{62}$$

and has the covariance matrix

$$\mathrm{V}(b_R) = \sigma^2 (S_c + S_R)^{-1}\,. \tag{63}$$

A variance comparison with the unbiased CLSE b_c, which discards the additional information of (56), gives

$$\mathrm{V}(b_c) - \mathrm{V}(b_R) = \sigma^2 S_c^{-1} X_R' (X_R S_c^{-1} X_R' + I)^{-1} X_R S_c^{-1} \tag{64}$$

which is nonnegative definite. Thus replacing missing values of X_* by some chosen imputation method results in a biased estimator b_R having smaller variance in the sense of (64) compared with the unbiased OLSE b_c. Hence a mean–squared–error–criterion appears to be a good device to weight the disadvantage of bias and the advantage of smaller variance.

3.3.2 MSE–Criteria

The mean squared error (MSE) of an estimator $\hat{\beta}$ is defined as

$$\mathrm{MSE}(\hat{\beta}, \beta) = \mathrm{V}(\hat{\beta}) + (\mathrm{bias}\,\hat{\beta}, \beta)(\mathrm{bias}\,\hat{\beta}, \beta)' \tag{65}$$

To compare the two estimators b_R and b_c with respect to their MSE or related functions of MSE we may apply the following criteria ([24]) to our problem.

3.3.3 MSE I–Criterion (Strong MSE–Superiority)

b_R is said to be MSE I–better than b_c if

$$\Delta(b_c, b_R) = \mathrm{MSE}(b_c, \beta) - \mathrm{MSE}(b_R, \beta) \qquad \text{n.n.d.} \tag{66}$$

Now we have

$$\mathrm{MSE}(b_c, \beta) = \sigma^2 S_c^{-1} \tag{67}$$

and, using (62) and (63),

$$\mathrm{MSE}(b_R, \beta) = \sigma^2 (S_c + S_R)^{-1} + (S_c + S_R)^{-1} X_R' \delta\delta' X_R (S_c + S_R)^{-1}. \tag{68}$$

By standard inversion formulae we get

$$(S_c + S_R)^{-1} = S_c^{-1} - S_c^{-1} X_R' (X_R S_c^{-1} X_R' + I)^{-1} X_R S_c^{-1} \tag{69}$$

and, therefore, it holds that

$$(S_c + S_R)^{-1} X_R' = S_c^{-1} X_R' (X_R S_c^{-1} X_R' + I)^{-1} = D\,. \tag{70}$$

As $X_R S_c^{-1} X_R' + I$ is p.d., we have the presentation

$$X_R S_c^{-1} X_R' + I = C'C\,, \tag{71}$$

where C is a regular matrix. Then (66) becomes

$$\begin{aligned}\Delta(b_c, b_R) &= \sigma^2 D\left[C'C - \sigma^{-2}\delta\delta'\right]D' \\ &= \sigma^2 DC'\left[I - \sigma^{-2}C'^{-1}\delta\delta'C^{-1}\right]CD' \qquad (72)\end{aligned}$$

which is n.n.d if and only if

$$I - \sigma^{-2}C'^{-1}\delta\delta'C^{-1} \qquad \text{n.n.d.} \qquad (73)$$

This holds if

$$\kappa = \sigma^{-2}\delta'C^{-1}C'^{-1}\delta = \sigma^{-2}\delta'(X_R S_c^{-1} X_R' + I)^{-1}\delta \leq 1\,, \qquad (74)$$

where κ is the non–centrality parameter of the test statistic

$$F = \frac{n-k}{J \cdot s^2}(y_* - X_R b_c)'\left[X_R S_c^{-1} X_R' + I\right]^{-1}(y_* - X_R b_c) \qquad (75)$$

with $s^2 = (y_* - X_R b_c)'(y_* - X_R b_c)$. F has an $F_{J,n-K}(\kappa)$–distribution under the null hypothesis $\mathrm{H}_0 : \kappa \leq 1$. The test statistic F can be used to provide a uniformly most powerful test which tests whether the restricted estimator b_R is MSE I–better than b_c ($\mathrm{H}_0 : \kappa \leq 1$) or not ($\mathrm{H}_1 : \kappa > 1$). Tabulation of $F_{J,n-K}(\kappa)$ for $\kappa = 1$ is given in [33].
Note. From a pre–testing standpoint one could use the PT–estimator

$$\hat{\beta} = \begin{cases} b_R & \text{if } \mathrm{H}_0 : \kappa \leq 1 \text{ is accepted,} \\ b_c & \text{otherwise} \end{cases}$$

(see [17]).

3.3.4 MSE II–Criterion (First Weak MSE–Criterion)

b_R is said to be MSE II–better than b_c if

$$\mathrm{tr}\Delta(b_c, b_R) \geq 0\,. \qquad (76)$$

Then (see [24])

$$\kappa \leq \kappa_{\min} S_c \cdot \mathrm{tr} S_c^{-1} X_R'(X_R S_c^{-1} X_R' + I)^{-1} X_R S_c^{-1} = \kappa_0\,, \qquad (77)$$

would be a sufficient condition. Moreover, testing the MSE II–superiority of b_R over b_c may be realized using the F–statistic (75), whereas $\mathrm{H}_0 : \kappa \leq \kappa_0$ [(77)] against $\mathrm{H}_1 : \kappa > \kappa_0$ is tested.

3.3.5 MSE III–Criterion (Second Weak MSE–Criterion)

Another weaker scalar MSE–criterion is derived by changing the parameter space. If one is interested in estimating $X_c\beta$, the conditional mean of y_c given X_c, instead of estimating β itself, then b_R is said to be MSE III–better than b_c if and only if

$$\mathrm{E}(X_c b_R - X_c\beta)^{'}(X_c b_R - X_c\beta) \leq \mathrm{E}(X_c b_c - X_c\beta)^{'}(X_c b_c - X_c\beta)\,, \quad (78)$$

i.e. if (see (72))

$$\mathrm{tr}S_c\Delta(b_c, b_R) = \sigma^2 \mathrm{tr}X_R S_c^{-1} X_R^{'} - \delta^{'} X_R S_c^{-1} X_R^{'}\delta \geq 0. \quad (79)$$

By using

$$\delta^{'}(X_R S_c^{-1} X_R^{'})\delta \leq \delta^{'}(X_R S_c^{-1} X_R^{'} + I)\delta = \sigma^{-2}\kappa \quad (80)$$

with κ [(74)] the noncentrality parameter of F (75), a sufficient condition for (79) to hold is

$$\kappa \leq \mathrm{tr}X_R S_c^{-1} X_R^{'}. \quad (81)$$

3.4 The Weighted Mixed Regression Framework

3.4.1 The Weighted Mixed Regression Estimator (WMRE)

The mixed estimator b_R (64) in the model (62) is the solution to the minimization problem

$$\min_\beta\{(y_c - X_c\beta)^{'}(y_c - X_c\beta) + (y_* - X_R\beta)^{'}(y_* - X_R\beta)\}. \quad (82)$$

To give the observed 'sample' matrix X_c a different weight than the nonobserved matrix X_R in estimating β, [27] suggested to solve

$$\min_\beta\{(y_c - X_c\beta)^{'}(y_c - X_c\beta) + \lambda(y_* - X_R\beta)^{'}(y_* - X_R\beta)\}\,, \quad (83)$$

where λ is a scalar factor. Differentiating (83) with respect to β and equating to zero gives the normal equation

$$(S_c + \lambda S_R)\beta - (X_c^{'} y_c + \lambda X_R^{'} y_*) = 0. \quad (84)$$

The solution may be called the weighted mixed regression estimator (WMRE) and is of the form

$$b(\lambda) = (S_c + \lambda S_R)^{-1}(X_c^{'} y_c + \lambda X_R^{'} y_*). \tag{85}$$

This estimator may be understood as the familiar mixed estimator in the model

$$\begin{pmatrix} y_c \\ \sqrt{\lambda} y_* \end{pmatrix} = \begin{pmatrix} X_c \\ \sqrt{\lambda} X_R \end{pmatrix} \beta + \begin{pmatrix} \epsilon_c \\ \sqrt{\lambda} v_* \end{pmatrix}. \tag{86}$$

Let

$$Z(\lambda) = Z = (S_c + \lambda S_R). \tag{87}$$

Then we have

$$\begin{aligned} b(\lambda) &= Z^{-1}(X_c^{'} X_c \beta + X_c^{'} \epsilon_c + \lambda X_R^{'} X_* \beta + \lambda X_R^{'} \epsilon_*) \\ &= \beta + \lambda Z^{-1} X_R^{'}(X_* - X_R)\beta + Z^{-1}(X_c^{'} \epsilon_c + \lambda X_R^{'} \epsilon_*). \end{aligned} \tag{88}$$

Let again

$$\delta = (X_* - X_R)\beta. \tag{89}$$

The WMRE is biased

$$\operatorname{bias} b(\lambda) = \lambda Z^{-1} X_R^{'} \delta \tag{90}$$

and has the covariance matrix

$$\mathrm{V}(b(\lambda)) = \sigma^2 Z^{-1}(S_c + \lambda^2 S_R) Z^{-1}. \tag{91}$$

3.4.2 Minimizing the MSEP

A reliable criterion to choose λ is to minimize the mean squared error of prediction (MSEP) with respect to λ.
Let

$$\tilde{y} = \tilde{x}^{'} \beta + \tilde{\epsilon}\,, \qquad \tilde{\epsilon} \sim (0, \sigma^2), \tag{92}$$

a nonobserved (future) realization of the regression model which is to be predicted by

$$p = \tilde{x}^{'} b(\lambda). \tag{93}$$

The MSEP of p is

$$\begin{aligned} \mathrm{E}(p - \tilde{y})^2 &= \mathrm{E}[\tilde{x}^{'}(b(\lambda) - \beta) - \tilde{\epsilon}]^2 \\ &= [\tilde{x}^{'} \operatorname{bias} b(\lambda)]^2 + \tilde{x}^{'} \mathrm{V}(b(\lambda)) \tilde{x}^{'} + \sigma^2. \end{aligned} \tag{94}$$

Minimizing with respect to λ gives the solution

$$\lambda = \frac{1}{1 + \sigma^{-2}\rho_1(\lambda)\rho_2^{-1}(\lambda)}, \qquad 0 \leq \lambda \leq 1, \tag{95}$$

where

$$\rho_1(\lambda) = \tilde{x}^{'} Z^{-1} S_c Z^{-1} X_R{}' \delta\delta^{'} X_R Z^{-1} \tilde{x}, \tag{96}$$

$$\rho_2(\lambda) = \tilde{x}^{'} Z^{-1} S_R Z^{-1} S_c Z^{-1} \tilde{x}. \tag{97}$$

Thus the optimal λ minimizing the MSEP (94) of $p = \tilde{x}^{'} b(\lambda)$ is solution of the relation (95). Noting that $Z = Z(\lambda)$ is a function of λ, also, solving ((95) for λ results in a procedure iterating the λ–values whereas σ^2 and δ are estimated by some procedure. The problem becomes somewhat simpler in the case that only one row of the regressor matrix is incompletely observed,

$$\begin{array}{ccccccc} y_* & = & x_* & \beta & + & \epsilon_* \\ (1,1) & & (1,p) & (p,1) & & (1,1) \end{array}, \qquad \epsilon_* \sim (0, \sigma^2). \tag{98}$$

Then we have $S_R = x_R x_R^{'}, \delta = (x_*^{'} - x_R^{'})\beta$ (a scalar) and

$$\rho_1(\lambda) = (\tilde{x}^{'} Z^{-1} S_c Z^{-1} x_R)(x_R^{'} Z^{-1} \tilde{x})\delta^2, \tag{99}$$

$$\rho_2(\lambda) = (\tilde{x}^{'} Z^{-1} x_R)(x_R^{'} Z^{-1} S_c Z^{-1} \tilde{x}). \tag{100}$$

So λ becomes

$$\lambda = \frac{1}{1 + \sigma^{-2}\delta^2} \tag{101}$$

Interpretation of the result:

(i) We note that $0 \leq \lambda \leq 1$, so that λ indeed is a weight given to the incompletely observed model.

(ii) $\lambda = 1$ holds for $\sigma^{-2}\delta^2 = 0$. If σ^2 is finite, then the incompletely observed but (by the replacement of x_* by x_R) 'repaired' model is given the same weight as the completely observed model in case of $\delta = 0$, only. Now, $\delta = (x_*^{'} - x_R^{'})\beta = 0$ means that the unknown expectation $\mathrm{E}y_* = x_*^{'}\beta$ of the dependent variable y_* is estimated exactly by $x_R^{'}\beta$ (for all β). Thus $\delta = 0$ is fulfilled when $x_* = x_R$, i.e. when missing values in x_* are re–estimated exactly (without error) by x_R.

This seems to be an interesting result to be taken in mind

in general mixed regression framework in the sense, that additional linear stochatic restrictions of type $r = R\beta + v_*$ should not be incorporated without posing on them a prior weight λ (and $\lambda < 1$ in general).

Furthermore, it may be conjectured that weighted mixed regression becomes equivalent (in a sense to be specified) to the familiar (unweighted) mixed regression, when the former is related to a strong MSE–criterion and the latter is related to a weaker MSE–criterion.

Now, $\lambda = 1$ may be caused by $\sigma^2 \to \infty$, also. As σ^2 is the variance common both to y_c and $y_*, \sigma^2 \to \infty$ leads to unreliable (imprecise) estimators in the complete model $y_c = X_c^{'}\beta + \epsilon$ as well as in the enlarged mixed model (59).

(iii) In general, an increasing δ decreases the weight λ of the additional stochastic relation $y_* = x_R^{'}\beta + v_*$. If $\delta \to \infty$, $\lambda \to 0$ and

$$\lim_{\lambda \to 0} b(\lambda) = b_c. \tag{102}$$

3.4.3 The Two–Stage WMRE

To bring the mixed estimator $b(\lambda)$ with λ from (101) in an operational form,σ^2 and δ have to be estimated by $\hat{\sigma}^2$ and $\hat{\delta}$ resulting in $\hat{\lambda} = 1/(1 + \hat{\sigma}^{-2}\hat{\delta}^2)$ and $b(\hat{\lambda})$.
Using the consistent estimators

$$\hat{\sigma}^2 = \frac{1}{m-p}(y_c - X_c b_c)^{'}(y_c - X_c b_c) \tag{103}$$

and

$$\hat{\delta} = y_* - x_R^{'} b_c, \tag{104}$$

what are then the properties of the resulting two–stage WMRE $b(\hat{\lambda})$. This will depend on the statistical properties (e.g. mean and variance) of $\hat{\lambda}$ itself. The bootstrap method ([11]) is one of the nonparametric methods in estimating variance and bias of a statistic of interest.

4 Regression Diagnostics to Identify Non–MCAR Processes

The different methods to deal with the design matrix with missing observations depend on the nature of missing data mechanism. The assumption that missing values are independent of the observed as well as unobserved data is more restrictive than the MAR process in which the missing values depend on the observed data.

Different diagnostic tests are available in the literature to identify the non–MCAR processes. One simple approach given by [8] is based on the sample means of observed and unobserved data on response variables. If the partitionary of y in y_c and y_* , based on missing values n_* in x_*, is random then it indicates that the process is MCAR. Another way to test the MCAR assumption is to compare the variance covariance matrices of estimates of β with complete and repaired data sets.

The "leave–one–out" strategy from sensitivity analysis ([4]) allows to detect the influential missingness of any particular observation. This strategy computes some scalar statistic based on complete data set or after eliminating any particular observation from the data set.

Let $\hat{\beta}_R$ be an estimator of β in the linear regression model

$$y = \begin{pmatrix} X_c \\ X_R \end{pmatrix} + \epsilon \tag{105}$$

where X_R is the matrix obtained after repairing x_* through a chosen imputing technique. Several diagnostic measures have been proposed based on this model. For example, using Cook's distance, one can compute

$$D = \frac{(\hat{\beta}_R - \hat{\beta}_c)' X' X (\hat{\beta}_R - \hat{\beta}_c)}{p s_c^2} \geq 0 \tag{106}$$

where s_c^2 is computed from the complete dataset. Another measure is based on residual sum of squares and requires to compute

$$DRSS = \frac{\frac{(RSS_R - RSS_c)}{J}}{\frac{RSS_c}{(n-m-p+1)}} \in (0, \infty). \tag{107}$$

Large values of DRSS are indicative of departure from MCAR process.

Based on kernel of Andrews–Pregibon statistics, the small values of the determinant

$$DXX = \frac{\mid X_c' X_c \mid}{\mid X'X \mid} \in [0,1] \tag{108}$$

indicate the violation of MCAR assumption.

The distributions of D, DRSS and DXX are required to test H_0: MCAR vs. H_1: non–MCAR, but they depend on x, β and s_c^2. One may obtain the solution through Monte–Carlo simulation with following steps:

- Fill missing values in x_* by suitable MCAR substitute.
- Using $\hat{\beta}_R$, s_c^2 and x_c, update y by calculating

$$y_*^s = x_f \hat{\beta}_c + \epsilon^s \text{ with } \epsilon \sim N(0, s_c^2 I). \tag{109}$$

  Here "s" stands for simulated values.

- Calculate the diagnostic measure based on this data set.
- Repeat the process N times with an updated ϵ^s in each step and estimate the null distribution to of the required diagnostic measure.

With thus obtained null distribution, the critical values obtained are the $N(1-\alpha)^{th}$ order statistics for D and DRSS and $N\alpha^{th}$ order statistics for DXX respectively. The decision rule is reject H_0 if D (or DRSS) $\geq f_{0,N(1-\alpha)}$ or if $DXX \leq f_{0,N_\alpha}$ respectively. For more details see for example [13], [28] or [29].

References

[1] Afifi, A. A. and Elashoff, R. M. (1996), Missing observations in multivariate statistics. Part I, review of the literature, Journal of the American Statistical Association, **61**, 595–604.

[2] Box, G. E. P. and Tiao, G. (1992, Bayesian Inference in Statistical Analysis. New York: Wiley.

[3] 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**, 302–307.

[4] Chatterjee, S. and Hadi, A. S. (1988), Sensitivity Analysis in Linear Regression. New York: Wiley.

[5] Chen, J. and Shao, J. (2000), Biases and variances of survey estimators based on nearest neighbor imputation. Tech. Rep., University of Wisconsin-Madison.

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

[7] Chu, C. K. and Cheng, P. E. (1995), Nonparametric regression estimation with missing data. Journal of Statistical Planning and Inference, **48**, 85–99.

[8] Cohen, J. and Cohen, P. (1983), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum, Hillsdale, NJ.

[9] Dagenais, M. G. (1973), The use of incomplete observations in multiple regression analysis: A generalized least squares approach. Journal of Econometrics, **1**, 317–328.

[10] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Ser. B, **43**, 1–22.

[11] Efron, B. (1979), Bootstrap methods. Another look at the jackknife, Annals of Statistics, **7**, 1–26.

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

[13] Fieger, A. (2000), Fehlende Kovariablenwerte bei Linearen Regressionsmodellen. Dissertation, Ludwig-Maximilians-Universität München.

[14] Gu, C. and Wahba, G. (1991), Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. SIAM Journal on Scientific and Statistical Computing, **12**, 383–398.

[15] Haitovsky, Y. (1968), Missing data in regression analysis. Journal of the Royal Statistical Society, Ser. B, **34**, 67–82.

[16] Hastie, T. and Tibshirani, R. J. (1990), Generalized Additive Models. London: Chapman and Hall.

[17] Hill, R. C. and Ziemer, R. F. (1983), Missing regressor values under conditions of multicollinearity. Communications in Statistics, Part A—Theory and Methods, **12**, 2557–2573.

[18] Ibrahim, J. G., Lipsitz, S. R. and Chen, M. H. (1999), Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. Journal of the Royal Statistical Society, Ser. B, **61**(1), 173–190.

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

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

[21] Nittner, T. (1999), Fehlende Daten im klassischen linearen Regressionsmodell – Eine Erweiterung existenter Verfahren auf diskrete Kovariablen. Diplomarbeit, Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80535 München, Germany.

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

[23] Oberhofer, W. and Kmenta, J. (1974), A general procedure for obtaining maximum likelihood estimates in generalized regression models. Econometrica, **42**, 579–590.

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

[25] Rubin, D. B. (1996), Multiple imputation after 18+ years. Journal of the American Statistical Association, **91**(434), 473–489.

[26] Schafer, J. L. (1997), Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

[27] Schaffrin, B. and Toutenburg, H. (1990), Weighted mixed regression. Zeitschrift für Angewandte Mathematik und Mechanik, **70**, 735–738.

[28] Simon, G. A. and Simonoff, J. S. (1986), Diagnostic plots for missing data in least squares regression. Journal of the American Statistical Association, **81**, 501–509.

[29] Simonoff, J. S. (1988), Regression diagnostics to detect nonrandom missingness in linear regression. Technometrics, **30**, 205–214.

[30] Teräsvirta, T. and Toutenburg, H. (1980), A note on the limits of a modified Theil estimator. Biometrical Journal, **22**, 561–562.

[31] Theil, H. (1963), On the use of incomplete prior information in regression analysis. Journal of the American Statistical Association, **58**, 401–414.

[32] Theil, H. and Goldberger, A. S. (1961), On pure and mixed estimation in econometrics. International Economic Review, **2**, 65–78.

[33] Toro-Vizcarrondo, C. and Wallace, T. D. (1968), A test of the mean square error criterion for restrictions in linear regression. Journal of the American Statistical Association, **63**, 558–572.

[34] Toutenburg, H. (1970), Probleme linearer Vorhersagen im allgemeinen linearen Regressionsmodell. Biometrische Zeitschrift, **12**, 242–252.

[35] Toutenburg, H. (2002), Lineare Modelle. 2. Auflage, Physica, Heidelberg.

[36] Toutenburg, H., Fieger, A. and Srivastava, V. K. (1999), Weighted modified first order regression procedures for estimation in linear models with missing X-observations. Statistical Papers, **40**, 351–361.

[37] Toutenburg, H., Heumann, C., Fieger, A. and Park, S. H. (1995), Missing values in regression: Mixed and weighted mixed estimation. Eds. V. Mammitzsch and H. Schneeweiß, Statistical Sciences, Proceedings of the 2nd Gauss Symposium, De Gruyter, Munich 1983, 289–301.

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

[39] Tutz, G. (2000), Die Analyse kategorialer Daten. Oldenbourg, München.

[40] Vach, W. (1994), Logistic Regression with Missing Values and Covariates. Lecture Notes in Statistics, **86**, Berlin: Springer–Verlag.

[41] Weisberg, S. (1980), Applied Linear Regression. New York: Wiley.

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

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