The Classical Linear Regression Model with one Incomplete Binary Variable

H. Toutenburg T. Nittner

December 1, 1999

Abstract

We present three different methods based on the conditional mean imputation when binary explanatory variables are incomplete. Apart from the single imputation and multiple imputation especially the so-called pi imputation is presented as a new procedure. Seven procedures are compared in a simulation experiment when missing data are confined to one independent binary variable: complete case analysis, zero order regression, categorical zero order regression, pi imputation, single imputation, multiple imputation, modified first order regression. After a brief theoretical description of the simulation experiment, MSE-ratio, variance and bias are used to illustrate differences within and between the approaches.

KEY WORDS: binary variables; imputation; incomplete data; logistic regression; simulation experiment;

1 Introduction

Statistical analysis with incomplete data is a common problem in practice. The linear regression as a main tool therefore often is affected by missing values within statistical analyses. Apart from other popular approaches of handling this problem also imputation methods were discussed in particular (see e.g. Hill and Ziemer (1983) or Little (1992)). Within these essays the statistical theory was also described as the possibility of a practical transfer, e.g. in simulation experiments. Various scientific analyses use linear regressions containing binary variables (nominally or ordinally scaled) which also may be incompletely observed. This paper presents a brief description of methods dealing with incomplete binary variables. Based on the well known notation of a linear model

$$y = X\beta + \epsilon \tag{1.1}$$

we assume a continuous and completely observed response vector y. The $(n \times K)$ design matrix X however contains m missing values in one binary regressor. For remaining assumptions and notations see e. g. (Toutenburg, 1992, p. 18ff).

As already insinuated we assume *one* incomplete variable. Without any restriction of generality we also assume a reorganization of the incomplete regressor according to

$$X_{K-1} = (x_c, x_{mis})' \quad ,$$

where the indices c and mis indicate the complete and missing part of the variable X_{K-1}. Thereby x_c is of dimension $((n-m)\times 1)$ and x_{mis} of $(m\times 1)$.

2 Standard Methods

Complete and available case analysis among others are the most popular approaches. Whereas the complete case analysis (CCA) discards all cases containing at least one missing value the available case analysis uses all cases having complete observations for variables which are used within the scope of the current analysis. Therefore the number of cases for different problems often varies and depends on the current variables involved and might exceed the $n-m$ cases of the CCA. Both, the complete- and the available case analysis take advantage exclusively of the observed data and their information. Hence, the scaling of the variables doesn't affect the process of analyzing or its technique, respectively.

The unconditional mean imputation, also known as zero order regression (ZOR) and first introduced by Wilks (1932), is also a common approach to missing values. Each missing value of a regressor is replaced by the sample mean of the observed values of the regressor computed from the complete cases (see e. g. Rao and Toutenburg (1999, ch. 8)). Its usage however strictly regarded requires an adaption to the non-continuous scaling. Except for treating ordinally scaled variables as continuous the median has to be imputed for ordinally scaled data and the mode for nominally scaled data instead of the mean.

Because of rather secondary attention this short illustration should be sufficient, for more information about this topic see e.g. Rao and Toutenburg (1999, ch. 8).

3 Methods based on the conditional mean imputation

The conditional mean imputation is also known as first order regression (FOR) or Buck's method (see (Buck, 1960)). An auxiliary regression based on the complete cases of the incompletely observed variable incorporates the structure of the design matrix X. All complete independent variables are the independent part of the regression, the incompletely observed variable is the new response vector. Toutenburg, Srivastava and Fieger (1996) extended the FOR by including the completely observed response vector as an additional independent variable in the auxiliary regression. This method is also known as modified first order regression (mFOR). In general, the auxiliary regression is formulated according to

$$x_{ij} = \theta_{0j} + \sum_{\mu=1,\mu\neq j}^{K} x_{i\mu}\theta_{\mu j} + u_{ij} \quad , \quad i \notin \Phi \tag{3.1}$$

where x_{ij} is the missing value and Φ the index set of the missing values. Therefore x_{ij} can be replaced by $\hat{x}_{ij}$ with $i \in \Phi$ and

$$\hat{x}_{ij} = \hat{\theta}_{0j} + \sum_{\mu=1,\mu\neq j}^{K} x_{i\mu}\hat{\theta}_{\mu j} + u_{ij} \quad , \quad i \notin \Phi \quad . \tag{3.2}$$

The assumption of having a binary x_{ij} however prevents the modelling of the auxiliary regression (3.1) using the classical linear regression because of its requirement of a continuous response variable. Thus we have to turn to the theory of the generalized linear models (GLM), in particular the modelling of the logistic regression. As described in Fahrmeir, Hamerle and Tutz (1996, p. 248), its main tool can be written as

$$P(y = 1 \mid x_i) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \pi_i \tag{3.3}$$

where $\eta = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{K-1} x_{i,K-1}$. Based on the so-called logit-link (3.3), several imputation methods can be built.

3.1 Pi imputation

The idea of a probability-imputation is realized in the hence called **pi imputation**. As within the classical prediction (see e.g. Toutenburg (1992, p. 158)) the estimate of β - here based on the complete case model - is used to get substitutes for the missing values with help of the corresponding values of the complete variables. Therefore simply the probability is imputed for a 0/1-value. This corresponds to $P(y = 1 \mid x_l)$ in (3.3), the probability that the missing value is '1' given the values of the complete variables for case l and has to be computed for all cases $l = n - m + 1, \ldots, n$ where we have missing values. There's no restriction for the complete variables concerning their scaling.

Example Let us assume a simple regression model where the continuous response is the income of an individual. In addition to the incomplete binary variable 'gender' ('0' = male, '1' = female) the model contains further $K - 1$ independent variables of any quantity and scaling. The response vector of the auxiliary regression then is equal to the $n - m$ observed cases of the variable 'gender'. All completely observed independent variables (using also the complete response vector leads to the estimate of the mFOR) are used to run the logistic regression and to compute the estimate $\hat{\beta}$ for the unknown parameter β. The probability

$$P(x_{l,K-1} = 1 \mid \underbrace{x_{l,K-2} = r, x_{l,K-3} = r, \ldots, x_{l,1} = r, x_{l,0} = 1)}_{x_{l\cdot}} \quad ,$$

with $l = n-m+1, \ldots, n$, $r \in I\!R$ and the assumption of X_0 as a constant vector of ones and X_{K-1} as the incomplete binary variable, is computed with the help of (3.3) leading to

$$\pi_l = P(x_{l,K-1} = 1 \mid x_{l\cdot}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{l,1} + \ldots + \hat{\beta}_{K-2} x_{l,K-2})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{l,1} + \ldots + \hat{\beta}_{K-2} x_{l,K-2})} \quad . \tag{3.4}$$

The probability π_l in (3.4) has to be calculated for every case of X_{K-1} which has missing values. In practice, e.g. $\pi = 0.47$ then reflects a probability of 47% that the individual who did not state its gender in this survey was female given its remaining data based on the individuals who answered each question. Thus we get m probabilities π_l which commonly are different.

3.2 Single imputation

The conditional probability π_l in (3.4) enables two more imputation methods. In this paragraph we will consider the so-called **single imputation**. Based on these probabilities the conditional distribution of the incomplete variable given the complete variables has to be estimated line-by-line, for each interviewed individual in terms of our example, respectively. From these distributions a random draw for each missing value has to be made. The conditional distribution is simply a binomial distribution with probability π_l. It may also be easy to manage the random draw in practice, e.g. in a simulation experiment. For every individual who refused the answer we have to draw a pseudorandom number t_l out of a continuous uniform distribution within [0;1]. All missing values having $t_l \leq \pi_l$ are substituted by '1', i.e. these individuals are 'supposed to be' female. When for example $t_n = 0.61$ and $\pi_n = 0.33$, the n-th interviewed person (who did not state its gender) now is assumed to be male.

3.3 Multiple imputation

All approaches imputing each missing value once imply an important defect. Neither the uncertainty of the imputed value nor the possible incorrectness of the specified model are taken into account. The **multiple imputation** tries to cope with this essential structural nuisance. This is managed by M imputations instead of one for each missing value. Thus we have M completed data sets and also M (mostly) different estimates for the unknown parameter β. Rubin (1996, p. 480) has shown that multiple imputation already brings useful results for less than five imputations.

In order to get the final estimate of the multiple imputation the average of the M parameter vectors has to be computed. A little more effort is needed computing the variance of the multiple imputation which consists of two components (see Little and Rubin (1987, p. 257)). Apart from the variance within the estimation which emerges during a simulation experiment between its replications also the variance between the estimations has to be computed. The variance between the estimations results out of the M different parameter estimates and is expected to reflect the uncertainty about the in practice unknown value. Combining these two variances results in the variance of the multiple imputation. In some simulation experiments the variance of the multiple imputation was realized as higher as these from the alternative methods (see especially Rubin (1996)). Within this context it may be noted that the realization of this algorithm doesn't require much effort when the procedure of the single imputation already has been performed. It is only required to run the single imputation M-times and to compute averages and averaged variances of the M resulting estimates.

3.4 Extension to multi-categorical scaling

The paragraph about the unconditional mean imputation already included extensions to nominally and ordinally scaled variables with more than two categories by suggesting mode and median as suitable parameters. Methods based on the auxiliary regression within a first order regression such as pi imputation, single imputation and multiple imputation can't be extended in such a simple way to the case of multi-categorical scaled variables. Difficulties are coming within the FOR because of the application of the multi-categorical logit link for nominally scaled variables and the so-called cumulative model for ordinal structure (see e. g. Fahrmeir et al. (1996, p. 262ff.)). Because we focus on an incomplete binary variable this short clue should be sufficient and in the following the simulation experiment is reviewed.

4 A simulation experiment

4.1 Introduction

This simulation experiment was realized using C++ programming language and Sun Solaris, CDE version 1.2. The main advantage of this language is the reuse of existing classes and their documentations which were produced within this institute (Statistical Institute of the Ludwig-Maximilians-University, Munich). These classes simplified the programming enormously. For more information about contents and specific tools see Fieger, Heumann, Kastner and Watzka (1997).

Two different models were simulated and discussed. The first model consists of three covariables - one constant, two binary variables (one of them incomplete) - and was considered for three percentages of missing values (10%, 30%, 50%). Besides the three variables of the first model the second model contains another covariable which is continuous, however. The second model was considered for the two percentages 10% and 30%. In both models the data were missing completely at random (MCAR, see Rubin (1987)). Because of similar results, or to be more exactly, because of no new perceptions the review is restricted to the first model.

The following settings were chosen:

`sample size` $n = 30$
$l = 100$ `loops`
`multiple imputations` $M = 3$
X_1 `and` X_2 `according to B(30, 0.6)`
`missing percentages` $m_p \in \{0.1, 0.3, 0.5\}$
`correlation between` X_1 `and` X_2 `according to` $\rho \in \{0.0, 0.3, 0.6, 0.9\}$

For creating the data we first chose a fixed parameter vector β containing only ones. The design matrix X resulted from standard normally distributed pseudorandom numbers which were grouped in order to get binomial distributed numbers according to B(30, 0.6). The correlation structure was involved by the coefficient of Bravais-Pearson within the normal distribution (an existing

function described in Fieger (1997) was used to manage this). A comparison of these fixed coefficients with the coefficients for the nominally scaled data showed just slight differences. After the creation of the error vector ϵ the response vector y could be computed by $y = X\beta + \epsilon$. The $l = 100$ loops were 100 creations of ϵ (given X and β) and therefore of new response vectors. Based on this 'true' model we discarded some cases according to the missing percentage and the MCAR-assumption. So we had the complete case model and a model with missing values which enabled us to develop the imputation methods. All approaches were analyzed by existing functions within the classes for the linear regression model.

Different settings (e.g. missing percentage) didn't influence the time of about ten minutes which took an experiment to run.

Apart from the 'true' data set (TR) with the original values complete case analysis, zero order regression (Z), categorical zero order regression (imputation of the mode, cZ), pi imputation (PI), single imputation (SI), multiple imputation (MI) and a single imputation based on a modified first order regression (mF) were considered. Because of the nonexistence of passing trends the illustration is confined to unique contexts. Base for the illustration are the adjusted R^2, the MSE-ratio[1], variance and bias. The bias (difference between the fixed parameter and the estimate of the considered method) was the computed average of the l replications, the variance was the variance between the l different estimates for each method (except for the multiple imputation where we also had the second component).

4.2 The adjusted R squared

Except the steady maximum of the modified first order regression no specific properties could be noticed. However, the maximum of the mFOR mustn't esteemed as astonishing because of the additional usage of the response vector y for the auxiliary regression which leads to an increase of the variance.

4.3 The MSE-Ratio

The following diagrams illustrate the MSE-ratios of all considered imputation methods for each percentage of missing values and each correlation structure. The x-coordinate displays the correlation, the ordinate the MSE-ratio. As easily can be seen (and was also seen during the analysis of the second model) the pi imputation is the only procedure which always has a minor MSE than the common complete case analysis. Therefore the pi imputation is the only method which presents a uniform attribute irrespective to correlation and percentage of missing values.

[1] ratio between the mean square error of the complete case analysis and the mean square error of the alternative method

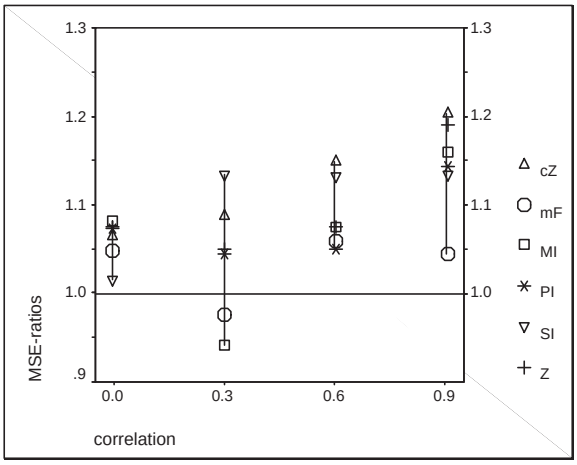

Figure 4.1: MSE-ratios of all imputation methods for a missing-percentages of 10%.

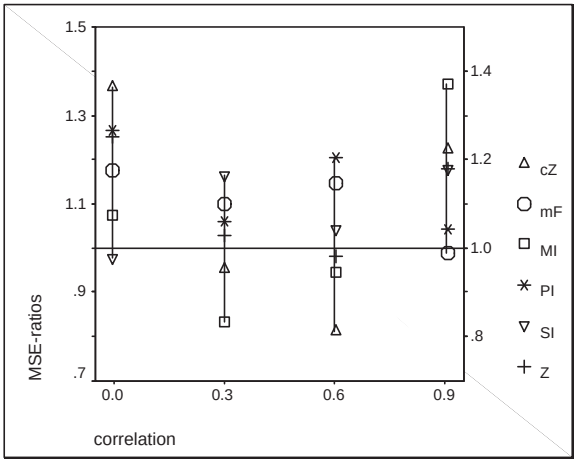

Figure 4.2: MSE-ratios of all imputation methods for a missing-percentage of 30%.

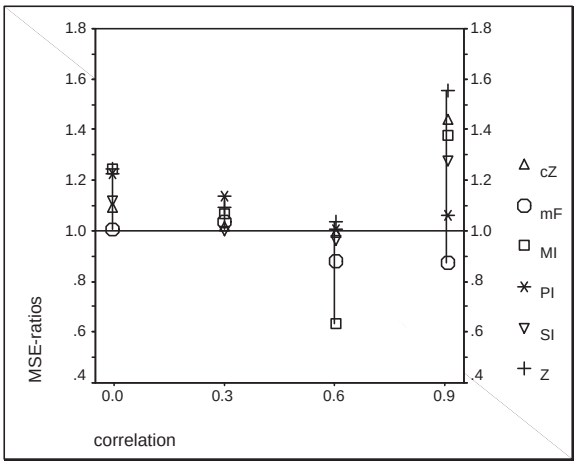

Figure 4.3: MSE-ratios of all imputation methods for a missing-percentage of 50%.

These results should be handled carefully, however. First of all, variance and bias have to be considered separately with respect to fluctuations which could explain apparent differences between the ratios. For example large variances of the complete case analysis could lead to MSE-superiorities of alternative methods which have smaller variances and relatively small biases. Secondly, model specific properties have to be taken into account when trying to discuss the results, e.g. an underestimation of the variance resulting from a zero order regression. Rather weaker trends could be seen from the 3-dimensional graphics. These diagrams show the evolution of the mean square error ratios (z-coordinate) in subject to correlation (x-coordinate) and percentage of missing values (y-coordinate).

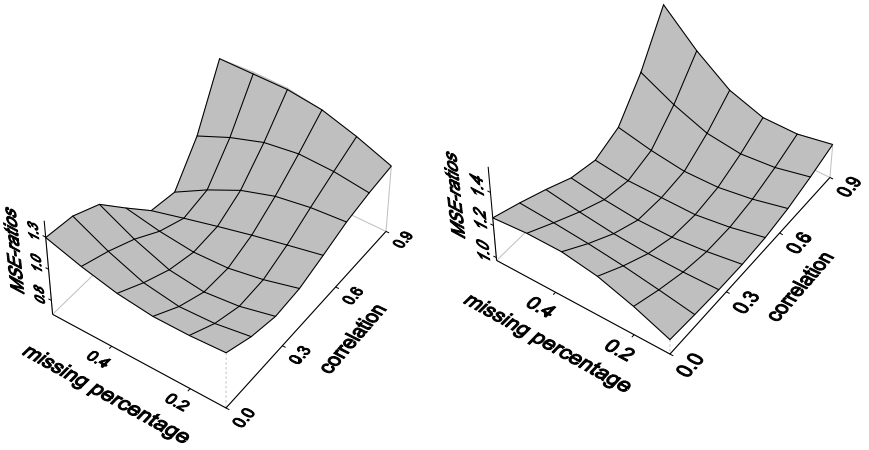

Figure 4.4: MSE-ratios of multiple imputation (left graphic) and ZOR (right graphic)

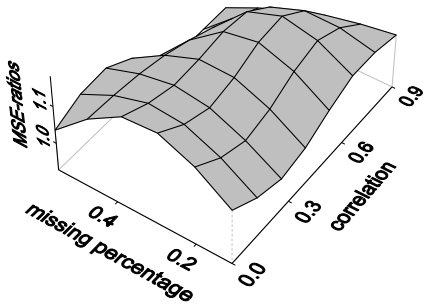

Figure 4.5: MSE-ratios of the mFOR

The multiple imputation always decreases its MSE within severe correlation. If the mean is imputed (zero order regression) an increase of the MSE-ratio can be observed with increasing percentages of missing values. The modified first order regression shows reverse trends, e.g. a 'qualitative' improvement during the raise of the missing-percentage up to 30% is followed by a deterioration during the raise up to 50%.

4.4 Variance

The modified first order regression shows maximal and categorical zero order regression shows minimal variances within severe correlation ($\rho = 0.9$) irrespective to the percentage of missing values. Results of similar uniqueness could be seen out of the situation for $\rho = 0.0$ where the true model had the smallest variance and again the modified first order regression had the maximum. Within the recent correlation structure such unique properties weren't found.

Probably the binary variables caused the relatively small variance of the multiple imputation which was expected to be larger than those of the alternatives. Therefore the usage of the single imputation instead of the multiple imputation seems to be justified within this scope. To be more precise, the rarely small differences between single and multiple imputation would recommend the use of the single imputation because of its minor effort. Procedures based on the unconditional mean imputation as the ZOR and the cZOR confirm some underestimation of the variance and therefore provide quantitative differences in comparison with the recent methods.

Despite more or less individual deviations it can be stated that variances will increase with an increasing percentage of missing values. This could be seen in figure 4.6 through 4.9. Complete case analysis, zero order regression, pi imputation and modified first order regression show this increase for the values itself in contrast to the alternatives where this phenomenon was determined only by analyzing the changes of maxima and minima. According to the previous

figures the correlation structure is situated on the x-coordinate, the missing-percentage on the y-coordinate and the variance of the corresponding method on the z-coordinate. Considering the bars parallelly to the y-coordinate verifies the increasing variance for increasing percentages of missing values. The recent methods differ just slightly from this trend. All methods have maximal variances within severe correlation without respect to the missing percentage.

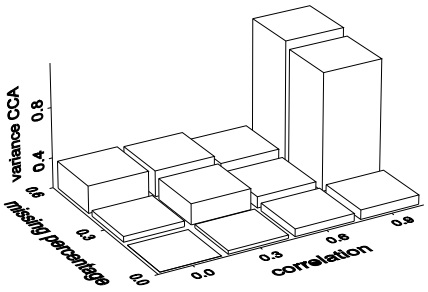

Figure 4.6: $V(\hat{\beta}_2)$ CCA

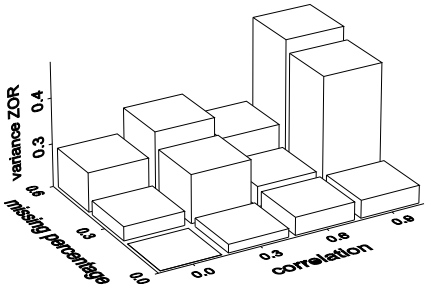

Figure 4.7: $V(\hat{\beta}_2)$ ZOR

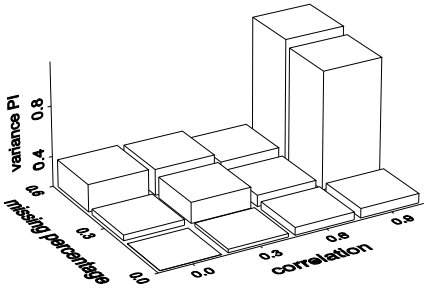

Figure 4.8: $V(\hat{\beta}_2)$ PI

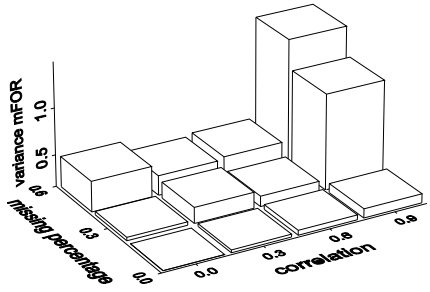

Figure 4.9: $V(\hat{\beta}_2)$ mFOR

4.5 Bias

Whereas the pi imputation has the smallest biases (apart from the unbiased complete case analysis) and also shows least deviates in its values the modified first order regression always overestimates more than the other methods. Single imputation and multiple imputation underestimate irrespectively to the missing-percentage. The degree of this underestimation generally is more severe for the multiple imputation.

5 Summary

The simulation experiment was realized especially for implementing different methods based on the conditional mean imputation and for comparing their results for different settings. Thereby the four methods pi imputation, single imputation, multiple imputation and a single imputation based on a modified first order regression were compared with standard methods (complete case analysis, mean imputation and imputation of the mode) and last but not least with the 'true' data set. Beginning with the discussion of the MSE-ratios as a kind of base for further analysis the pi imputation emerged as the only method representing a steady trend across the different settings. This phenomenon was also observed by analyzing the second model and therefore emphasized the unique position of the pi imputation. Apart from individual local properties, an increase of the variances with increasing percentages of missing values had been observed. A general characteristic could be seen in the maximal variances within severe correlation across all methods irrespective to the missing-percentage. Additional to its unique trend by considering the MSE-ratios the pi imputation showed smallest biases and is thus emphasizing its impact within this experiment. Though small biases are one reason for small mean square errors, they could often be considered exclusively because of their low influence to the MSE in comparison to the values of the variance. Considering the bias, similar results were seen between single and multiple imputation which consequently favorite the single imputation because of its minor effort and the nonexistence of larger variances of the multiple imputation which should reflect the uncertainty of the unknown value and therefore improve estimates based on the single imputation.

Summarizing the simulation experiment it has to be mentioned that the different imputation methods were implemented without much effort. Conclusions based on the chosen settings and their results however were confined to individual properties except the pi imputation.

A simulation experiment in general has to be considered as a kind of closed system which always has to be discussed with regard to the assumptions and conditions which had been made. The aim of a simulation study could moreover not be the verification of a theoretical provable coherence. In fact it should illustrate that theoretically complex methods can be implemented with relatively small effort. Therefore miscellaneous situations could be simulated and exemplarily discussed, sometimes in accordance with theoretical already known properties.

References

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.

Fahrmeir, L., Hamerle, A. and Tutz, G. (eds) (1996). *Multivariate statistische Verfahren*, 2 edn, de Gruyter, Berlin.

Fieger, A. (1997). C++ Klassen zur Linearen Regression bei fehlenden Kovariablen, *SFB386 – Discussion Paper 61*, Ludwig-Maximilians-Universität München.

Fieger, A., Heumann, C., Kastner, C. and Watzka, K. (1997). Generische Bibliothek zur Linearen Algebra und zur Simulation in C++, *SFB386 – Discussion Paper 63*, Ludwig-Maximilians-Universität München.

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.

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

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

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.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Sample Surveys*, Wiley, New York.

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

Toutenburg, H. (1992). *Lineare Modelle*, Physica, Heidelberg.

Toutenburg, H., Srivastava, V. K. and Fieger, A. (1996). Estimation of parameters in multiple regression with missing X-observations using first order regression procedure, *SFB386—Discussion Paper 38*, Ludwig-Maximilians-Universität München, Munich.

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