Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FB Psychologie
Psychologische Methoden

Discussion Paper

No. 1

Multiple imputation of predictor variables using generalized additive models

June 2013

Roel de Jong, Stef van Buuren & Martin Spiess

Psychologische Methoden, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

Multiple imputation of predictor variables using generalized additive models

Roel de Jong, Stef van Buuren & Martin Spiess

Abstract

The sensitivity of multiple imputation methods to deviations from their distributional assumptions is investigated using simulations, where the parameters of scientific interest are the coefficients of a linear regression model, and values in predictor variables are missing at random. The performance of a newly proposed imputation method based on generalized additive models for location, scale and shape (GAMLSS) is investigated. Although imputation methods based on predictive mean matching are virtually unbiased, they suffer from mild to moderate under-coverage, even in the experiment where all variables are jointly normal distributed. The GAMLSS method features better coverage than currently available methods.

1 Introduction

In the social sciences, the parameters of scientific interest are often the regression coefficients of a linear model (LM) which are predominantly estimated using the (OLS) estimator. When there are missing values in predictor variables, and the probability of a missing value depends on the response variable given the covariates, a complete case analysis (CCA) will generally lead to invalid inference.

Multiple imputation (MI; Rubin, 1987) is a statistical mode of inference to draw conclusions from incomplete data sets. It involves generating plausible values, called imputations, for each missing datum in the data set.

These imputations are generated by an imputation method, based on the observed part of the data set and assumptions made by the imputer. The resulting imputed data set, which is free of missing data, is used to estimate the parameters of scientific interest using standard methods. Unfortunately, treating observed and imputed data on equal footing generally leads to invalid inference, since the analysis does not take into account the additional uncertainty about the imputed data. MI is designed to solve this deficit; in contrast with single imputation, MI requires the imputation and analysis step to be performed at least two times, after which the resulting analyses are aggregated or 'pooled' to form the final inference using simple rules.

The validity of MI based inference depends on the degree to which the assumptions posited by the imputer are met. He and Raghunathan (2009) investigated the sensitivity of several imputation methods to deviations from their distributional assumptions. In a setting with three variables and missing data which are missing completely at random (MCAR), they demonstrate that with respect to the estimation of regression coefficients, currently used MI procedures can in fact give worse performance than complete case analyses under seemingly innocuous deviations from standard (multivariate normality) simulation conditions.

In this paper, the sensitivity of several imputation methods is investigated when the values in the predictor variables are missing at random (MAR). Additionally, the performance of a newly proposed imputation method based on generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005) is investigated.

2 Data Generating Process and Analysis Model

Since we are interested in the sensitivity of imputation methods to violations of assumptions about the imputation model, we assume throughout that the true data generating process and the analysis model coincide. In particular, we assume that the data are generated and analyzed according to the linear model

$$y_i = \alpha + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + u_i \qquad i = 1, \ldots, N, \tag{1}$$

with $\boldsymbol{\beta}$ being the $(k \times 1)$ parameter vector of main interest and u_i a latent independent and identically distributed (iid) error variable with $E(u_i) = 0$, which is independently generated from the predictors $\mathbf{x}_i$. The unknown parameters will be estimated by OLS.

In the simulation study, we will generate missing values in the first predictor x_{i1}, where the probability of a missing value in x_{i1} is conditionally independent of x_{i1} given all other variables $\mathbf{w}_i = \left(y \ , \ \mathbf{x}_{i,-1}^{\mathrm{T}}\right)^{\mathrm{T}}$. It is important to note that in imputation models, x_{i1} becomes the response variable for which imputations, i.e. predictions, are generated conditional on $\mathbf{w}_i$.

3 Imputation Methods

For reasons of scientific reproducibility, transparency, and practical relevancy, only imputation methods implemented and made available for the open-source statistical environment and programming language R (R Development Core Team, 2011) are considered. In particular, we tested the functions `mice`, version `2.10` (van Buuren and Groothuis-Oudshoorn, 2010) with imputation methods ‘linear model’ (LM) as a special case of the class of generalized linear models (GLM) and ‘predictive mean matching’ (PMM), `mi`, version `0.09-14` (Gelman et al., 2010) with imputation method GLM and `Hmisc` version `3.9-0` using function `aregImpute` (Harrell, 2010) with a PMM method based on a non linear regression imputation method denoted as nlinPMM. The newly proposed imputation method based on GAMLSS will be described in 3.4.

3.1 Generalized linear model (GLM)

In generalized linear models (McCullagh and Nelder, 1989) the response variable x_1 – for simplicity, index i is dropped in what follows – is assumed to be generated from a (conditional) distribution in the exponential family. The conditional expectation and variance of x_1 are related to the linear predictor through the inverse of a link function $g(\cdot)$:

$$\begin{aligned}\mu(\mathbf{w}) &:= \mathrm{E}\left(x_1|\mathbf{w}\right) = g^{-1}(\tilde{\alpha} + \mathbf{w}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}) \\ \sigma^2(\mathbf{w}) &:= \mathrm{Var}(x_1|\mathbf{w}) = v(g^{-1}(\tilde{\alpha} + \mathbf{w}^{\mathrm{T}}\tilde{\boldsymbol{\beta}})),\end{aligned} \tag{2}$$

where $v(\cdot)$ is the scedastic function mapping the predicted mean to the conditional variance; its form follows from the specified distribution and link function.

The GLM requires specification of a conditional distribution for x_1 and link function $g(\cdot)$. If values in x_1 are MAR, then the usual assumption is that the conditional distribution of the observed x_1 given $\mathbf{w}$, $f(x_1|\mathbf{w})$, is the same as the conditional distribution of the missing x_1. However, specification of $f(x_1|\mathbf{w})$ is typically based on the observed range of values of x_1 and it is often implicitly assumed that the marginal distribution $f(x_1)$ belongs to the same family as the conditional distribution $f(x_1|\mathbf{w})$. For example, for continuous variables x_1 often the identity link together with a conditional normal distribution is assumed to generate imputations. However, the congruency of the conditional that needs to be modeled and the marginal distribution which is only of interest if x_1 does not dependent on all other variables, is only true in some special cases. One example is when x_1 and y are distributed bivariate normal conditional on $\mathbf{x}_{-1}$, in which case $\mathrm{E}\left(x_1|\mathbf{w}\right) = \tilde{\alpha} + \mathbf{w}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ and $\mathrm{Var}(x_1|\mathbf{w}) = \sigma^2$, which corresponds to the LM method for imputing continuous data. Note that linearity and homoscedasticity does not hold if x_1 and y are not (conditionally) bivariate normally distributed (Spanos, 1995). Another case where the marginal and the conditional distribution families coincide is when x_1 is binary (Efron, 1975). On the other hand, suppose $x_1 \sim \mathrm{Poisson}(\lambda)$. Then the conditional distribution of $x_1|\mathbf{w}$ can neither be a Poisson nor a Negative Binomial distribution. Instead, the imputation model for x_1 should allow for under-dispersion (for details, see de Jong, 2012).

The generalized linear model is implemented in almost all imputation software, and together with predictive mean matching remain one of the most widely used imputation methods. A disadvantage of the method is that the model may be too restrictive for the data at hand, hence leading to imputations that distort the information, and thus lead to invalid inferences.

3.2 Predictive Mean Matching (PMM)

PMM was first proposed in the seminal book of Rubin (1987) and in Little (1988). A comparison with an imputation method based on the linear model when estimating the marginal mean and marginal distribution of a variable with missing values was undertaken in a simulation study by Schenker and Welsh (1988). None of the articles mentioned above derived the large-sample properties of the method, and only Schenker and Welsh (1988) tested the method empirically, although with respect to marginal statistics. Despite this, the method has been found to work well in simulation studies (e.g., Schenker and Taylor, 1996; Andridge and Little, 2010; Yu, Burton and Rivero-Arias, 2007) and is currently adopted as the standard method in the widely used `mice` package for multiple imputation inference with respect to $\boldsymbol{\beta}$.

PMM can be seen as a type of random k-nearest-neighbor method. Given a metric $d : \mathbb{R}^{2k} \to \mathbb{R}$ and a query point for which x_1 is missing, the p nearest neighbors of the query point are sought to obtain a set of p donor values from which an imputation is randomly drawn. What differentiates PMM from nearest neighbor methods is the metric used, which is defined in terms of the linear predictor of the reverse linear regression:

$$d_{PMM}(\mathbf{a}, \mathbf{b}) = |\mathbf{a}\dot{\boldsymbol{\beta}} - \mathbf{b}\dot{\boldsymbol{\beta}}| = |(\mathbf{a} - \mathbf{b})\dot{\boldsymbol{\beta}}|, \tag{3}$$

where $\mathbf{a}$ and $\mathbf{b}$ are realizations of $\mathbf{w}$, and $\dot{\boldsymbol{\beta}}$ are (approximated) draws from the posterior distribution of the parameters of the reverse linear regression, i.e. the regression of x_1 on $\mathbf{w}$. Since matching is done using the linear predictor and imputed values are 'live' or observed, the method can also be used for the imputation of non-continuous data without the need for iterative maximum likelihood fitting.

Problems may occur when regions of the sample space are sparsely populated, possibly due to the missing data mechanism. Because of the low number of observed values of x_1, the same donors are considered for each missing value, which might result in underestimation of the variance of the multiple imputation based estimator of $\boldsymbol{\beta}$. Further, PMM is unable to extrapolate correctly from observed values to truncated regions, leading to a biased estimate of the regression slope. Although often heralded for imputing 'realistic' values, the resulting inability to properly inter- and extrapolate can be a serious weakness of the method, especially when the missing data mechanism is selective, for example, if it deletes observations in the tails of the distribution.

3.3 PMM Based on Nonlinear Regression (nlinPMM)

The function `aregImpute` in the package `Hmisc` is another readily available alternative for end users. The corresponding imputation method has not been published: there are no large-sample results available. The primary source of information, apart from the program code itself, is the documentation contained in Harrell (2010). The imputation method is similar to PMM, and allows for nonlinear transformation in the construction of the matching metric.

First, the algorithm finds those transformations of predictors $f_j(\mathbf{w})$ which lead to optimal prediction of a linear transformation of the observed values of the variable to be imputed, x_1, in the following additive model:

$$\tilde{c} + x_1\tilde{d} = \tilde{\alpha} + \sum_{j=1}^{k} f_j(w_j)\tilde{\boldsymbol{\beta}}_j + \nu, \tag{4}$$

where the $f_j(\cdot)$ are restricted cubic spline basis functions of the predictor variables w_j, $j = 1, \ldots, k$ with a user specified fixed number of knots. After estimation of (4), a variant of PMM using weighted probability sampling of donor values is used to generate imputations, where the weights are inversely proportional to the following distance function:

$$d_{areg}(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{k} |(f_j(a_j) - f_j(b_j))\tilde{\boldsymbol{\beta}}_j|, \tag{5}$$

and where $\mathbf{a} = (a_1, \ldots, a_k)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_k)^{\mathrm{T}}$ are realizations of $\mathbf{w}$. The method uses the simple non-parametric bootstrap to approximate draws from the Bayesian posterior distribution of the parameters of the imputation model. Since the final imputed values are produced using PMM, `aregImpute` can also be used for the imputation of non-continuous data and will be denoted as nlinPMM in what follows.

3.4 Generalized Additive Models for Location, Scale and Shape (GAMLSS)

In the literature, several methods are available which jointly model the conditional expectation and conditional variance, and iteratively estimate both using nonparametric techniques. For example, Yu and Jones

(2004) propose a local linear regression method with estimators based on the local normal likelihood. Rigby and Stasinopoulos (1996) propose a similar idea using semi-parametric additive models based on the penalized normal likelihood. Both approaches involve first fitting the conditional mean using local linear regression or a smoother while holding the conditional variance fixed, and then fitting the conditional variance using local linear regression or smoother while holding the conditional mean fixed. Rigby and Stasinopoulos (2005) propose the GAMLSS model, which allows for relaxation of the normality assumption and the specification of arbitrary families of conditional distributions for x_1, even ones outside of the exponential family.

3.4.1 The Imputation Model

In the GAMLSS imputation method, at least the mean and dispersion parameters of a specified distribution $\mathcal{D}$ are modeled using additive terms:

$$\begin{aligned} \mu =& g_1^{-1}(\tilde{\alpha}_{(1)} + \sum_{j=1}^{k} h_{1j}(w_j)) \\ \sigma =& g_2^{-1}(\tilde{\alpha}_{(2)} + \sum_{j=1}^{k} h_{2j}(w_j)), \end{aligned} \tag{6}$$

where $g_p^{-1}(\cdot)$ are inverse monotonic link functions which relate the parameters of the conditional distribution $\mathcal{D}$ to the predictor variables w_j, $j = 1, \ldots, k$, and h_{pj} represents the type of effect of the jth covariate (Rigby and Stasinopoulos 1996, 2005). This distribution defaults to normal for continuous data, but alternatives can be chosen from a a broad range of alternatives. This enables users in combination with a suitable link function to restrict the drawn imputations to a certain range by specifying for example a truncated normal distribution, and allows for easy generalization of the method to discrete and count data.

If besides location and scale $\mathcal{D}$ has up to two shape parameters $\{\nu, \tau\}$ and the sample size is relatively large, we can extend (6) by modeling these parameters additively:

$$\begin{aligned} \nu =& g_3^{-1}(\tilde{\alpha}_{(3)} + \sum_{j=1}^{k} h_{3j}(w_j)) \\ \tau =& g_4^{-1}(\tilde{\alpha}_{(4)} + \sum_{j=1}^{k} h_{4j}(w_j)). \end{aligned} \tag{7}$$

Since this extended model portrays the conditional distribution $f(x_1|\mathbf{w})$ more accurately due to the fact that instead of only two parameters, i.e. mean and spread, four parameters, i.e., in addition, skewness and kurtosis, are modeled. Hence, the resulting imputations may be of higher quality compared to those whose imputation model solely consists of (6).

3.4.2 Implementation

The `R` implementation of the imputation method uses the `gamlss` package (Rigby and Stasinopoulos, 2005) to fit model (6). Rigby and Stasinopoulos (2005) provide a description of the algorithms used by this package. Implemented smoothing terms h_{pj} include cubic smoothing splines, penalized splines, and local regression. In principle, any smoother can be used; however, penalized B-splines Eilers and Marx (1996) proved to be computationally the most stable. More specifically, the smoother used in the simulation studies in Section 4 consists of a penalized B-spline with 20 knots, a piecewise polynomial of the second degree with a second order penalty, and automatic selection of the smoothing parameter using the Local Maximum Likelihood criterion. For high amounts of smoothing, the fit of this smoother approaches linearity.

Let obs denote the fully observed cases, and mis be the cases where x_1 is missing. Further, let $\mathbf{x}_{1,\text{obs}}$ be the vector of observed $x_{1,i}$'s, $\mathbf{W}_{\text{obs}}$ the matrix of all $\mathbf{w}_i$'s for which x_i is observed and, correspondingly, let $\mathbf{x}_{1,\text{mis}}$ be the vector of missing $x_{1,i}$'s and $\mathbf{W}_{\text{mis}}$ be the matrix of $\mathbf{w}_i$'s for which x_i is missing.

Unfortunately, the package `gamlss` does not support Bayesian inference. Therefore, it is impossible to obtain multiple imputations by drawing from the posterior predictive distribution. To incorporate the added variance due to non-response into the multiple imputation inference, the posterior predictive distribution of the missing values is approximated by the bootstrap predictive distribution (Harris, 1989):

$$f^*(\mathbf{x}_{1,\text{mis}}|\mathbf{x}_{1,\text{obs}}, \mathbf{W}) = \int f(\mathbf{x}_{1,\text{mis}}|\tilde{\boldsymbol{\eta}}, \mathbf{W}_{\text{mis}}) \times f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{x}_{1,\text{obs}}, \mathbf{W}_{\text{obs}}))d\tilde{\boldsymbol{\eta}}, \tag{8}$$

where $\tilde{\boldsymbol{\eta}}$ denote the possible values of the imputation model parameters, $\hat{\boldsymbol{\eta}}(\mathbf{x}_{1,\text{obs}}, \mathbf{W}_{\text{obs}})$ is the estimator of said parameters, and $f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{x}_{1,\text{obs}}, \mathbf{W}_{\text{obs}})$ is the sampling distribution of the imputation parameters evaluated

Algorithm 1 GAMLSS imputation

```
1. Fit Model (6), possibly extended to (7), using the observed data $\{\mathbf{x}_{1,\text{obs}}, \mathbf{W}_{\text{obs}}\}$. This is the estimated
   model used to resample $\mathbf{x}_{1,\text{obs}}$ in step (2).

2. Resample $\mathbf{x}_{1,\text{obs}}$ as follows:

       $\mathbf{x}^*_{1,\text{obs}} \sim \mathcal{D}(\hat{\mu}, \hat{\sigma}) \quad \text{or} \quad \mathbf{x}^*_{1,\text{obs}} \sim \mathcal{D}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$

   Define a bootstrap sample $B := \left\{\mathbf{x}^*_{1,\text{obs}}, \mathbf{W}_{\text{obs}}\right\}$

3. Refit model (6) or (6) and (7) using $B$. Draw $n_{\text{mis}}$ imputations for $\mathbf{x}_{1,\text{mis}}$ as follows:

       $\tilde{\mathbf{x}}_{1,\text{mis}} \sim \mathcal{D}(\dot{\mu}, \dot{\sigma}) \quad \text{or} \quad \tilde{\mathbf{x}}_{1,\text{mis}} \sim \mathcal{D}(\dot{\mu}, \dot{\sigma}, \dot{\nu}, \dot{\tau})$

4. Repeat step (2) and (3) $m$ independent times, where $m$ is the number of imputations.
```

at the estimated values of the parameters. $f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{x}_{1,\text{obs}}, \mathbf{W}_{\text{obs}}))$ is simulated by performing the parametric bootstrap, and acts as a surrogate for the posterior distribution of the parameters of the imputation model. A full description of the algorithm is given in Algorithm 1. An advantage compared to a fully Bayesian approach is that no prior information – which is typically lacking – needs to be specified.

Even though the implementation of penalized smoothing splines in the package `gamlss` is considered to be the most stable, in some cases Algorithm 1 may fail to converge. This is frequently traced to the algorithm which selects the smoothing parameter. The implementation of the imputation method catches such an occurrence, and then falls back to a cubic smoothing spline (function `cs` in the `gamlss` package) with a fixed smoothing parameter consisting of one additional degrees of freedom on top of the linear term, which indicates a very large amount of smoothing. Even with these measures in place, GAMLSS may fail to converge in some scenarios, especially for low sample sizes, as will become apparent in the simulation studies.

A difference between `aregImpute` and the proposed imputation method is that the former fixes the number of knots of the transformations f_j to a default fixed value, while GAMLSS optimizes the smoothing parameter of h_{ij} using cross-validation; after all, the performance of a smoother is extremely sensitive to the appropriateness of the chosen smoothing parameter. Also, `aregImpute` draws by default imputations from the observed values using PMM, while GAMLSS samples imputations from $\mathcal{D}$ using the estimated parameter values

3.4.3 Discussion of Assumptions

When the IDGP contains many predictors, a problem that non-parametric smoothers face is the 'curse of dimensionality', where the volume of predictor space grows so fast that the available data becomes sparse; this generally leads to an explosion of the variance of the non-parametric estimator, and computational problems. One strategy to cope with the curse is to force predictors to enter the model additively as in (6). Although the additivity assumption allows for the incorporation of a moderate number of predictors in the imputation model, it cannot capture the effects of potential interactions between the predictors of the imputations model. When interaction terms need to be included in the model, they should be explicitly specified; even if the direct regression is additive, it is generally unknown if the functional relationship relating the predictors of the imputation method to the parameters of the specified conditional distribution is also additive. Another possible limitation is that the functions $h_{ij}(\cdot)$ to be estimated in (6) and (7) should be sufficiently smooth; functions with pronounced discontinuities might lead to imputations that cannot adequately reflect such jumps. Finally, estimating arbitrary smooth functions using flexible non-parametric estimators requires more data than required for a linear regression, and the GAMLSS imputation method might not be appropriate for small samples.

4 Simulation Experiments

First we will empirically investigate the performance of all the imputation methods described in Section 3 in the context of a simple linear regression model following (1) with a single predictor variable x with missing values. The true parameter values are $\alpha = 0$ and $\beta = 1$; since the intercept is usually of no scientific interest, only the regression slope β is reported in the results. It should be emphasized that the predictor x and response variable y swap roles in the imputation model.

Further note that constraining the scope of the experiments to a single variable with missing values allows for the isolation of defects in the imputation methods from possible confounding issues stemming from the fully conditional specification framework (c.f. van Buuren, Brand, Groothuis-Oudshoorn and Rubin, 2006) in which the imputation methods are ultimately embedded. A further constraint is the limitation to a single predictor variable in the imputation model. However, obtaining acceptable performance in this basic setting is not a trivial task, and acceptable performance is a prerequisite for more involved scenarios. Further, this basic scenario simplifies experimenting with the distribution of the predictor x and other parameters of the simulation, because computational cost is less high than with multiple predictors. Finally, two simulation studies are performed with multiple predictor variables.

The three simulation parameters which are systematically varied are the distribution of x, the coefficient of determination R^2, and sample size N. All studies have 1000 replications and $m = 10$ imputations, and a normal distribution for the errors of the complete data model (1). A very important factor is the distribution of the predictor x. The following continuous densities are considered: Standard normal, skew-normal, which generalizes the normal distribution by introducing an additional skewness parameter, which is set here to $\kappa = 5$ leading to a positively skewed distribution, uniform on $[0, 1]$, squared uniform or standard beta distribution with shape parameters 0.5 and 1, and the T distribution with three degrees of freedom as an example of a heavy-tailed distribution. Further, as an example of a non-continuous distribution the Poisson distribution with $\lambda = 2$ is considered. The study is performed for all combinations of the distributions listed above and the factor levels $R^2 \in \{.25, .50, .75\}$ and $N \in \{200, 500, 1000\}$. Since the conclusions from the simulation experiment are not affected if the results for $N = 500$ or $R^2 = 0.5$ are shown, they are omitted in the tables, but are available upon request.

For all studies, the following missing data mechanism is imposed:

$$p(s|y) = \begin{cases} (\varphi_1)^{1-s}(1-\varphi_1)^s & \text{if } y < \tilde{y} \\ \\ (\varphi_2)^{1-s}(1-\varphi_2)^s & \text{if } y \geqq \tilde{y}, \end{cases} \tag{9}$$

where $\tilde{y}$ is the sample median. Setting $\varphi_1 = .1, \varphi_2 = .7$ results in 40% missing data in x, where the chance of a missing datum in x is .1 when the corresponding value of y is smaller than the sample median of y, and .7 when it is larger than the median; note that this missing data mechanism is MAR (Rubin, 1987). While holding the missing data mechanism constant at (9), the coefficient of determination determines the extent to which the missing values are MAR, with R^2 approaching 0 implying the missing data are in fact MCAR and evenly spread, and a high value of R^2 giving rise to a strongly systematic missing data mechanism with the potential of thinning out selected regions of the sample space. To replicate parts of the study of He and Raghunathan (2009), two simulation studies are conducted where 40% of the values of x_1 are MCAR, and $f(x)$ is the normal or the Beta distribution; all other simulation parameters are equal to those in the MAR experiments.

The results of the simulation study are reported in tables which all share common elements. The columns identify the mode of inference, where COM stands for the complete data analysis; this is the analysis on the complete data, before any cases are deleted, and should be taken as the golden standard. CCA denotes complete case analysis, which represents the analysis on the completely observed cases only. All other entries are multiple imputation inferences using the indicated imputation method (LM, GLM, PMM, nlinPMM, GAMLSS). Further, the output of the package `mi` is suppressed in most Tables since it performs very similar to `mice`.

All imputation methods are assessed on four criteria. (1) Number of simulations which failed due to computational problems (FAIL), (2) bias, given as the difference of the true parameter value, which is 1, and the mean of the estimates ($m(\hat{\beta})$) over 1000 simulations, (3) the mean of the estimated standard errors ($m(\widehat{\mathrm{SD}})$) and (4) coverage (COVER), which is equal to the proportion of replications where the true parameter lays inside the 95% confidence interval as produced by the multiple imputation inference.

Two additional simulation experiments were conducted with four jointly independent normally distributed predictor variables, each having unit variance and associated regression coefficients $\boldsymbol{\beta} = \left[1, \sqrt{2/3}, \sqrt{2/3}, \sqrt{2/3}\right]$, and unit residual variance. Note that also in these experiments, only x_1 is afflicted by missing values. Further, R^2 is fixed at .75; the regression coefficients, residual variance, and covariance matrix of the predictor are chosen such that the amount of variance explained by x_1 equals $R^2_{x_1} = .50$, which is canonical in the sense that it is the middle value of the set of coefficients of determination $\{.25, .50, .75\}$ in the simple experiments with only x_1.

For the studies with multiple predictors, the following missing data mechanism is imposed:

$$p(s|lp) = \begin{cases} (\varphi_1)^{1-s}(1-\varphi_1)^s & \text{if } lp < \tilde{lp} \\ \\ (\varphi_2)^{1-s}(1-\varphi_2)^s & \text{if } lp \geqq \tilde{lp}, \end{cases} \tag{10}$$

where

$$lp = -.4x_2 - .4x_3 - .4x_4 + .50y \tag{11}$$

is an approximation of the reverse linear predictor, and $\tilde{lp}$ is the sample median of lp. Again, setting $\varphi_1 = .1$ and $\varphi_2 = .7$ results in 40% missing data in x_1.

5 Results

5.1 Normal

The first simulation scenario features a normal distribution for the predictor variable x, which implies that x and y are distributed bivariate normal; this is a standard simulation condition for assessing the performance of imputation methods. Results of the simulation study are presented in Table 1. Since the missing data are MAR, CCA is biased, which leads to under-coverage. The under-coverage of CCA seems unaffected by the coefficient of determination, but becomes worse with increasing sample size.

Since x and y are distributed bivariate normal, both the direct and reverse regression are linear, and the linear model (LM) method is expected to be perfectly adequate; in fact, when the missing values are MAR, the use of LM is only warranted when the observed data are distributed according to a multivariate normal distribution. This is confirmed in the simulation results, where the LM based estimates are virtually unbiased and have nominal coverage. As is to be expected, the aggregated standard errors are larger than those of the golden standard set by the complete data analysis. This loss of precision is due to the missing values; multiple imputation does not make up data.

The adopted missing data mechanism does not truncate the sample space, but thins out the sample space for large values of y. Even in this case PMM suffers from mild to moderate under-coverage, with coverage rates ranging between .892 and .915. The standard errors are slightly smaller than those of the LM imputation method, which is counter-intuitive since PMM is more flexible and uses less information, i.e. that the errors are normally distributed, external to the data than the LM method. With respect to bias, PMM performs roughly equal to LM, which means very limited empirical bias. The unsatisfactory performance of PMM did not arise in the simulation studies of He and Raghunathan (2009), probably because they simulated a MCAR missing data mechanism which does not thin out the sample space as selectively as missing mechanism (9). The performance of nlinPMM is comparable or slightly worse than that of PMM, with coverages ranging between .866 ($R^2 = 0.5$) and .916 ($R^2 = 0.5$). A possible explanation is that nlinPMM also performs a predictive mean matching step, and thereby suffers from the same problem as the `mice` implementation of the PMM algorithm. When the imputation model features multiple predictors and with a $R^2_{x_1} = .50$, the coverages of PMM for β_1 range from .892 to .904, and the coverages of nlinPMM range from .909 to .924; although the coverages are slightly better than in the single predictor study, they are still clearly below the nominal level.

GAMLSS is expected to give unbiased results, albeit with some loss of efficiency compared to LM. The conditional distribution $\mathcal{D}$ is specified to be normal. Looking at the results in Table 1, bias is comparable to that of LM and thus negligible, although the standard errors are moderately larger than those of LM; this is the price to pay for the greater flexibility of the model. However, for larger sample sizes, the difference in efficiency diminishes. GAMLSS unfortunately fails to converge in a total of three cases for the lowest sample size condition; however, in the larger sample conditions no problems arise.

Lastly, the performance of GAMLSS in the study with multiple predictors (not shown, but available upon request) is comparable to the study with a single predictor; this indicates that, as expected, if the additivity assumption is correct, GAMLSS seems to successfully circumvents the curse of dimensionality.

Include Table 1 here

5.2 Skew-Normal and Uniform

The second and third simulation study feature a marginal skew-normal and uniform distribution for the predictor variable x, respectively. It can be shown that the reverse conditional expectation of x given y is non-linear, and the conditional variance is heteroscedastic; thus, the LM method is expected to fail. Indeed, as the results in Table 2 indicate, the LM method breaks down with coverages ranging between 0.455 and 0.918. The under-coverage seems primarily due to substantial empirical biases ranging from .051 to .076 ($R^2 = 0.5$), which are comparable to those of the CCA. Although PMM and nlinPMM have negligible bias, their coverage rates are equal to those of the normal study, and remain to show mild to moderate under-coverage. The performance of PMM and nlinPMM continues to be substandard irrespective of the conditional distribution of x, and will not be addressed in the discussion of the remaining studies with a single predictor.

For the GAMLSS approach, the conditional distribution $\mathcal{D}$ of x is specified as normal. Since the conditional mean is not restricted to be a linear function of the predictors, and the conditional variance is not restricted to be constant, the GAMLSS approach is expected to offer robust performance in the skew-normal scenario. Generally speaking, these expectations are fulfilled: only the case with $n = 200$ and $R^2 = 0.75$ features slight undercoverage. Comparison of estimated standard error size is meaningless since only GAMLSS features adequate coverage.

A simulation experiment has also been conducted with x_1 distributed skew-normal with a standardized third cumulant of .85, and $\mathbf{x}_{-1}$ distributed standard normal, with all predictors jointly independent. Again in this study, the results with respect to β_1 are similar to the results with just one predictor x_1. Further, GAMLSS performs comparable to the study with a single predictor. Thus, the results are omitted but are available upon request.

Include Table 2 here

In the case of x having a standard uniform distribution, it may be desirable to restrict the imputed values to lay between zero and one; this can be accomplished by letting $\mathcal{D}$ be the (Generalized) Beta distribution. Finally, we test the normal distribution, denoted by GAMLSS (normal) in the table, even though this choice of $\mathcal{D}$ leads to the imputation of potentially unrealistic values.

Include Table 3 here

The results of the GAMLSS imputation method in Table 3 when x has the uniform distribution indicate that imputing under the Normal and Generalized Beta distribution gives comparable and adequate results. Imputation under the normal model (LM) also gives negligible empirical bias, demonstrating that, apart from the first two moments, imputations for x do not need to match the exact shape of the conditional distribution $f(x|y)$. Finally, the good performance of imputations under the Generalized Beta distribution demonstrates that the goals of generating plausible imputations and consistent imputations are not necessarily incompatible with each other.

5.3 Uniform Squared (Beta)

This simulation study can be interpreted as an assessment of the transform, then impute strategy as described in von Hippel (2009), where the predictor x is created by taking the square of the original predictor variable z, and where z is standard uniformly distributed. In this case, the conditional expectation of x given y deviates significantly from linearity; moreover, the skewness of the conditional distribution of x given y leads to the occurrence of outliers in x for large values of y. Arguably, this scenario features a conditional distribution whose features are difficult to estimate, with the attrition of the missing data mechanism (9) further exacerbating the situation. Given that the values of x lay in the $(0,1)$ interval, one might want to impute only realistic values; therefore, $\mathcal{D}$ is chosen to be a generalized beta distribution as in Section 5.2.

Include Table 4 here

Unfortunately, GAMLSS with a generalized Beta distribution breaks down for high values of the coefficient of determination. When $R^2 = .75$ there is moderate undercoverage, which becomes worse for larger sample sizes. Apparently, for smaller sample sizes, the larger standard errors camouflage the empirical bias. The results might indicate that for this scenario, imputation model (6) is not adequate. Performance might improve with a varying bandwidth; unfortunately, this feature has not been implemented yet in the GAMLSS package. Despite producing potential unrealistic values in the form of negative imputations, GAMLSS with a normal distribution is on target, with the empirical bias being quite acceptable.

5.4 Student's T

The fourth simulation study feature a marginal T distribution with three degrees of freedom for the predictor variable x. For the GAMLSS method, $\mathcal{D}$ is specified to be normal. In the simulation study of He and Raghunathan (2009), all tested imputation methods broke down when the distribution of x was strongly heavy tailed. As the results in Table 5 indicate, this is also true for the GAMLSS method, which features biases which are systematically bigger than the LM method, and coverage rates ranging between .893 ($R^2 = 0.5$) and .943. The results of this study suggest that the GAMLSS method, despite its flexibility, is unsuitable for imputation when x has a heavy tailed distributions. While of all methods the coverage rates of GAMLSS are closest to the nominal level, this seems largely due to inflated standard errors.

Include Table 5 here

5.5 Poisson

The Poisson simulation study features the mi package, which offers an implementation of the generalized linear model (GLM) method where the conditional distribution of x is specified to be a Poisson distribution (POISSON). It can be shown that if x follows a Poisson distribution, then the true conditional distribution of $f(x|y)$ is an under-dispersed distribution (de Jong, 2012); using Poisson regression is therefore expected to fail. A polytomous regression model may also be adopted to generate imputations for count data (POLYT); in this study we used mice which provides a GLM where x has a conditional categorical distribution. Because GAMLSS only implements a Poisson distribution and overdispersed count distributions, $\mathcal{D}$ is choosen to be Normal, which results in the imputation of 'unrealistic' values.

As expected, the results in Table 6 show very low coverages for the polytomous and Poisson regression methods. In contrast, the GAMLSS method seems to offer good performance, although the empirical bias for the case when $R^2 = .25$ is somewhat disquieting.

Include Table 6 here

6 Conclusion

The LM and GLM are parametric regression models and pose restrictions on the functional form of the conditional mean and variance of the variable with missing values. These restrictions, if not correct, may lead to inconsistent estimation of the parameters of scientific interest, and ultimately to invalid multiple imputation inferences; therefore, it is expected that imputation methods which jointly estimate the conditional expectation and conditional variance using non-parametric techniques offer better performance. The proposed GAMLSS method models parameters of a specified distribution $\mathcal{D}$ using additive smoother terms, which in combination with a suitable link allows for easy generalization of the method to discrete and count data.

This paper reports the results of a simulation study where the data generating process of scientific interest assumed by the analyst consists of a linear regression model with missing values in a single predictor variable, and a strongly systematic MAR mechanism. Experimental conditions include the marginal distribution of the predictor with missing values, the coefficient of determination, and sample size. Although the PMM and nlinPMM imputation methods are virtually unbiased, they suffer from mild to moderate under-coverage in all conducted experiments, including the experiment where all variables are jointly normal distributed. The LM method performs excellent when the variables are jointly normal distributed, but breaks down in most cases when the distribution of the predictor deviates from normality, and the reverse regression becomes non-linear; performance is worst when the coefficient of determination is high. In contrast, the GAMLSS method features better coverage than currently available methods.

In this simulation study we restricted attention to models with one dependent and one independent variable. However, as mentioned in sections 4 and 5.1 we also ran simulations with four covariates. The results of this limited study – in the scenario considered, all the variables are normally distributed – imply that the proposed GAMLSS method works also well in larger models.

Thus, based on the simulation results, we recommend to use the GAMLSS imputation method if there is doubt with respect to the parametric imputation models, which may even be the case in standard situations, like imputing a continuous variable that potentially may be used as a covariate in an analysis model. In addition, GAMLSS provides a suitable alternative to PMM in situations where PMM is problematic, e.g. when the number of potential donors is small, or when imputations should extend beyond the data values, as is censoring. Further, it is recommended to impute all continuous variables using the normal distribution (without rounding), even if this means that the resulting imputations are unrealistic if the GAMLSS imputation method is used.

References

Andridge, R.R. & Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review, 78,* 40–64.

de Jong, R. (2012). *Robust Multiple Imputation.* PhD thesis, University of Hamburg. http://ediss.sub.uni-hamburg.de/volltexte/2012/5971/

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association, 70(352),* 892–898.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11(2),* 89–121.

Gelman, A., Hill, J., Su, Y.-S., Yajima, M., & Pittau, M.G. (2010). *mi: Missing Data Imputation and Model Checking.* R package version 0.09-11.

Harrell, F.E. (2010). *Hmisc: Harrell Miscellaneous.* R package version 3.8-3.

Harris, I. (1989). Predictive fit for natural exponential families. *Biometrika, 76(4),* 675–684.

He, Y. & Raghunathan, T. (2009). On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions. *Communications in Statistics – Simulation and Computation, 38(4),* 856–883.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6(3),* 287–296.

McCullagh, P. & Nelder, J. (1989, 2nd ed.). *Generalized Linear Models.* Taylor and Francis.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rigby, R.A. & Stasinopoulos, D.M. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing, 6(1),* 57–65.

Rigby, R.A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3),* 507–554.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

Schenker, N. & Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis, 22,* 425–446.

Schenker, N. & Welsh, A. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics, 16(4),* 1550–1566.

Spanos, A. (1995). On normality and the linear regression model. *Econometric Reviews, 14(2),* 195–203.

van Buuren, S., Brand, J., Groothuis-Oudshoorn, C. & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76(12),* 1049–1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2010). *MICE: Multivariate Imputation by Chained Equations in R.* R package version 2.10.

von Hippel, P.T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology, 39(1),* 265–291.

Yu, L.-M., Burton, A. & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data, *Statistical Methods in Medical Research, 16,* 243–258.

Yu, K. & Jones, M. C. (2004). Likelihood-Based Local Linear Estimation of the Conditional Variance Function. *Journal of the American Statistical Association, 99(465),* 139–144.

Table 1: Normal distribution

		COM	CCA	LM	PMM	nlinPMM	GAMLSS
$N = 200$ $R^2 = 0.25$	$m(\hat{\beta})$	0.996	0.865	0.980	0.974	0.974	0.999
	$m(\widehat{\text{SD}})$	0.123	0.152	0.155	0.154	0.141	0.175
	COVER	0.949	0.833	0.949	0.896	0.882	0.950
	FAIL	0	0	0	0	0	2
$N = 200$ $R^2 = 0.75$	$m(\hat{\beta})$	1.001	0.956	1.001	1.011	0.994	1.006
	$m(\widehat{\text{SD}})$	0.041	0.056	0.051	0.049	0.048	0.062
	COVER	0.964	0.867	0.950	0.892	0.868	0.948
	FAIL	0	0	0	0	0	0
$N = 1000$ $R^2 = 0.25$	$m(\hat{\beta})$	0.999	0.871	0.994	0.995	0.993	1.000
	$m(\widehat{\text{SD}})$	0.055	0.068	0.068	0.066	0.062	0.072
	COVER	0.947	0.525	0.956	0.910	0.896	0.950
	FAIL	0	0	0	0	0	0
$N = 1000$ $R^2 = 0.75$	$m(\hat{\beta})$	1.000	0.955	1.000	1.003	1.000	1.002
	$m(\widehat{\text{SD}})$	0.018	0.025	0.023	0.022	0.021	0.024
	COVER	0.952	0.544	0.944	0.915	0.901	0.944
	FAIL	0	0	0	0	0	0

Table 2: Skew - Normal distribution

		COM	CCA	LM	PMM	nlinPMM	GAMLSS
$N = 200$ $R^2 = 0.25$	$m(\hat{\beta})$	1.002	0.924	1.058	0.985	0.975	0.978
	$m(\widehat{\mathrm{SD}})$	0.124	0.164	0.170	0.159	0.143	0.202
	COVER	0.940	0.904	0.918	0.907	0.865	0.953
	FAIL	0	0	0	0	0	0
$N = 200$ $R^2 = 0.75$	$m(\hat{\beta})$	0.999	0.974	1.051	1.019	0.988	1.020
	$m(\widehat{\mathrm{SD}})$	0.041	0.060	0.055	0.054	0.051	0.069
	COVER	0.949	0.925	0.853	0.878	0.830	0.917
	FAIL	0	0	0	0	0	1
$N = 1000$ $R^2 = 0.25$	$m(\hat{\beta})$	1.002	0.924	1.070	0.999	0.996	0.984
	$m(\widehat{\mathrm{SD}})$	0.055	0.073	0.074	0.065	0.062	0.085
	COVER	0.954	0.796	0.835	0.891	0.883	0.952
	FAIL	0	0	0	0	0	0
$N = 1000$ $R^2 = 0.75$	$m(\hat{\beta})$	1.000	0.977	1.052	1.006	1.003	1.007
	$m(\widehat{\mathrm{SD}})$	0.018	0.026	0.024	0.024	0.022	0.027
	COVER	0.957	0.841	0.455	0.889	0.885	0.946
	FAIL	0	0	0	0	0	0

Table 3: Uniform distribution

		COM	CCA	LM	PMM	nlinPMM	GAMLSS (Normal)	GAMLSS (Gen. Beta)
	$m(\hat{\beta})$	1.004	0.872	0.989	0.990	0.987	1.015	1.004
$N = 200$	$m(\widehat{\text{SD}})$	0.123	0.153	0.156	0.153	0.138	0.163	0.165
$R^2 = 0.25$	COVER	0.959	0.856	0.958	0.920	0.879	0.943	0.949
	FAIL	0	0	0	0	0	0	0
	$m(\hat{\beta})$	1.001	0.956	1.006	1.004	0.997	1.018	0.997
$N = 200$	$m(\widehat{\text{SD}})$	0.041	0.056	0.052	0.046	0.045	0.053	0.050
$R^2 = 0.75$	COVER	0.953	0.881	0.957	0.921	0.906	0.940	0.954
	FAIL	0	0	0	0	0	0	2
	$m(\hat{\beta})$	0.999	0.869	0.992	0.997	0.991	0.998	0.990
$N = 1000$	$m(\widehat{\text{SD}})$	0.055	0.068	0.068	0.067	0.062	0.071	0.071
$R^2 = 0.25$	COVER	0.959	0.517	0.959	0.924	0.903	0.959	0.953
	FAIL	0	0	0	0	0	0	0
	$m(\hat{\beta})$	1.000	0.958	1.008	1.002	1.001	1.008	0.994
$N = 1000$	$m(\widehat{\text{SD}})$	0.018	0.025	0.023	0.020	0.020	0.023	0.022
$R^2 = 0.75$	COVER	0.933	0.588	0.933	0.920	0.914	0.948	0.941
	FAIL	0	0	0	0	0	0	0

Table 4: Uniform Squared (Beta) distribution

		COM	CCA	LM	PMM	nlinPMM	GAMLSS (Normal)	GAMLSS (Gen. Beta)
$N = 200$ $R^2 = 0.25$	$m(\hat{\beta})$	0.990	0.907	1.039	0.964	0.960	0.972	0.976
	$m(\widehat{\text{SD}})$	0.123	0.162	0.167	0.155	0.139	0.190	0.168
	COVER	0.942	0.876	0.925	0.902	0.854	0.959	0.960
	FAIL	0	0	0	0	0	0	3
$N = 200$ $R^2 = 0.75$	$m(\hat{\beta})$	0.997	0.992	1.069	1.003	0.993	1.025	0.983
	$m(\widehat{\text{SD}})$	0.041	0.060	0.056	0.049	0.047	0.060	0.057
	COVER	0.943	0.932	0.779	0.889	0.845	0.944	0.934
	FAIL	0	0	0	0	0	0	0
$N = 1000$ $R^2 = 0.25$	$m(\hat{\beta})$	0.998	0.918	1.059	0.994	0.990	0.993	0.977
	$m(\widehat{\text{SD}})$	0.055	0.072	0.073	0.067	0.061	0.080	0.072
	COVER	0.956	0.785	0.866	0.919	0.905	0.940	0.946
	FAIL	0	0	0	0	0	0	1
$N = 1000$ $R^2 = 0.75$	$m(\hat{\beta})$	1.000	0.994	1.071	1.000	1.000	1.011	0.980
	$m(\widehat{\text{SD}})$	0.018	0.027	0.025	0.022	0.021	0.026	0.024
	COVER	0.947	0.938	0.174	0.905	0.897	0.948	0.860
	FAIL	0	0	0	0	0	0	0

Table 5: T distribution

		COM	CCA	LM	PMM	nlinPMM	GAMLSS
	$m(\hat{\beta})$	1.001	0.892	1.012	0.992	0.962	1.032
$N = 200$	$m(\widehat{\text{SD}})$	0.129	0.161	0.166	0.168	0.168	0.206
$R^2 = 0.25$	COVER	0.947	0.892	0.933	0.935	0.914	0.939
	FAIL	0	0	0	0	0	0
	$m(\hat{\beta})$	1.000	0.950	0.999	1.032	0.986	1.017
$N = 200$	$m(\widehat{\text{SD}})$	0.044	0.058	0.054	0.060	0.060	0.085
$R^2 = 0.75$	COVER	0.950	0.866	0.888	0.885	0.896	0.911
	FAIL	0	0	0	0	0	1
	$m(\hat{\beta})$	1.000	0.893	1.013	1.000	0.984	1.013
$N = 1000$	$m(\widehat{\text{SD}})$	0.056	0.069	0.070	0.069	0.070	0.097
$R^2 = 0.25$	COVER	0.950	0.649	0.912	0.878	0.891	0.940
	FAIL	0	0	0	0	0	8
	$m(\hat{\beta})$	0.999	0.955	0.995	1.023	0.998	1.003
$N = 1000$	$m(\widehat{\text{SD}})$	0.019	0.025	0.023	0.025	0.024	0.042
$R^2 = 0.75$	COVER	0.945	0.546	0.812	0.797	0.890	0.923
	FAIL	0	0	0	0	0	10

Table 6: Poisson distribution

		COM	CCA	POLYT	PMM	nlinPMM	POISSON	GAMLSS
	$m(\hat{\beta})$	0.998	0.907	0.547	0.976	0.966	0.860	0.974
$N = 200$	$m(\widehat{\mathrm{SD}})$	0.123	0.162	0.181	0.157	0.143	0.146	0.200
$R^2 = 0.25$	COVER	0.962	0.896	0.218	0.915	0.855	0.933	0.973
	FAIL	0	0	0	0	0	0	0
	$m(\hat{\beta})$	1.000	0.965	0.583	1.020	0.991	0.629	1.015
$N = 200$	$m(\widehat{\mathrm{SD}})$	0.041	0.059	0.108	0.053	0.051	0.081	0.068
$R^2 = 0.75$	COVER	0.935	0.885	0.002	0.872	0.838	0.003	0.924
	FAIL	0	0	0	0	0	0	0
	$m(\hat{\beta})$	0.999	0.911	0.546	0.995	0.992	0.877	0.980
$N = 1000$	$m(\widehat{\mathrm{SD}})$	0.055	0.072	0.080	0.066	0.062	0.063	0.082
$R^2 = 0.25$	COVER	0.948	0.757	0.000	0.887	0.880	0.530	0.956
	FAIL	0	0	0	0	0	0	0
	$m(\hat{\beta})$	1.001	0.969	0.581	1.006	1.002	0.635	1.005
$N = 1000$	$m(\widehat{\mathrm{SD}})$	0.018	0.026	0.048	0.023	0.022	0.038	0.026
$R^2 = 0.75$	COVER	0.952	0.776	0.000	0.892	0.873	0.000	0.944
	FAIL	0	0	0	0	0	0	0

Current Titles Discussion Papers FB Psychologie, Psychologische Methoden

1 de Jong, R., van Buuren, S. & Spiess, M. (2013). *Multiple imputation of predictor variables using generalized additive models.*