Multiple Imputation of Incomplete Ordinary and Overdispersed Count Data

Kristian Kleinke*[1], Roel de Jong [2], Martin Spiess [2], & Jost Reinecke [1]

[1] *Bielefeld University, Faculty of Sociology & Centre for Statistics*
[2] *University of Hamburg, Department of Psychology*

Abstract

Throughout the last couple of years multiple imputation (MI) has become a popular and widely accepted method to address the missing data problem. However, MI solutions for incomplete count data are still not available in most statistical packages. We present count data imputation add-ons for the popular `mice` software in R (van Buuren & Groothuis-Oudshoorn, 2011). Our add-on functions allow to create multiple imputations of incomplete ordinary and overdispersed count data following the chained equations approach of creating multiple imputations (cf. Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren & Groothuis-Oudshoorn, 2011). We furthermore present evaluations of these solutions regarding their ability to produce unbiased parameter estimates and standard errors as well as their ability to cope with missing not at random mechanisms.

Keywords: missing data, multiple imputation, count data
Running head: Count Data Imputation

*Correspondence should be addressed to Kristian Kleinke, Bielefeld University, Faculty of Sociology, Postfach 10 01 31, D–33501 Bielefeld. kristian.kleinke@uni-bielefeld.de

1 Introduction

Throughout the last couple of years, multiple imputation has become an increasingly popular method to handle the missing data problem. Together with full information maximum likelihood estimation (FIML), it has been enumerated among the state of the art procedures to analyze incomplete data (Schafer & Graham, 2002). However, multiple imputation software at the moment has very limited capabilities to impute incomplete count data. In R, for example, the `mi` package allows to generate multiple imputations under the Poisson model (Su, Gelman, Hill, & Yajima, 2009). There are also other software solutions for statistical packages like IVEware for SAS or ice for STATA that support *basic* count data imputation procedures. However, more advanced models like zero-inflation models or multilevel count data models are currently not supported. To remedy this lack of general count data support, we developed a comprehensive count data imputation package called `countimp` (Kleinke & Reinecke, 2011), which allows to create multiple imputations of incomplete ordinary, overdispersed, zero-inflated and multilevel count data. The package is available from the first author. In this paper we focus on ordinary and overdispersed count data.

We begin by giving a brief introduction to missing data and multiple imputation in general and a brief introduction to the chained equations MI approach. We then introduce our imputation procedures for ordinary and overdispersed count data and present two evaluation studies: The first one generally tests our algorithms' ability to plausibly impute missing values under different scenarios and to yield widely unbiased parameter estimates. The second study tests the algorithms' sensitivity to missing not at random processes, a problem that typically affects any missing data procedure and that generally endangers the correct estimation of statistical parameters. Finally, we conclude with a discussion of our findings and name fruitful avenues for future research.

2 Theoretical Background

2.1 Missing Data

Missing data are a nuisance. Whether or not missing data may bias estimations of population quantities depends on the interaction of three factors (Schafer, 1997a): the number of missing values in the data set, the missing data procedure that is being used and the missing data mechanism. Firstly, it is quite easily understandable that the more information is unobserved and thus the more information has to be estimated the more uncertain one can be about one's statistical inferences. In fact, the more values are missing, the larger could be the bias in parameter estimates. Secondly, there are better and worse procedures to deal with missing data, depending on how well the respective

procedure is suited to estimate population parameters and standard errors correctly. For an overview and an evaluation of simple and more sophisticated procedures, see Schafer & Graham (2002) or Kleinke, Stemmler, Reinecke, & Lösel (2011) for example. Thirdly, the randomness or non-randomness of the missing data process determines, how severe bias might be and which missing data procedure should be used: Rubin (1976) introduced important terminology to describe the randomness of missing data processes. He regards the missingness of a value as a probabilistic phenomenon, which means that a missing value occurs in the data set with a certain probability. Let Y be a $n \times p$ data matrix and let R be a matrix with the same dimensions as Y, indicating for every value in Y, if it is missing or not. Y can be split into an observed part Y_{obs} and an unobserved part Y_{mis}. $P(Y)$ is the distribution of Y, depending on unknown parameters θ. $P(R)$ denotes the distribution of R, depending on Y_{obs}, Y_{mis} and unknown parameters ξ. Of special interest are the possible distributions of R, the so-called missing data processes or missing data mechanisms. Rubin (1976) distinguishes three mechanisms:

$$P(R|Y,\xi) = P(R|\xi)$$

$$P(R|Y,\xi) = P(R|Y_{obs},\xi)$$

$$P(R|Y,\xi) = P(R|Y_{obs},Y_{mis},\xi)$$

In the first case, missingness does not depend on observed or unobserved information. The occurrence of a missing value is a completely random phenomenon and missing values are therefore called *missing completely at random* (MCAR), which means that missing values can be regarded as a random sample of Y. Note that all ad hoc missing data solutions like listwise or pairwise deletion or unconditional mean imputation require the mechanism to be MCAR to produce unbiased parameter estimates. The second mechanism allows missingness to depend on Y_{obs}. This mechanism is called *missing at random* (MAR): After controlling for observed information Y_{obs}, missingness in fact is random. FIML and MI procedures typically allow missingness to be MAR. In the third case, missingness depends on unobserved information. Missing information in this case is very hard or even impossible to estimate. This mechanism is called *missing not at random* (MNAR). Note that there is no way to diagnose a MNAR mechanism. One has to thoroughly discuss, if it is likely that statistical estimates are distorted by MNAR mechanisms and to what extent they are biased. Van Buuren & Groothuis-Oudshoorn (2011) for example demonstrate, how a MNAR sensitivity analysis can be conducted in that regard. Although multiple imputation procedures had been designed to work under MCAR and MAR, it has been argued that they are to some extent robust to mild violations of the MAR assumption (e.g. Schafer, 1997a). If MNAR mechanisms

are feared to be present in one's empirical data, the general advice for the practitioner is to find as many variables as possible that are strongly related both to variables with missing values and to missingness itself and include them in the imputation model. This makes missing information to some extent more predictable and thus the missing data mechanism more likely to be MAR (Collins, Schafer, & Kam, 2001; Graham, 2009; Schafer, 1997a; Schafer & Graham, 2002). Note that also special MNAR techniques have been developed. However, they require extensive modeling and guessing (typically both Y and R have to be modeled and some untestable assumptions have to be made). A brief overview of MNAR procedures is given in Schafer (1997a). More thorough introductions and discussions may be found in Enders (2010, 2011), Hedeker & Mermelstein (2011) and van Buuren (2011). Most practitioners however would want to impute under the MAR assumption using a rich imputation model.

2.2 Multiple Imputation in a Nutshell

The basic idea of multiple imputation is that each missing value is replaced by not only one, but by $m > 1$ plausible values. Typically between 5 and 20 imputations are created. The resulting m completed data sets are then analyzed separately and the m statistical results are combined into one overall result using Rubin's (1987) formula for MI inference. While the estimated population quantity is simply the mean of all m parameter estimates, standard errors are calculated using the combination of two variance components: the variation between and within the imputed data sets. This combined variation is supposed to reflect the additional uncertainty in parameter estimation due to missing data in an adequate way. Typically multiple imputation produces parameter estimates that are more accurate than estimates from traditional single imputation procedures like unconditional mean imputation or other ad hoc procedures like case deletion, especially when values are not missing completely at random (Little & Rubin, 1987; Schafer & Graham, 2002; Kleinke et al., 2011). The most commonly used frameworks for creating multiple imputations are joint modeling (JM, e.g. Schafer, 1997a,b) and sequential regression multiple imputation (also called MI by chained equations, e.g. Raghunathan et al., 2001; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren, 2007). Introductions to joint modeling may be found in Allison (2001); Graham (2009); Graham, Cumsille, & Elek-Fisk (2003); Schafer (1997a); Schafer & Graham (2002), an introduction to the sequential regression framework is given by van Buuren & Groothuis-Oudshoorn (2011). Joint Modeling has a strong theoretical foundation in Bayesian statistics. However, the procedure makes it necessary to specify a joint probability model for the data set as a whole, which is sometimes unpractical or impossible in real life (Gelman & Raghunathan, 2001). Currently existing software allows to create imputations under the multivariate normal model (`norm`) for approxi-

mately multivariate normal data, the log-linear model for data sets that contain only categorical data (`cat`), the general location model for data sets that contain both continuous and categorical data (`mix`) and under a multivariate linear mixed effects model for clustered or panel data (`pan`) (Schafer, 1997a,b). Note however that the `pan` procedure only supports linear relationships and only two hierarchical levels. Generalized linear mixed effects models are not supported.

A more flexible alternative to JM, especially in cases where no suitable multivariate distribution may be specified is the sequential regressions framework. Instead of trying to find a joint model, each incompletely observed variable is modeled separately.

2.3 Multiple Imputation by Chained Equations

Multiply imputing data with the sequential regressions approach, which is also called fully conditional specification (FCS), or multiple imputation by chained equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011), means that for each incompletely observed variable Y_j in data set Y, a conditional model $P(Y_j|Y_s, \theta_j)$ is specified, with Y_s being a subset of Y containing some or all of the variables in Y except from Y_j. This subset is used to model Y_j and predict missing information in Y_j. Having specified these separate models, imputations of missing values in Y_j are then generated from $P(Y_j|Y_s, \theta_j)$, assuming that values are missing at random in the sense of Rubin (1976). The actual imputation process is a three-step process: First, the posterior distribution of parameters θ is calculated based on the observed data: $P(\theta|Y_{obs})$. Secondly, new parameters θ^* are drawn from $P(\theta|Y_{obs})$. Thirdly, imputations Y^* are generated from $P(Y_{mis}|Y_{obs}, \theta^*)$. Steps 2 and 3 are repeated m times to obtain the m imputations. From a mathematical and computational point of view, the problem is to determine the distributions to draw from. The `mice` software in R (van Buuren & Groothuis-Oudshoorn, 2011), which we use as basis for our count data imputation procedures, approximates this problem by iteratively sampling from the conditional distributions using a Gibbs sampler:

$$
\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1|Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}) \\
Y_1^{*(t)} &\sim P(Y_1|Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p|Y_p^{obs}, Y_1^{(t)}, \ldots, Y_{p-1}^{(t)}) \\
Y_p^{*(t)} &\sim P(Y_p|Y_p^{obs}, Y_1^{(t)}, \ldots, Y_p^{(t)}, \theta_p^{*(t)}).
\end{aligned}
$$

`obs` and `mis` refer to the observed and missing data respectively, t denotes the iteration number, and p stands for the number of variables in the imputation model. Each iteration consists of one cycle through all Y_j. `mice` executes these iterations m times in

parallel to generate the m imputations. One theoretical disadvantage of using chained equations could be that from a theoretical point of view the specified conditional densities can be incompatible and the implicit joint distribution to which the Gibbs sampler attempts to converge may not exist. However, despite this lack of theoretical corroboration, chained equation approaches seem to work well and imputations can be regarded as plausible (van Buuren et al., 2006; Horton & Lipsitz, 2001; Yu, Burton, & Rivero-Arias, 2007). The interested reader may find further implementations of chained equation approaches, applications and discussions in the following references: Brand (1999); Gelman (2004); Heckerman, Chickering, Meek, Rounthwaite, & Kadie (2001); Kennickell (1991); Raghunathan et al. (2001); Rubin (2003); van Buuren (2007); van Buuren, Boshuizen, & Knook (1999).

2.4 The `mice` Software in R

Before we introduce our count data imputation procedures we give the reader some basic information about the `mice` software in R, which is needed to call our functions. Here, we only focus on the information that is absolutely necessary for the reader and practitioner to understand, how our procedures work and how `mice` is used to create the imputations. We assume that the readers are already familiar with the R language. Otherwise, a comprehensive introduction to R may be found in Adler (2010). The `mice` software is described in details in van Buuren & Groothuis-Oudshoorn (2011).

When calling the main function `mice()` to multiply impute incomplete data, the user can specify certain arguments:

```
mice(data, m = 5,
    method = vector("character",length=ncol(data)),
    predictorMatrix = (1 - diag(1, ncol(data))),
    visitSequence = (1:ncol(data))[apply(is.na(data),2,any)],
    post = vector("character", length = ncol(data)),
    defaultMethod = c("pmm","logreg","polyreg"),
    maxit = 5,
    diagnostics = TRUE,
    printFlag = TRUE,
    seed = NA,
    imputationMethod = NULL,
    defaultImputationMethod = NULL
    )
```

The imputation procedure with which to impute missing data in a certain variable is specified via the `method` argument. A not exhaustive overview of imputation methods

that are already implemented in `mice` is given in Table 1. The procedures are all described in detail in van Buuren & Groothuis-Oudshoorn (2011). `method` must be a character vector of length equal to the number of variables in the data set, indicating the missing data procedures with which each variable is to be imputed[1]. Completely observed variables have method `""`, indicating that they need not be imputed. The command `imp<-mice(data,method=c("","norm","pmm","logreg"))` would multiply impute the data stored in the object `data` and store the results in an object called `imp`. The first variable in that data set would not be imputed, the second one would be imputed using Bayesian linear regression (`norm`), the third variable would be filled in by predictive mean matching (`pmm`), and the last one by Bayesian logistic regression (`logreg`). The respective imputation functions are stored internally under the name `mice.impute.name`, where `name` identifies the respective imputation function. Thus specifying `"logreg"` as imputation method for a variable internally calls the function `mice.impute.logreg()`. This is important to know when programming self-written imputation procedures. That these can be called by `mice()`, they have to be called `mice.impute.name`, where the `name` part can be any combination of characters. Our count data imputation functions are called `mice.impute.pois` and `mice.impute.qpois` to impute missing data under the Poisson or Quasi-Poisson model. These functions can be called by setting the respective entry in the `method` vector to `"pois"` or `"qpois"`.

Table 1: Overview of imputation procedures in MICE

Name	Description	Scale
pmm	predictive mean matching	numeric
norm	Bayesian linear regression	numeric
2l.norm	2-level linear mixed effects model	numeric
logreg	Bayesian logistic regression	factor (2 levels)
polyreg	polytomous regression	factor ($>$ 2 levels)
sample	random sample from observed data	any

Selecting the subsets of predictors for each incompletely observed variable is done via the `predictorMatrix` argument. `predictorMatrix` must be a rectangular matrix of dimensions equal to the number of variables in the data set. An example is shown in Table 2. Each row i in that matrix denotes the imputation model of variable V_i. The zeros and ones indicate (0 = no; 1 = yes), if the respective variable V_j is used to predict missing data in V_i. Using the information from the `predictorMatrix`, `mice` automatically creates three objects that are passed on to the `mice.impute.name` sub-

[1]If `method` is not specified the `defaultMethod` is used, which is `pmm`, predictive mean matching for continuous data, `logreg`, i.e. Bayesian logistic regression for factors with two levels and `polyreg`, polytomous logistic regression for factors with more than two levels

function: `y`, `x` and `ry`. `y` is an incomplete data vector of length n, the dependent variable in the imputation regression model and `x` is an $n \times p$ matrix of predictors, those variables that were specified via the respective row in the `predictorMatrix`. `ry` is the response indicator of vector `y`, indicating if a value in `y` has been observed ($\texttt{ry}_i = \texttt{TRUE}$), or not ($\texttt{ry}_i = \texttt{FALSE}$).

Table 2: Specification of imputation models in MICE: The predictor matrix.

	V1	V2	V3	V4	V5	V6	V7
V1	0	1	1	1	1	1	1
V2	1	0	1	1	1	0	1
V3	0	0	0	0	0	0	0
V4	0	0	0	0	0	0	0
V5	0	0	0	0	0	0	0
V6	0	0	0	0	0	0	0
V7	0	0	0	0	0	0	0

Note. Each row i denotes the imputation model for incompletely observed variable V_i in the data set. The zeros and ones indicate, if variable V_j, with $j \in 1, \ldots k$, where k is the number of variables in the data set, is part of the imputation model of V_i (1 = yes, 0 = no).

2.5 Count Data Analysis and Imputation

We present multiple imputation procedures for incomplete count data within a sequential regression (chained equations) MI framework. The imputation procedures are based on a Poisson or Quasi-Poisson regression model. For a comprehensive overview on count data models, see Zeileis, Kleiber, & Jackman (2008). We only give a brief overview: The classical way to analyze count data is to fit a Poisson model under the generalized linear modeling (GLM) framework (Nelder & Wedderburn, 1972): GLMs describe the dependence of a scalar variable y_i on a vector of regressors x_i. The conditional distribution of $y_i|x_i$ is a linear exponential family with probability density

$$f(y; \lambda, \delta) = \exp\left(\frac{y\lambda - b(\lambda)}{\delta} + c(y, \delta)\right),$$

with λ being the canonical parameter that depends on x_i via a linear predictor, δ being a dispersion parameter, and $b(\cdot)$ and $c(\cdot)$ being functions that determine, which member of the family (e.g. Poisson) is used. The mean is determined by $E[y_i|x_i] = \mu_i = b'(\lambda_i)$, the variance by $V[y_i|x_i] = \phi b''(\lambda_i)$. The dependence of $E[y_i|x_i] = \mu_i$ on x_i is specified via a link function $g(\cdot)$. The classical poisson model

$$f(y; \mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

with link function $g(\mu) = \log(\mu)$ assumes that the variance $V(\mu)$ is equal to the mean μ (thus dispersion parameter δ has the value 1).

However, the restriction of equidispersion (that the variance is equal to the mean) is often violated in real life. Very often empirical data are overdispersed, which means that the variance is larger in comparison to the mean. Analyzing overdispersed data using classical Poisson regression leads to an underestimation of the variation in the data and model based tests are thus more liberal (cf. Zeileis et al., 2008). To end up with proper parameter estimates and standard errors, some adjustments need to be made. One popular solution is to estimate dispersion parameter δ from the data rather than fixing it to 1. In the S and R languages for statistical computing this can be done by fitting a `quasipoisson` model, which is identical to an ordinary Poisson model except that δ is estimated from the data (cf. McCullagh & Nelder, 1989).

Our imputation procedures for ordinary and overdispersed count data are based on Poisson and Quasi-Poisson regression. However, we do not impute deterministically, i.e. we do not directly return the fitted value. This would cause all imputations to lie directly on the regression line, which leads to an under-estimation of standard errors. Instead, we follow the stochastic imputation approach described in Rubin (1987). In his Bayesian logistic regression approach, Rubin (1987) fits a logit model to the data and computes parameters $\widehat{\theta}$, the posterior mean and $\widehat{V}(\widehat{\theta})$, the posterior variance of θ. $\widehat{\theta}$ is estimated by maximum likelihood and is defined by

$$\Pi_{i \in obs} f(Y_i|X_i, \widehat{\theta}) \geq \Pi_{i \in obs} f(Y_i|X_i, \theta) \, \forall \, \theta,$$

whereas $\widehat{V}(\widehat{\theta})$ is defined by the negative inverse of the second derivative matrix of the log-posterior distribution at $\theta = \widehat{\theta}$

$$\widehat{V}(\widehat{\theta}) = - \left[\frac{\partial^2}{\partial \theta \partial \theta} \log \Pi_{i \in obs} f(Y_i|X_i, \theta) \Big|_{\theta = \widehat{\theta}} \right]^{-1}.$$

Having calculated $\widehat{\theta}$ and $V(\widehat{\theta})$, new parameters θ^* are drawn from $N(\widehat{\theta}, \widehat{V}(\widehat{\theta}))$. For every missing case the fitted value $p_i = \text{logit}^{-1}(X_i\theta^*)$ is computed. Finally independent uniform (0,1) random numbers u_i are drawn, with $i \in mis$. If $u_i > p_i$ the imputed value is $y_i = 0$, else $y_i = 1$. These steps are repeated m times with new draws of random numbers to end up with m imputations of each missing value.

Analogous, our Poisson imputation approach fits a Poisson model and calculates $\widehat{\theta}$ and $\widehat{V}(\widehat{\theta})$, the posterior mean and the posterior variance of θ respectively. Then, new parameters θ^* are drawn from $N(\widehat{\theta}, \widehat{V}(\widehat{\theta}))$. With these new parameters, predicted scores are computed for each participant with missing values in y: $p_i = \exp(X_i\theta^*)$. Finally, imputations are simulated from $y_i \sim \text{Pois}(p_i)$.

The quasi-poisson imputation approach is quite similar. The only difference to

Poisson imputation is that we use the `quasipoisson` family instead of the `poisson` family and imputations are simulated from $y_i \sim NB(\mu = p_i; \text{size} = p_i/(\delta - 1))$, if the estimate of δ is larger than 1 and $y_i \sim \text{Pois}(p_i)$, if $\delta \leq 1$. The negative binomial (NB) distribution is well suited to simulate overdispersed count data. For further information about the negative binomial distribution and different parametrizations, see Hilbe (2007). If data are equidispersed, imputations can as well be drawn directly from the Poisson distribution. Underdispersion is quite rare in empirical data and not supported by either the Poisson or NB distribution.

We now present two evaluation studies regarding these procedures' ability to produce plausible imputations, unbiased parameter estimates and standard errors (Study 1) and their robustness to MNAR mechanisms (Study 2).

3 Study 1: Estimation Precision of Poisson and Quasi-Poisson imputation

We tested the Poisson and Quasi-Poisson imputation approach in a couple of Monte Carlo simulations. We first simulated incomplete Poisson data following either a MCAR or MAR pattern with 50% missing values in the dependent count variable. Secondly, we simulated incomplete overdispersed count data, again following either a MCAR or MAR pattern with 50% missing values in the dependent variable. We evaluated, how Poisson and Quasi-Poisson imputation were able to cope with incomplete ordinary and overdispersed count data.

3.1 Hypotheses

We assume that multiple Poisson imputation only produces proper imputations when the variable in fact is approximately Poisson distributed. We also hypothesize that when data are overdispersed, imputations by the Poisson method will no longer be adequate and that Quasi-Poisson imputation produces far better results than the Poisson approach.

3.2 Monte Carlo Simulations

In the first simulation, we created 1000 data sets with sample sizes $N = 200$, $N = 500$, and $N = 1000$ respectively, each containing four variables, one dependent variable y, and three predictors x_1, x_2, and x_3. Continuous variables x_1–x_3 were drawn from standard normal distributions, y was drawn from a poisson distribution

$$y_i \sim \text{Pois}(\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}))$$

with parameters $\beta_0 = 1.00$, $\beta_1 = 0.50$, $\beta_2 = -0.75$, and $\beta_3 = .25$.

The design of the second simulation was identical to the first one except that the count variable y was drawn from a negative binomial distribution

$$y_i \sim NB(\mu_i; \text{size} = \frac{\mu_i}{\delta - 1})$$

with means $\mu_i = \exp(1 + 0.50x_{1i} - 0.75x_{2i} + 0.25x_{3i})$ and dispersion parameter $\delta = 2$.

In both simulations we then deleted values. 50% of all data in y were made missing, x_1–x_3 were completely observed. Under the MCAR conditions, cases to receive a missing value were a random sample of all cases. In the MAR conditions, missing values in y were introduced according to the following rule:

$$P(y_i \in mis) = \begin{cases} .2 * .5 = .1 & \text{if} \quad x_{1i} < \bar{x}_1 \\ .8 * .5 = .4 & \text{if} \quad x_{1i} > \bar{x}_1 \end{cases} ,$$

with x_1 being the "cause of missingness" and $\bar{x}_1$ denoting the mean of x_1. Thus the probability of a certain y_i to be missing was .1 if the corresponding x_1 value was below $\bar{x}_1$ and .4, if it was above the mean.

In each simulation, we ended up with 1000 data sets with sample size $N = 200$, 1000 data sets with sample size $N = 500$ and 1000 data sets with sample size $N = 1000$ that followed a MCAR pattern, and with the same kind of data sets that followed the described MAR pattern.

3.3 Missing Data Imputation and Data Analysis

We imputed incomplete Poisson data using our multiple Poisson imputation routine. Incomplete overdispersed count data were imputed using both the Poisson method and the Quasi-Poisson procedure. We hypothesized that in the latter case Quasi-Poisson imputation would yield better results than the Poisson method. We then fitted a generalized linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, using the `quasipoisson` family of distributions with a log link to get an estimation for dispersion parameter δ in addition to the other parameter estimates.

3.4 Quality Criteria

To evaluate the quality of our missing data procedures we followed criteria established by Neyman & Pearson (1933) and Neyman (1937), which were used in previous missing data research (e.g. Schafer & Graham, 2002; Kleinke et al., 2011). Let Q be the population parameter of interest and $\widehat{Q}$ the estimate of Q based on the sampled data and based on the applied missing data procedure. The quality of the missing data procedure

can be deemed good if the difference between the true parameter and the estimated parameter is small. Bias is defined as the difference between Q and the average of $\widehat{Q}$ across the replicated samples and should be close to zero. Not only should the bias in the parameter estimates be as small as possible, also measures of uncertainty need to be accurate. Too large standard errors are undesirable, as the risk of a Type II error increases. Underestimation of standard errors is also a serious problem, as confidence intervals might be too narrow to include the true parameter. Coverage rate is defined as the percentage of confidence intervals that cover the true parameter. Significant bias in calculations of standard errors shows up as a reduction in coverage rates. Coverage rate in fact is a useful quantity that indicates both bias in parameter estimation and bias in measures of uncertainty. Significant bias in parameter estimation might also indicate a decrease of coverage, because the intervals would be too far to the left or too far to the right to cover the true parameter. A missing data procedure is good if the 95% confidence intervals cover the true parameter with probability close to 95%. Lower coverage rates increase the probability of Type I error. We deem coverage below 90% as seriously low, as it corresponds to a doubling of the nominal error rate (cf. Schafer & Graham, 2002).

In sum, bias should be close to zero and standard errors should be reasonably small with corresponding coverage above 90%.

4 Results

Results regarding multiple Poisson imputation are displayed in Table 3. As can be seen, coverage is always close to 95% and parameter estimates are close to the "true" population values regardless of the sample size and regardless of the missing data mechanism (MCAR or MAR).

However, if the data are overdispersed, using multiple Poisson imputation yields suboptimal estimates, as can be seen in Table 4. Regression coefficients are still being estimated quite well with somewhat larger standard errors in comparison to the first simulation. Consistency in parameter estimation was not so good, as most coverage rates were below 90%, however still in the high eighties. Note that dispersion parameter δ was severely underestimated by around 25%. Again, results are independent from sample size and the simulated missing data mechanism. Quasi-Poisson imputation was supposed to produce better results with overdispersed data, and this is supported by our simulation results. Results regarding Quasi-Poisson imputation are displayed in Table 5. As can be seen, bias in parameter estimation is always negligible with corresponding coverage rates well over 90%. Dispersion is also always estimated well.

5 Study 2: Robustness to MNAR Processes

5.1 Overview and Hypotheses

The purpose of the second study was to evaluate the algorithms' robustness to missing not at random processes. Like in Study 1, we looked at two scenarios, one with Poisson data and one with NB data. Typically, MI procedures were designed for MAR mechanism, but are believed to be – at least to some extent – robust to violations of the MAR assumption (e.g. Schafer, 1997a). As bias in parameter estimation both depends on the missing data mechanism and the amount of missing data, we assumed that our algorithms can cope with MNAR mechanisms if not too many data are missing.

Secondly, inclusion of strong auxiliary variables may buffer bias introduced by MNAR mechanisms as they make the mechanism a little bit more MAR (Collins et al., 2001). Thus we assumed that having strong auxiliary variables in the data set will buffer the ill effects of MNAR mechanisms more strongly than having only weak predictors in the data set. Thus, having a large quantity of missing data, a strong MNAR mechanism, and only few or weak auxiliary variables was assumed to be the worst case and assumed to lead to severe biases.

5.2 Method

In the first scenario we created a Poisson distributed dependent variable y and three normally distributed predictors x_1–x_3. We varied the sample size $N = 200$, $N = 500$ and $N = 1000$ and the percentage of missing values p_{mis} in y respectively, with $p_{mis} = 5\%$, $p_{mis} = 10\%$, $p_{mis} = 20\%$ and $p_{mis} = 30\%$. Population parameters were set to $\beta_0 = 1.00$, $\beta_1 = 0.50$, $\beta_2 = -0.75$, and $\beta_3 = .25$.

To simulate a MNAR mechanism we chose y itself as the "cause of missingness" Z. The simulated MNAR mechanism was quite strong with missingness probabilities

$$P(y_i \in mis) = \begin{cases} p_{mis} & \text{if} \quad y_i < \bar{y} \\ 0 & \text{if} \quad y_i > \bar{y} \end{cases} .$$

Again, like in Study 1, we simulated 1000 data sets in each condition. The incomplete data sets were multiply imputed and generalized linear models using the `quasipoisson` family were fitted.

The second scenario was identical to the first one, except that we simulated overdispersion and drew y from a negative binomial distribution:

$$y_i \sim NB(\mu_i = \exp(1 + 0.15x_{1i} - 0.20x_{2i} + 0.25x_{3i}), \text{size} = \frac{\mu_i}{\delta - 1}),$$

with $\delta = 2$. We also chose somewhat weaker predictors in comparison to the first sce-

nario. Predictors that are strongly related to y and / or that are related to missingness may buffer bias due to MNAR processes (Collins et al., 2001). The question is, how robust our algorithms are, if only variables with weak or moderate relations to y are present in the data set.

5.3 Results

Results of the first scenario are summarized in Table 6, results regarding the second scenario are displayed in Table 7. In scenario 1, with only 5% missing values, parameter estimates were unbiased and coverage rates lay well within the acceptable range regardless of the sample size. With 10% missing values, we observed about 2% bias in estimations of the intercepts. Intercepts were estimated higher than they actually are. With increasing sample sizes ($N = 500$ and $N = 1000$) coverage regarding intercept estimates fell below the 90% line. The lowest coverage rate was 84.4%. With 20% missing data in y, bias regarding intercepts increased to 4% – 5% and further climbed up to 8%, when 30% of cases in y were unobserved. Depending on the sample size, coverage ranged from 84.6% – 55.2%, when 20% of values were missing and between 72.7% and 21.4%, when 30% values were missing. This is clearly an unacceptable coverage rate. Note that when more than 20% values in y were unobserved, also some regression coefficients of the completely observed predictors were wrongly estimated.

With regard to the first scenario, we can conclude that with small amounts of missing data, bias due to MNAR processes is negligible and our proposed procedures might be used. If up to 10% cases are missing and the underlying MNAR mechanism is really strong, some but still rather small bias is likely. If much more than 10% of cases are missing, rather severe mis-estimation may occur. In that case, we would urge researchers to conduct a thorough MNAR sensitivity analysis (for details, see van Buuren & Groothuis-Oudshoorn, 2011) and discuss statistical results with utter caution.

We now turn to the results of the second scenario. With only 5% missing values in y and weaker predictors in comparison to the first scenario, we observed about 3% bias in estimations of intercepts regardless of the sample size. However, standard errors decreased with increasing sample size, which lead to quite noticeable undercoverage of 80.3% under a sample size of $N = 1000$. Dispersion parameter δ was estimated quite okay with only -1.5% bias. Things looked quite differently, when the percentage of missing values increased: With 10% missing cases, we observed 6% bias in estimations of intercepts and this rapidly increased to 13%–14% under 20% missings and 21%–23% with 30% missings, which is unacceptably large. The corresponding coverage rates ranged from 87.6%–40.9%, depending on the sample size, with 10% missings and between 50% and 0.4% with 20% missings and between 12.3% and 0% with 30%

missings, which implies that this missing data procedure does not work under theses conditions. Analogous, mis-estimation of δ was negligible with 5% missings (Bias = -1.5%), but got more severe, the more values were unobserved: Bias rapidly increased from -2.5% to -6.5% and finally to -12.5% under 10%, 20% and 30% missing values respectively.

We conclude, that our missing data procedures might be used under MNAR, when the missing data problem is only minor (around 5%). If more values are missing, the missing data mechanism is really strong and only weak predictors are present in the data set, rather severe biases are to be expected.

6 Discussion

We have tested two multiple imputation procedures for incomplete count data that work as an add-on to the `mice` software in R: Multiple Poisson imputation for variables that are approximately Poisson distributed, and multiple Quasi-Poisson imputation for overdispersed count data. Our simulations have shown that incompletely observed variables that in fact follow the assumed distributions, can very well be imputed with our procedures.

However, empirical count data do more or less deviate from these theoretical models and we would assume that the more empirical count data deviate from these statistical convenient distributions, the more inadequate imputations will become and the more bias is to be expected. Our algorithms are expected to work well, if the model fit is good. If the Poisson or Quasi-Poisson model as an imputation model for count data is a misspecified imputation model, then the imputations can be expected to be not proper anymore and the inferences of scientific interest may be biased. A semi- or non-parametric imputation approach could be a solution in this case. Further simulations need to clarify, to what extent our procedures are robust in that regard.

Another point that needs to be addressed in the near future is the problem of incomplete predictors. In our simulations we had a complete set of predictors. This is seldom the case in empirical data. `mice` handles this problem by repeatedly cycling through each incompletely observed variable, imputing it and then using the filled-in information to estimate the next variable and so on. Imputation models are specified on a variable to variable basis and imputation results are expected to be good, if theoretically and mathematically sound models are specified and good predictors have been found (cf. Collins et al., 2001). Nevertheless, if there is a huge quantity of missing data in the predictors as well, there is not much substance to estimate missing information and bias is to be expected. A fruitful avenue for future research is to see, if semi- or nonparametric procedures are more robust in that regard. Generally, future simula-

tions should show how much information must be present in the data – or how much missing information may be allowed to still obtain reasonable and justifiable parameter estimates.

Furture evaluations should also look at more complex models with higher order relationships and interactions.

A further problem that often occurs when analyzing count data are behavioral events that seldom occur. This leads to distributions with large amounts of zeros. Such zero-inflated data are typically analyzed with special models like the zero-inflated Poisson model for example (Lambert, 1992). The procedures presented here cannot handle excess zeros. This problem, as well as imputation of multilevel count data is addressed by other imputation procedures in the `countimp` package (Kleinke & Reinecke, 2011).

References

Adler, J. (2010). *R in a nutshell*. Beijing: O'Reilly.

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Doctoral Dissertation, Erasmus University, Rotterdam, The Netherlands.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. *Psychological Methods*, *6*, 330–351.

Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Enders, C. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, *99*(466), 537–545.

Gelman, A., & Raghunathan, T. E. (2001). [Conditionally Specified Distributions: An Introduction]: Comment. *Statistical Science*, *16*(3), pp. 268-269.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Volume 2. Research methods in psychology* (pp. 87–114). Hoboken, NJ: Wiley & Sons.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., & Kadie, C. (2001). Dependency networks for inference, collaborative filtering, and data visualization. *The Journal of Machine Learning Research*, *1*, 49–75.

Hedeker, D., & Mermelstein, R. J. (2011). Multilevel analysis of ordinal outcomes related to survival data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 115–136). New York, NY: Taylor & Francis.

Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge: Cambridge University Press.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, *55*, 244–254.

Kennickell, A. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. In American Statistical Association (Ed.), *Proceedings of the survey research methods section,* (pp. 1–10). Alexandria, VA: American Statistical Association.

Kleinke, K., & Reinecke, J. (2011). *COUNTIMP – A multiple imputation package for incomplete count data* (Technical Report). Bielefeld: Bielefeld University, Faculty of Sociology.

Kleinke, K., Stemmler, M., Reinecke, J., & Lösel, F. (2011). Efficient ways to impute incomplete panel data. *Advances in Statistical Analysis*, *95*(4), 351–373.

Lambert, D. (1992). Zero-inflated poisson regression. With an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, *135*, 370–384.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A*, *236*, 333–380.

Neyman, J., & Pearson, E. S. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, *57*(1), 3–18.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. (1997b). *Imputation of missing covariates under a general linear mixed model* (Technical Report 97-10). University Park: Pennsylvania State University, The Methodology Center.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2009). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, *20*(1), 1–27.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York, NY: Taylor & Francis.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*(6), 681–694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & 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. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

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.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1–25.

Table 3: Performance of multiple Poisson imputation with Poisson distributed data

		MCAR					MAR				
		β_0	β_1	β_2	β_3	δ	β_0	β_1	β_2	β_3	δ
$N = 200$	$\widehat{Q}$	0.99	0.50	-0.75	0.25	1.00	0.99	0.50	-0.75	0.25	1.00
	SE	0.07	0.05	0.06	0.05		0.08	0.06	0.06	0.06	
	CR	93.90	95.00	94.30	96.00		94.60	93.80	93.80	95.30	
$N = 500$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.00	0.99	0.50	-0.75	0.25	1.01
	SE	0.04	0.03	0.03	0.03		0.08	0.06	0.06	0.06	
	CR	96.10	94.00	95.10	95.60		94.70	95.50	94.60	95.30	
$N = 1000$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.00	0.99	0.50	-0.75	0.24	1.00
	SE	0.03	0.02	0.02	0.02		0.08	0.07	0.06	0.06	
	CR	94.90	95.10	92.80	93.80		94.30	93.00	96.40	94.90	

Note. Monte Carlo Simulations with sample sizes $N = 200$, $N = 500$ and $N = 1000$ with 1000 replications respectively. The population parameters were set as: $y = 1 + 0.50x_1 - 0.75x_2 + 0.25x_3 + \varepsilon$. $\delta = 1$.

δ: dispersion parameter

$\widehat{Q}$: estimated average population quantity across the 1000 replications

SE: average standard error

CR: coverage rate (Use of boldface type indicates low coverage, i.e. CR $< 90\%$)

Table 4: Performance of multiple Poisson imputation with overdispersed Negative Binomial distributed data

		MCAR					MAR				
		β_0	β_1	β_2	β_3	δ	β_0	β_1	β_2	β_3	δ
$N = 200$	$\widehat{Q}$	0.99	0.51	-0.75	0.25	1.49	0.99	0.50	-0.75	0.25	1.48
	SE	0.08	0.06	0.06	0.06		0.08	0.07	0.07	0.07	
	CR	**88.70**	**89.10**	**87.10**	90.20		**88.80**	**86.90**	**89.30**	**88.60**	
$N = 500$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.49	0.99	0.49	-0.75	0.25	1.48
	SE	0.05	0.04	0.04	0.04		0.08	0.07	0.07	0.07	
	CR	**88.60**	**87.80**	**87.80**	**88.50**		90.10	**87.70**	**88.80**	**88.70**	
$N = 1000$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.50	0.99	0.50	-0.75	0.25	1.48
	SE	0.03	0.03	0.03	0.03		0.08	0.07	0.07	0.07	
	CR	**87.00**	**88.90**	**88.10**	**88.20**		**89.10**	**88.90**	**88.60**	**88.20**	

Note. Monte Carlo Simulations with sample sizes $N = 200$, $N = 500$ and $N = 1000$ with 1000 replications respectively. The population parameters were set as: $y = 1 + 0.50x_1 - 0.75x_2 + 0.25x_3 + \varepsilon$. $\delta = 2$.

δ: dispersion parameter

$\widehat{Q}$: estimated average population quantity across the 1000 replications

SE: average standard error

CR: coverage rate (Use of boldface type indicates low coverage, i.e. CR $< 90\%$)

Table 5: Performance of multiple Quasi-Poisson imputation with overdispersed Negative Binomial distributed data

		MCAR					MAR				
		β_0	β_1	β_2	β_3	δ	β_0	β_1	β_2	β_3	δ
$N = 200$	$\widehat{Q}$	0.99	0.51	-0.75	0.25	1.97	0.98	0.50	-0.75	0.25	1.95
	SE	0.09	0.07	0.07	0.07		0.10	0.08	0.08	0.08	
	CR	93.50	92.10	91.80	93.20		93.30	91.20	91.00	91.00	
$N = 500$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.99	0.98	0.50	-0.75	0.25	1.95
	SE	0.06	0.04	0.04	0.04		0.10	0.08	0.08	0.08	
	CR	92.20	94.00	93.00	93.20		93.70	91.40	92.30	92.60	
$N = 1000$	$\widehat{Q}$	1.00	0.50	-0.75	0.25	1.99	0.99	0.50	-0.75	0.25	1.96
	SE	0.04	0.03	0.03	0.03		0.10	0.08	0.08	0.08	
	CR	93.00	93.90	92.20	93.50		91.60	91.60	91.60	92.10	

Note. Monte Carlo Simulations with sample sizes $N = 200$, $N = 500$ and $N = 1000$ with 1000 replications respectively. The population parameters were set as: $y = 1 + 0.50x_1 - 0.75x_2 + 0.25x_3 + \varepsilon$. $\delta = 2$.

δ: dispersion parameter

$\widehat{Q}$: estimated average population quantity across the 1000 replications

SE: average standard error

CR: coverage rate (Use of boldface type indicates low coverage, i.e. CR $< 90\%$)

Table 6: Performance of multiple Poisson imputation when the data are missing not at random

		β_0	β_1	β_2	β_3	δ	β_0	β_1	β_2	β_3	δ
		5% NAs					20% NAs				
$N = 200$	$\widehat{Q}$	1.01	0.50	-0.75	0.25	1.00	1.04	0.49	-0.74	0.24	1.00
	SE	0.05	0.04	0.04	0.04		0.05	0.04	0.04	0.04	
	CR	95.10	94.20	95.30	96.20		**84.60**	93.80	93.40	96.40	
$N = 500$	$\widehat{Q}$	1.01	0.50	-0.75	0.25	1.00	1.05	0.49	-0.73	0.25	1.00
	SE	0.03	0.02	0.02	0.02		0.03	0.02	0.02	0.02	
	CR	92.80	94.70	94.50	94.70		**71.50**	92.90	**89.50**	94.10	
$N = 1000$	$\widehat{Q}$	1.01	0.50	-0.75	0.25	1.00	1.04	0.49	-0.74	0.25	0.99
	SE	0.02	0.02	0.02	0.02		0.02	0.02	0.02	0.02	
	CR	92.00	94.50	95.00	94.80		**55.30**	90.80	**85.30**	94.40	
		10% NAs					30% NAs				
$N = 200$	$\widehat{Q}$	1.02	0.50	-0.74	0.25	1.00	1.08	0.48	-0.72	0.24	0.99
	SE	0.05	0.04	0.04	0.04		0.06	0.04	0.04	0.04	
	CR	92.90	94.80	94.10	95.40		**72.70**	92.80	**87.50**	94.80	
$N = 500$	$\widehat{Q}$	1.02	0.49	-0.74	0.25	1.00	1.08	0.48	-0.72	0.24	0.99
	SE	0.03	0.02	0.02	0.02		0.04	0.02	0.02	0.02	
	CR	**88.80**	94.00	94.40	94.60		**44.50**	**88.80**	**79.60**	93.30	
$N = 1000$	$\widehat{Q}$	1.02	0.50	-0.74	0.25	1.00	1.08	0.48	-0.72	0.24	0.99
	SE	0.02	0.02	0.02	0.02		0.03	0.02	0.02	0.02	
	CR	**84.40**	94.80	93.30	94.00		**21.40**	**83.70**	**70.00**	93.10	

Note. Monte Carlo Simulations with sample sizes $N = 200$, $N = 500$ and $N = 1000$ with 1000 replications respectively. The population parameters were set as: $y = 1 + 0.50x_1 - 0.75x_2 + 0.25x_3 + \varepsilon$. $\delta = 1$.

δ: dispersion parameter

$\widehat{Q}$: estimated average population quantity across the 1000 replications

SE: average standard error

CR: coverage rate (Use of boldface type indicates low coverage, i.e. CR $< 90\%$)

Table 7: Performance of multiple Quasi-Poisson imputation when the data are missing not at random and predictors are "weak"

		β_0	β_1	β_2	β_3	δ	β_0	β_1	β_2	β_3	δ
		5% NAs					20% NAs				
$N = 200$	$\widehat{Q}$	1.03	0.14	-0.20	0.25	1.97	1.13	0.14	-0.18	0.23	1.89
	SE	0.06	0.06	0.06	0.06		0.06	0.06	0.06	0.06	
	CR	93.30	95.00	95.20	95.20		**50.00**	94.00	93.20	92.90	
$N = 500$	$\widehat{Q}$	1.03	0.15	-0.20	0.25	1.97	1.13	0.13	-0.18	0.23	1.87
	SE	0.04	0.04	0.04	0.04		0.04	0.04	0.04	0.04	
	CR	90.50	96.00	94.30	94.00		**12.70**	93.30	91.60	**89.50**	
$N = 1000$	$\widehat{Q}$	1.03	0.15	-0.20	0.24	1.97	1.14	0.13	-0.18	0.22	1.87
	SE	0.03	0.03	0.03	0.03		0.03	0.03	0.03	0.03	
	CR	**80.30**	94.40	94.50	94.70		**0.40**	91.00	**86.30**	**82.40**	
		10% NAs					30% NAs				
$N = 200$	$\widehat{Q}$	1.06	0.14	-0.19	0.24	1.95	1.21	0.12	-0.17	0.21	1.77
	SE	0.06	0.06	0.06	0.06		0.06	0.06	0.06	0.06	
	CR	**87.60**	95.00	94.90	94.90		**12.30**	91.80	**89.60**	**88.30**	
$N = 500$	$\widehat{Q}$	1.06	0.14	-0.19	0.24	1.95	1.22	0.12	-0.17	0.21	1.75
	SE	0.04	0.04	0.04	0.04		0.04	0.04	0.04	0.04	
	CR	**70.60**	95.50	94.30	94.70		**0.50**	**88.60**	**84.30**	**78.30**	
$N = 1000$	$\widehat{Q}$	1.06	0.14	-0.19	0.24	1.95	1.23	0.12	-0.16	0.20	1.75
	SE	0.03	0.03	0.03	0.03		0.03	0.03	0.03	0.03	
	CR	**40.90**	94.40	94.70	93.60		**0.00**	**81.40**	**68.90**	**58.50**	

Note. Monte Carlo Simulations with sample sizes $N = 200$, $N = 500$ and $N = 1000$ with 1000 replications respectively. The population parameters were set as: $y = 1 + 0.15x_1 - 0.20x_2 + 0.25x_3 + \varepsilon$. $\delta = 2$.

δ: dispersion parameter

$\widehat{Q}$: estimated average population quantity across the 1000 replications

SE: average standard error

CR: coverage rate (Use of boldface type indicates low coverage, i.e. CR $< 90\%$)