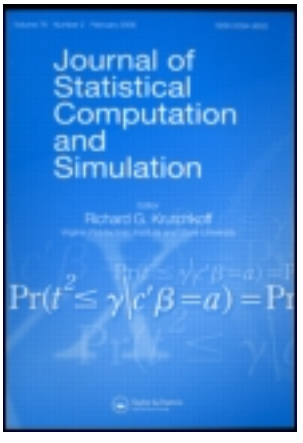

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Examining the robustness of fully synthetic data techniques for data with binary variables

Gregory J. Matthews [a], Ofer Harel [a] & Robert H. Aseltine [b]
[a] Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT, 06269-4120, United States
[b] University of Connecticut Health Center, 99 Ash Street, East Hartford, CT, 06108, United States
Published online: 05 Mar 2009.

To cite this article: Gregory J. Matthews, Ofer Harel & Robert H. Aseltine (2010) Examining the robustness of fully synthetic data techniques for data with binary variables, Journal of Statistical Computation and Simulation, 80:6, 609-624, DOI: 10.1080/00949650902744438

To link to this article: http://dx.doi.org/10.1080/00949650902744438

Journal of Statistical Computation and Simulation
Vol. 80, No. 6, June 2010, 609–624

Examining the robustness of fully synthetic data techniques for data with binary variables

Gregory J. Matthews[a], Ofer Harel[a]* and Robert H. Aseltine[b]

[a]*Department of Statistics, University of Connecticut , 215 Glenbrook Road, Unit 4120, Storrs, CT, 06269-4120, United States;* [b]*University of Connecticut Health Center, 99 Ash Street, East Hartford, CT, 06108, United States*

(*Received 18 July 2008; final version received 13 January 2009*)

There is a growing demand for public use data while at the same time there are increasing concerns about the privacy of personal information. One proposed method for accomplishing both goals is to release data sets that do not contain real values but yield the same inferences as the actual data. The idea is to view confidential data as missing and use multiple imputation techniques to create synthetic data sets. In this article, we compare techniques for creating synthetic data sets in simple scenarios with a binary variable.

Keywords: synthetic data; disclosure limitation; confidentiality; multiple imputation

AMS Subject Classification: 65C60

1. Introduction

There is an increasing demand for access to economic, medical, educational, and human services data while at the same time there are growing concerns about confidentiality for the individuals whose information is being collected. Often it is unethical to release private information to the public, and for certain medical and educational data, it is illegal. Currently, in many situations, researchers have to go through a lengthy process to gain access to such data and, even after being approved, are only granted access to limited data sets or data that has been perturbed in some way. Although this may protect individuals' privacy and confidentiality, it may severely limit the utility of the information. Somehow, a balance must be struck between the release of data for research purposes and the risk of disclosing private information.

When agencies want to release sensitive data to the public for research, variables that would identify an individual must be removed. However, simply removing obvious individual identifiers, like name or home address, still poses privacy risks. For example, Sweeney [1] used public voting records to identify individuals in publicly available health data released by the Massachusetts Group Insurance Commission that was believed to be anonymous. Thus, agencies that release public use data need to go further than simply removing obvious identifiers. To combat disclosure

*Corresponding author. Email: oharel@stat.uconn.edu

ISSN 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/00949650902744438
http://www.informaworld.com

risks agencies often employ simple methods for protecting sensitive data, including truncation [2], data swapping [3], and row deletion [4]. One major drawback to these methods is that, while they may achieve a higher level of confidentiality, much useful information is lost, diminishing the overall utility of this data.

Other methods for preserving confidentiality using masking procedures, including matrix masking and the addition of white noise are discussed in Cox [5,6] and Fienberg *et al.* [7]. However, in order to analyse masked data, knowledge of the masking technique must be known and specific statistical software may have to be used, complicating the analysis for the average user. Analysis of masked data is discussed in Duncan and Lambert [8], Keller and Bethlehem [9], and, for analysing the addition of white noise, Little [10].

Another proposed method for preserving confidentiality involves synthetic data sets. Synthetic data sets replace the actual data with imputed values but still yield the same statistical inference as the actual data set. The idea, first proposed by Rubin [11], is to create a synthetic data set by viewing data that are sensitive or confidential as missing and replacing these values using standard multiple imputation techniques [12–14]. In partially synthetic data, only values considered to be at high risk of disclosure are replaced with imputed values, while the rest of the data remain intact. Partially synthetic data methods are discussed in Kennickell [15], Abowd and Woodcock [16], Liu and Little [17], and Reiter [18,19]. In fully synthetic data sets [11,20–23], which are discussed in this article, all data are replaced with imputed values . In practice, multiple synthetic data sets would be created and released to the public, and valid inference can be obtained using the combining rules set forth in Raghunathan *et al.* [24].

Raghunathan *et al.* [24] showed that fully synthetic data generated from a multivariate normal (MVM) model when, in fact, it is the true model that maintains a high level of utility. Bernaards *et al.* [25] showed that imputing binary variables under the assumption of multivariate normality combined with rounding techniques performs well. By combining the work from Raghunathan *et al.* [24] and Bernaards *et al.* [25], we tested how well different methods of creating synthetic data performed in some simple cases where a binary variable is present in the data.

We created synthetic data using two general methods. In the first, we assumed the data follow a MVN distribution for use in creating synthetic data sets along with the suggestions for rounding the binary variable set forth in Bernaards *et al.* [25]. They suggested three rounding techniques, including simple rounding (SR) and adaptive rounding (AR), for returning the imputed values of the binary variable to binary values.

The second approach we used to create synthetic data sets was a fully conditional specification (FCS) [26] in which a univariate model was built for each variable conditional on the other variables in the data set. By iteratively sampling from these univariate conditional distributions, synthetic data sets were created by approximating draws from the posterior predictive distribution. When necessary, imputed values of the binary variable were rounded as suggested in Bernaards *et al.* [25].

In Section 2, we review the main methods for synthetic data sets. Section 3 describes some simulation studies and Section 4 contains the results of these simulations. In Section 5, we conclude with discussion and implications of the results.

2. Synthetic data

An actual data set, $\boldsymbol{D}$, is a sample of size n from a finite population $\boldsymbol{P}$ of size N. We assume, for simplicity, that there were no missing data in our population. In practice, we almost never know the true population, and we base our inferences on the sample $\boldsymbol{D}$. Using the random sample $\boldsymbol{D}$ from $\boldsymbol{P}$, M synthetic populations are created $\boldsymbol{P}^{(l)}$, $l = (1, 2, ..., M)$ by drawing from the posterior predictive distribution, $\Pr(\boldsymbol{P}^{(l)}|\boldsymbol{D})$.

Often the size of the population is too large to be released, and a simple random sample of size k is drawn from each of these synthetic populations. This collection of M simple random samples becomes the released synthetic data $\boldsymbol{D}_{\text{Syn}} = \boldsymbol{D}^{(l)}, l = 1, 2, \ldots, M$. It should be noted that instead of releasing $\boldsymbol{D}_{\text{Syn}}$, one could just release the distribution from which it was created and achieve the same level of privacy. However, it is almost always simpler for the end user to analyse the draws from the distribution, rather than the distribution itself.

From the synthetic data sets the analyst wants to make inferences about some quantity of interest, $\mathbf{Q}$ (e.g. regression coefficients). $\mathbf{Q}$ can be estimated by a point estimate $\boldsymbol{q}$ with an associated measure of uncertainty $\boldsymbol{v}$. For each synthetic data set (1,2,. . .,M), $\boldsymbol{q}$ and $\boldsymbol{v}$ can be estimated, yielding M estimates $(\boldsymbol{q}^{(l)}, \boldsymbol{v}^{(l)}), l = (1, 2, \ldots, M)$.

As described in Raghunathan *et al.* [24], the population parameter, $\mathbf{Q}|\boldsymbol{D}_{\text{Syn}}$ can be approximated by a normal distribution with mean

$$\bar{q}_M = \sum_l q^{(l)}/M$$

and variance

$$T_M = (1 + M^{-1})b_M - \bar{v}_M$$

where $\bar{v}_M = \sum_l v^{(l)}/M$ and $b_M = \sum_l (q^{(l)} - \bar{q}_M)^2/(M - 1)$.

This approximation for variance of $\mathbf{Q}$ is suitable for large M and simplifies computation. The exact variance can be obtained using numerical procedures and evaluating the integrals described in Raghunathan *et al.* [24]. While this approximation may be simplifying, one drawback to its use is the possibility of negative variance estimates. In spite of this drawback, the variance approximation is used here because of the complexity of calculating T_M exactly. Often the occurrence of negative estimates is reduced by increasing M, the number of synthetic data sets. In this article, negative estimates of variance are dealt with in two ways, either by replacing negative estimates of T_M with $\bar{v}_M$ as suggested by Reiter [23] or by simply dropping the observations with negative variance estimates and reporting the number of simulated samples that require this drop. These methods of dealing with negative estimates are compared in the results section.

3. Simulation studies

Simulations were conducted to compare the inferences garnered from the synthetic data sets and the actual population parameters. Each method of creating synthetic data was then evaluated based on how closely the synthetic data estimated the true population parameters.

Synthetic data sets were created using two methods. First, synthetic data sets were created by independently drawing from the posterior predictive distribution under the assumption of multivariate normality. This process was implemented using R [27]. Binary variables were rounded to 0 or 1 using either SR or AR, as described in Bernaards *et al.* [25]. A second method of creating synthetic data sets used a FCS. This method was implemented using the R package MICE [28]. If warranted, binary variables are rounded to 0 or 1 based on suggestions from Bernaards *et al.* [25].

A population of size $N = 1000$ was created with all observations from the population being random draws from a MVN population. 500 observations were drawn from a MVN distribution with mean vector $\mu_1 = (35\ 5\ 5)^{\text{T}}$, another 500 observations from a MVN distribution with mean vector $\mu_2 = (25\ 5\ 5)^{\text{T}}$ and both having covariance matrix Σ. Two different matrices were considered for Σ. The first covariance matrix, Σ_1, had variances of 4, 2, and 2 for the first, second, and third variable respectively, with a common covariance of 1.5, while the second covariance matrix considered was $\Sigma_2 = 10 * \Sigma_1$.

An indicator specifying which mean vector was used in the generation of the observation was added to the data. The indicator was 1 if drawn from the distribution with mean vector μ_1 and 0 if drawn from the distribution with mean vector μ_2. In this simple setup, the indicator signals which subjects are in the treatment group and which are in the control.

From the population, simple random samples were taken of size $n = 100$. For each sample, M synthetic data populations were created of size $N_{\text{Syn}} = 1000$. From the M synthetic data populations, a simple random sample of size $k = 100$ was drawn from each synthetic population. These M simple random samples, $\mathbf{D}_{\text{Syn}} = \mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(M)}$, are the released data sets. This process was simulated 1000 times. Simulations were run using $M = 10$ and $M = 100$ for each method of creating synthetic data, under each covariance structure.

One common analysis performed by researchers is regression analysis, thus, we evaluated each of these methods based on how well the synthetic data produced by each method estimated the actual regression coefficients. The model of interest here was the first variable regressed against the remaining three.

3.1. *Multivariate normality*

In the first method, fully synthetic data sets were created by sampling from the joint posterior predictive distribution of $\theta = (\mu, \Sigma)$. A model was fit assuming that the data follow a MVN

Table 1. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 10$ and Σ_1. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.88	29.89	30.32	30.67	28.15
	variance	1.85	1.84	0.47	1.82	35.97
	bias	−0.74	−0.74	−0.30	0.05	−2.47
	std bias	−73.61	−72.57	−58.44	45.20	−44.04
	MSE	1.78	1.75	0.77	0.90	10.00
	length	4.98	4.97	2.45	4.93	22.37
	coverage (%)	87.70	87.20	76.80	95.40	99.70
β_1	estimate	0.35	0.35	0.51	0.25	0.51
	variance	0.16	0.16	0.06	0.14	1.63
	bias	−0.07	−0.07	0.09	−0.17	0.09
	std bias	−17.71	−21.38	58.17	−80.36	8.49
	MSE	0.11	0.11	0.05	0.13	0.21
	length	1.47	1.47	0.86	1.35	4.75
	coverage (%)	91.70	91.70	87.90	88.00	1.00
β_2	estimate	0.46	0.46	0.39	0.41	−0.18
	variance	0.14	0.14	0.05	0.13	1.14
	bias	0.03	0.03	−0.04	−0.02	−0.60
	std bias	10.17	12.67	−26.16	−4.68	−62.45
	MSE	0.09	0.09	0.02	0.08	0.53
	length	1.38	1.38	0.82	1.31	3.97
	coverage (%)	92.60	92.50	89.20	93.90	98.70
β_3	estimate	−7.87	−7.88	−9.62	−7.94	0.08
	variance	0.56	0.56	0.17	0.53	48.18
	bias	1.87	1.85	0.11	1.80	9.81
	std bias	351.94	392.69	38.81	339.64	150.44
	MSE	3.64	3.59	0.14	3.38	99.94
	length	2.73	2.73	1.50	2.67	26.72
	coverage (%)	22.60	22.60	88.80	22.80	85.30

distribution with unknown mean vector, μ, and unknown covariance matrix, Σ. A noninformative prior $p(\mu, \Sigma) \propto |\Sigma|^{-1/2}$ was used.

Synthetic data sets were then created by using the steps described in Raghunathan *et al.* [24]. To sample from the posterior predictive distribution, we generated a random draw from a random variable, W, with a Wishart distribution with $n-1$ degrees of freedom and scale matrix $[1/(n-1)]S^{-1}$, where S is the sample covariance matrix. Let $\Sigma^\star = W^{-1}$. Next, a random vector, $\mu^\star$ was generated from a MVN with mean $\bar{y}$, the sample mean, and covariance $\Sigma^\star/n$. Next, we generated N_{syn} random vectors from the MVN with mean $\mu^\star$ and covariance matrix $\Sigma^\star$. This process was repeated M times, and we generated M synthetic populations of size N_{syn} and obtained a simple random sample of size k from each to create M synthetic populations.

After drawing from the posterior predictive, the imputed values of the indicator variable were rounded to either 0 or 1 based on SR or AR as described in Bernaards *et al.* [25]. In either case, values of the imputed indicator variable that are less than 0 are set to 0 and values greater than 1 are set to 1. Using SR, imputed values are rounded to either 0 or 1 based on which is closer. AR, suggested in Bernaards *et al.* [25], is similar to SR, but uses a different threshold. Based on the normal approximation to the binomial distribution, the threshold for AR is $\bar{\omega} - \Phi^{-1}(\bar{\omega})\sqrt{\bar{\omega}(1-\bar{\omega})}$, where $\bar{\omega}$ is the mean value of a single imputed binary variable in the data and Φ^{-1} is the quantile function of the normal distribution. Values larger than the threshold were rounded to 1 with values not exceeding the threshold rounded down to 0.

Table 2. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 100$ and Σ_1. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.87	29.87	30.41	30.72	28.26
	variance	1.10	1.08	0.34	1.10	36.33
	bias	−0.76	−0.75	−0.21	0.10	−2.12
	std bias	−86.18	−84.10	−37.41	10.97	−162.24
	MSE	1.39	1.37	0.64	0.47	5.85
	length	4.01	3.96	2.17	3.93	21.18
	coverage (%)	88.10	88.10	79.40	97.60	100.00
β_1	estimate	0.35	0.35	0.49	0.26	0.50
	variance	0.09	0.09	0.05	0.08	1.42
	bias	−0.07	−0.07	0.07	−0.16	0.08
	std bias	−26.84	−28.32	33.11	−71.58	23.57
	MSE	0.09	0.09	0.05	0.07	0.09
	length	1.14	1.15	0.81	1.03	4.61
	coverage (%)	92.90	92.40	94.90	87.70	100.00
β_2	estimate	0.46	0.46	0.39	0.40	−0.22
	variance	0.08	0.08	0.04	0.07	1.08
	bias	0.04	0.04	−0.04	−0.03	−0.64
	std bias	13.30	13.51	−24.11	−11.80	−189.88
	MSE	0.07	0.07	0.03	0.04	0.48
	length	1.08	1.09	0.76	0.99	4.02
	coverage (%)	93.30	93.00	96.20	96.20	100.00
β_3	estimate	−7.84	−7.86	−9.61	−7.97	0.02
	variance	0.38	0.38	0.13	0.41	43.19
	bias	1.89	1.87	0.12	1.76	9.75
	std bias	330.99	335.09	31.11	293.23	978.35
	MSE	3.65	3.60	0.12	3.17	95.51
	length	2.37	2.36	1.39	2.46	25.54
	coverage (%)	2.60	3.10	93.10	3.70	95.80

3.2. *Fully conditional specification*

To create synthetic data sets under a FCS, our goal, as before, was to sample from the posterior predictive distribution. Here, synthetic data sets were created using a FCS as described in Van Buuren *et al.* [26].

In this approach, a model must be specified for each variable, conditional on the other variables. The goal of FCS is to approximate draws from the posterior distribution by sampling from the simpler, conditional distributions of the form

$$P(y_1|y_2, y_3, \ldots, y_j, \theta_1)$$

$$\vdots$$

$$P(y_j|y_1, y_2, \ldots, y_{j-1}, \theta j)$$

where $y_1, \ldots, y_j$ are the j variables and $\theta_i, i = 1, \ldots, j$ are the appropriate parameters of each conditional distribution assigned to the variable. Imputations were created by iteratively sampling from the conditional univariate distributions via Gibbs sampling [29]. As pointed out in Van

Table 3. Results evaluating the performance of regression coefficients estimated using synthetic data with M=10, Σ_1, using only observations with $T_m > 0$. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.90	29.89	30.30	30.72	28.15
	variance	1.81	1.81	0.49	1.84	35.97
	bias	−0.72	−0.73	−0.33	0.10	−2.47
	std bias	−77.14	−77.60	−63.37	56.84	−44.04
	MSE	1.76	1.77	0.79	0.92	10.57
	length	4.84	4.82	2.51	4.89	22.37
	coverage (%)	85.40	84.50	76.90	94.50	99.90
	$T_m > 0$	811	793	918	840	1000
β_1	estimate	0.36	0.35	0.51	0.25	0.51
	variance	0.16	0.16	0.06	0.13	1.63
	bias	−0.07	−0.07	0.09	−0.18	0.085
	std bias	−17.56	−23.40	60.12	−90.69	8.44
	MSE	0.11	0.11	0.05	0.13	0.21
	length	1.42	1.43	0.87	1.31	4.75
	coverage (%)	89.70	89.90	87.50	85.80	100.0
	$T_m > 0$	783	790	925	787	998
β_2	estimate	0.46	0.47	0.39	0.40	−0.18
	variance	0.14	0.14	0.05	0.12	1.15
	bias	0.04	0.04	−0.04	−0.03	−0.60
	std bias	12.26	16.39	−25.52	−6.34	−162.37
	MSE	0.09	0.09	0.04	0.08	0.53
	length	1.34	1.34	0.83	1.27	3.96
	coverage (%)	90.10	90.70	89.10	92.10	98.70
	$T_m > 0$	792	796	925	761	999
β_3	estimate	−7.87	−7.88	−9.62	−7.94	0.08
	variance	0.57	0.57	0.18	0.54	48.18
	bias	1.87	1.85	0.11	1.79	9.81
	std bias	367.72	417.35	39.64	351.97	150.44
	MSE	3.63	3.58	0.14	3.35	99.94
	length	2.72	2.71	1.52	2.67	26.72
	coverage (%)	24.40	24.40	88.60	24.20	85.30
	$T_m > 0$	846	840	912	848	1000

Buuren *et al.* [26], "the parameters $\theta_1, \ldots, \theta_j$ are treated as specific to the respective conditional densities and are not necessarily the product of some factorization of the 'true' joint distribution."

When dealing with a small number of variables, each variable can be conditioned on all of the remaining variables. However, as the number of variables increases, this method becomes tedious, as each variable must have an individual model specified. Three methods of creating synthetic data using this approach were considered in this article: predictive mean matching (PMM), normal Bayesian regression with SR on the indicator, and normal Bayesian regression for the continuous variables and logistic regression for the indicator variable. Algorithms used to impute based on these methods are detailed in Van Buuren *et al.* [26].

4. Results

We tested five different methods of creating synthetic data, under two different covariance structures, using two different values of M (10 and 100), and two different variance estimates (throwing out negative estimates and replacing negative estimates). For each synthetic data scenario,

Table 4. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 100$, Σ_1, using only observations with $T_m > 0$. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.87	29.88	30.40	30.72	28.26
	variance	1.10	1.07	0.34	1.06	36.33
	bias	−0.76	−0.74	−0.22	0.10	−2.12
	std bias	−86.18	−84.04	−39.65	10.97	−162.24
	MSE	1.39	1.36	0.64	0.47	5.85
	length	4.01	3.95	2.17	3.93	23.32
	coverage (%)	88.10	88.00	79.40	97.60	100.0
	$T_m > 0$	999	995	985	1000	1000
β_1	estimate	0.35	0.35	0.49	0.26	0.50
	variance	0.09	0.09	0.05	0.07	1.42
	bias	−0.07	−0.07	0.07	−0.16	0.08
	std bias	−27.02	−28.59	33.11	−71.77	23.57
	MSE	0.09	0.09	0.05	0.07	0.09
	length	1.14	1.15	0.81	1.02	4.60
	coverage (%)	92.90	92.40	94.90	87.00	100.0
	$T_m > 0$	996	994	1000	992	1000
β_2	estimate	0.46	0.46	0.39	0.40	−0.22
	variance	0.08	0.08	0.04	0.07	1.08
	bias	0.03	0.04	−0.04	−0.03	−0.64
	std bias	13.24	13.44	−24.11	−11.84	−189.88
	MSE	0.07	0.07	0.03	0.04	0.48
	length	1.08	1.09	0.77	0.982	4.01
	coverage (%)	93.20	93.00	96.20	96.20	100.0
	$T_m > 0$	996	997	1000	992	1000
β_3	estimate	−7.84	−7.86	−9.61	−7.97	0.02
	variance	0.38	0.38	0.13	0.41	43.19
	bias	1.89	1.88	0.118	1.76	9.75
	std bias	331.13	335.09	31.11	293.23	978.35
	MSE	3.65	3.60	0.12	3.17	95.51
	length	2.37	2.36	1.39	2.46	25.54
	coverage (%)	2.50	3.10	93.10	3.70	95.80
	$T_m > 0$	999	1000	1000	1000	1000

regression coefficients were computed and compared with the population parameters. Confidence intervals ($\alpha = 0.05$) were computed using the normal approximation. We measured the accuracy of the results by looking at the bias, standard bias (std bias), mean squared error (MSE), confidence interval length (length), and confidence interval coverage (coverage). For confidence interval coverage, the regression coefficients' intervals estimated from the synthetic data were compared with the actual population parameters. Standard bias was computed using the formula $(100)(q - \hat{q}_M)/\sqrt{T_M}$. In scenarios where negatively estimated variance were dropped, the table notes how many out of the 1000 runs were estimated at greater than zero.

In Tables 1–8, we present statistics for the regression coefficients estimated by the synthetic data. Each column presents a different method of creating the synthetic data. The first two columns of all eight tables contain statistics obtained from synthetic data created with the joint MVN model with SR and AR, respectively. The last three columns show statistics for the estimated regression coefficients from the synthetic data created using FCS. The three methods tested here that use a FCS were PMM, normal regression for all variables with SR for the indicator (All-Norm), and normal regression for continuous variables with logistic regression for the indicator (Norm-Log).

Tables 1–4 are the results when the smaller covariance structure, Σ_1, was used, while Tables 5–8 show results for the larger covariance structure, Σ_2. In Tables 1, 2, 5, and 6, negative estimates of

Table 5. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 10$ and Σ_2. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.54	29.55	30.21	29.59	30.32
	variance	1.37	1.34	1.11	1.35	1.19
	bias	−0.82	−0.80	0.06	−0.76	−0.03
	std bias	−91.21	−94.78	7.17	−90.78	−8.92
	MSE	1.67	1.64	1.18	1.64	0.83
	length	4.31	4.25	3.83	4.23	3.90
	coverage (%)	84.60	84.80	88.30	83.70	89.20
β_1	estimate	0.41	0.41	0.43	0.40	0.38
	variance	0.06	0.06	0.04	0.06	0.05
	bias	−0.02	−0.02	0.01	−0.02	−0.05
	std bias	−10.21	−12.32	10.80	−32.46	−33.16
	MSE	0.05	0.05	0.03	0.05	0.05
	length	0.87	0.87	0.70	0.85	0.78
	coverage (%)	89.80	89.00	88.80	87.90	81.80
β_2	estimate	0.44	0.44	0.46	0.43	0.48
	variance	0.05	0.05	0.04	0.05	0.04
	bias	0.01	0.01	−0.01	0.01	0.05
	std bias	4.36	5.03	−6.52	4.15	25.98
	MSE	0.04	0.04	0.03	0.04	0.05
	length	0.84	0.84	0.73	0.83	0.71
	coverage (%)	91.60	90.70	88.40	91.40	84.00
β_3	estimate	−7.41	−7.43	−10.14	−7.50	−9.08
	variance	1.46	1.47	1.32	1.39	1.35
	bias	1.74	1.72	−0.20	1.66	0.08
	std bias	205.07	216.09	−27.58	222.38	10.36
	MSE	3.96	3.92	1.32	3.69	1.04
	length	4.40	4.40	4.20	4.30	4.33
	coverage (%)	62.10	63.10	85.40	62.50	89.20

Table 6. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 100$ and Σ_2. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.51	29.52	30.20	29.59	30.42
	variance	0.96	0.92	0.79	0.93	0.83
	bias	−0.85	−0.84	0.05	−0.76	0.07
	std bias	−93.90	−94.83	8.46	−69.82	7.17
	MSE	1.42	1.40	1.31	1.16	0.84
	length	3.76	3.69	3.73	3.71	3.52
	coverage (%)	85.80	85.10	86.30	92.50	93.00
β_1	estimate	0.40	0.40	0.44	0.38	0.41
	variance	0.04	0.04	0.03	0.05	0.03
	bias	−0.02	−0.02	0.02	0.04	−0.01
	std bias	−11.88	−11.87	13.66	−20.61	−7.23
	MSE	0.04	0.04	0.031	0.02	0.04
	length	0.78	0.78	0.65	0.81	0.71
	coverage (%)	92.40	92.60	91.20	96.80	92.10
β_2	estimate	0.44	0.44	0.45	0.47	0.43
	variance	0.04	0.04	0.03	0.04	0.03
	bias	0.02	0.02	−0.02	0.04	0.01
	std bias	7.42	7.35	−8.19	19.46	2.20
	MSE	0.03	0.03	0.03	0.02	0.04
	length	0.74	0.75	0.61	0.77	0.68
	coverage (%)	93.00	93.60	90.50	97.60	92.80
β_3	estimate	−7.36	−7.37	−10.04	−7.62	−9.13
	variance	0.96	0.96	0.89	0.92	0.98
	bias	1.79	1.74	−0.10	1.54	0.02
	std bias	197.32	196.95	−13.10	126.95	3.54
	MSE	3.85	3.81	1.55	2.74	0.98
	length	3.78	3.77	3.64	3.71	3.83
	coverage (%)	53.30	54.00	84.60	77.20	92.50

variance were replaced with $\bar{v}_m$, while Tables 3, 4, 7, and 8 display results using only observations with positive variance estimates, while noting the number of positive variance estimates out of 1000.

Figures 1–5 show scatter plots of the regression coefficient of the indicator variable from the actual data compared with the corresponding regression estimate from the synthetic data for the larger covariance structure and $M = 100$, which corresponds to the data used to create Table 6.

4.1. *Joint multivariate normal model*

Using a joint MVN model to produce synthetic data sets has been shown to be useful when the model truly fits the data [24]. As a result of the departures from normality in our data, we observed poorer results.

Using the smaller covariance matrix, Σ_1, with $M = 10$ and $M = 100$, both SR and AR, we observed large biases all larger than 1.8 and low coverages, all of which were less than 25%, for the regression coefficient of the indicator variable, as seen in Tables 1 and 2. When replacing negative variance estimates under both SR and AR with $M = 10$, we observed a coverage percentage of 22.6, and when $M = 100$ the coverage percentage drops to 2.6 and 3.1% for SR and AR, respectively. When only using positive variance estimates, the coverage for the indicator is only

Table 7. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 10$, Σ_2, using only observations with $T_m > 0$. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.54	29.55	30.3	29.6	30.34
	variance	1.40	1.36	1.14	1.37	1.22
	bias	−0.82	−0.81	0.05	−0.75	−0.02
	std bias	−94.47	−98.64	6.66	−93.94	−6.57
	MSE	1.68	1.65	1.18	1.65	0.84
	length	4.32	4.25	3.85	4.21	3.91
	coverage (%)	83.30	83.50	88.00	82.00	88.00
	$T_m > 0$	857	853	859	843	903
β_1	estimate	0.41	0.41	0.43	0.40	0.37
	variance	0.06	0.06	0.04	0.06	0.05
	bias	−0.02	−0.02	0.013	−0.02	−0.06
	std bias	−10.83	−12.98	13.56	−34.40	−39.31
	MSE	0.05	0.05	0.03	0.05	0.06
	length	0.88	0.87	0.70	0.85	0.78
	coverage (%)	89.10	88.30	87.60	87.20	79.60
	$T_m > 0$	882	888	871	891	876
β_2	estimate	0.44	0.44	0.45	0.43	0.48
	variance	0.05	0.05	0.04	0.05	0.04
	bias	0.01	0.01	−0.02	0.003	0.06
	std bias	6.05	6.69	−10.65	3.24	25.78
	MSE	0.04	0.04	0.03	0.04	0.06
	length	0.85	0.84	0.74	0.83	0.71
	coverage (%)	90.70	89.90	87.40	91.00	81.40
	$T_m > 0$	880	882	890	880	859
β_3	estimate	−7.42	−7.45	−10.11	−7.49	−9.08
	variance	1.47	1.47	1.34	1.37	1.38
	bias	1.73	1.70	−0.16	1.67	0.07
	std bias	213.36	226.34	−24.53	243.37	10.79
	MSE	3.95	3.89	1.32	3.69	0.89
	length	4.35	4.35	4.22	4.19	4.34
	coverage (%)	60.60	61.30	84.50	59.90	88.60
	$T_m > 0$	862	853	914	802	832

slightly better when $M = 10$ at 24.4%, and when $M = 100$ we observe the same coverages as before since nearly all variance estimates are positive. While the estimates of the indicator perform poorly, the regression coefficients for the continuous variables, under both SR and AR, all have coverage percentages slightly above 90% for both $M = 10$ and $M = 100$.

As seen in Tables 5 and 6, using the larger covariance structure, Σ_2, for both $M = 10$ and $M = 100$, the estimate of the regression coefficient of the indicator variable performs better. We observed a smaller bias and increased coverage. For $M = 10$ we observed a coverage of 62.1% and 63.1% for SR and AR respectively, and when $M = 100$ we observed a coverage of 53.3% and 54.0%, both of which are higher than their respective coverage percentages when using Σ_1 as the covariance structure.

From Tables 3 and 4, we observed that with SR and AR and $M = 10$ nearly 80% of our variance estimates were positive, and when $M = 100$ nearly all of the variance estimates were positive. From Tables 7 and 8 we observe slightly improved results under the larger covariance structure. When $M = 10$ over 85% of variance estimates are positive and when $M = 100$ there were no negative estimates of variance under either SR or AR.

Table 8. Results evaluating the performance of regression coefficients estimated using synthetic data with $M = 100$, Σ_2, using only observations with $T_m > 0$. Each column presents a different method used to create synthetic data. β_0 is the model intercept, β_1 and β_2 are the coefficients for the continuous variables, and β_3 is the coefficient for the indicator.

		Multivariate normal		Fully conditional specification		
		SR	AR	PMM	All-Norm	Norm-Log
β_0	estimate	29.51	29.52	30.2	29.59	30.42
	variance	0.96	0.92	0.79	0.93	0.83
	bias	−0.85	−0.84	0.05	−0.76	0.07
	std bias	−93.90	−94.83	8.46	−85.24	7.17
	MSE	1.42	1.40	1.31	1.16	0.84
	length	3.76	3.69	3.73	3.71	3.52
	coverage (%)	85.80	85.10	86.00	93.00	93.00
	$T_m > 0$	1000	1000	1000	999	1000
β_1	estimate	0.40	0.40	0.44	0.38	0.41
	variance	0.04	0.04	0.03	0.05	0.03
	bias	−0.02	−0.02	0.02	−0.04	−0.01
	std bias	−11.88	−11.87	13.66	−20.53	−7.23
	MSE	0.04	0.04	0.03	0.02	0.038
	length	0.78	0.78	0.65	0.81	0.71
	coverage (%)	92.40	92.60	91.20	96.80	92.10
	$T_m > 0$	1000	1000	1000	1000	1000
β_2	estimate	0.44	0.44	0.45	0.47	0.43
	variance	0.04	0.04	0.03	0.04	0.03
	bias	0.02	0.02	−0.02	0.04	0.01
	std bias	7.42	7.35	−8.19	19.60	2.20
	MSE	0.03	0.03	0.03	0.02	0.04
	length	0.74	0.75	0.61	0.78	0.68
	coverage (%)	93.00	93.60	90.50	97.60	92.80
	$T_m > 0$	1000	1000	1000	1000	1000
β_3	estimate	−7.36	−7.37	−10.04	−7.62	−9.13
	variance	0.96	0.96	0.89	0.92	0.98
	bias	1.79	1.78	−0.10	1.54	0.02
	std bias	197.32	196.95	−13.10	169.68	3.54
	MSE	3.85	3.77	1.55	2.74	0.98
	length	3.78	3.77	3.64	3.71	3.83
	coverage (%)	53.30	54.00	84.60	77.20	92.50
	$T_m > 0$	1000	1000	1000	999	1000

Figures 1 and 2 show the scatter plot of the synthetic estimates versus the actual estimates for the regression coefficient of the indicator when synthetic data sets were created using SR and AR, respectively. Here we can see graphically the bias introduced when we created synthetic data sets from this data using these methods.

4.2. *Fully conditional specification*

Of the three methods for creating synthetic data sets under a FCS, PMM appears to be best overall. However, in certain situations, namely with the larger covariance structure, Norm-Log is better. All-Norm performs the worst of the three FCS methods.

As seen in Tables 1 and 2, PMM performs very well based on measures of bias and coverage, with the smaller covariance matrix, Σ_1. PMM also consistently produced the smallest variance and MSE in the estimates. With the smaller covariance matrix, the coverage of the estimate of the regression coefficient for the indicator variable was 88.8% and 93.1% for $M = 10$ and $M = 100$, respectively. That is in comparison for All-Norm that has coverages of the indicator estimate of 22.8% and 3.0%, again, for $M = 10$ and $M = 100$, respectively. One can also see

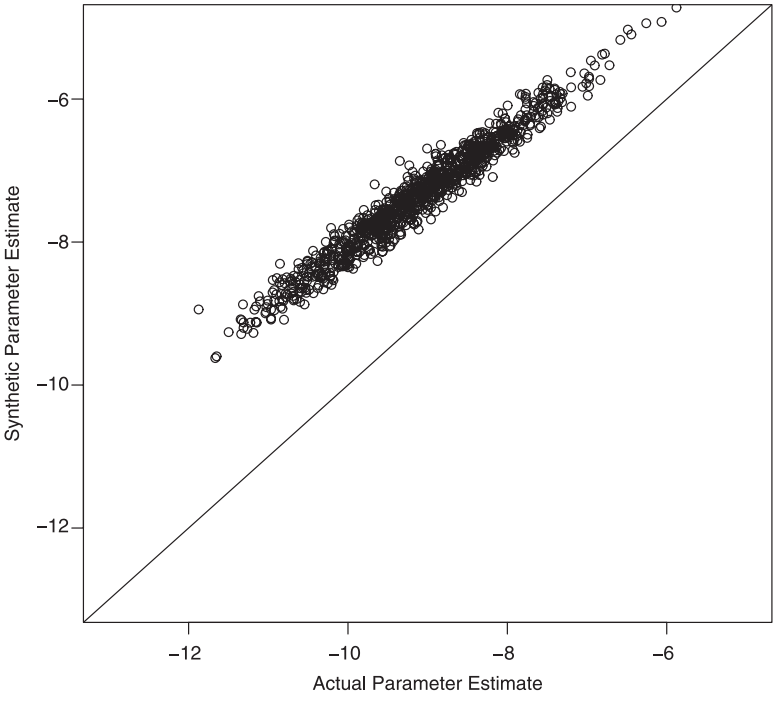

Figure 1. Synthetic regression parameters estimates versus actual regression parameter estimates with synthetic data created using SR, $M = 100$, and Σ_2.

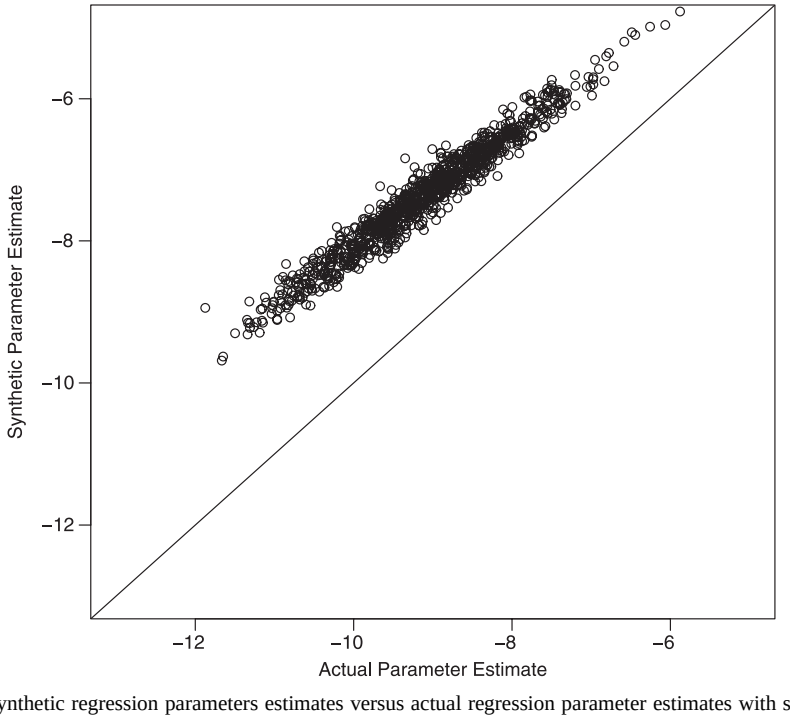

Figure 2. Synthetic regression parameters estimates versus actual regression parameter estimates with synthetic data created using AR, $M = 100$, and Σ_2.

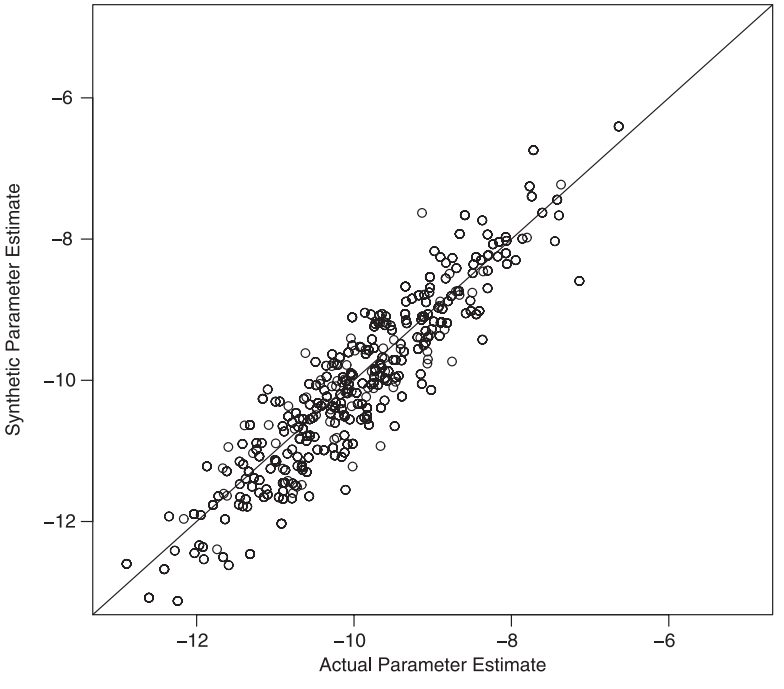

Figure 3. Synthetic regression parameters estimates versus actual regression parameter estimates with synthetic data created using PMM, $M = 100$, and Σ_2.

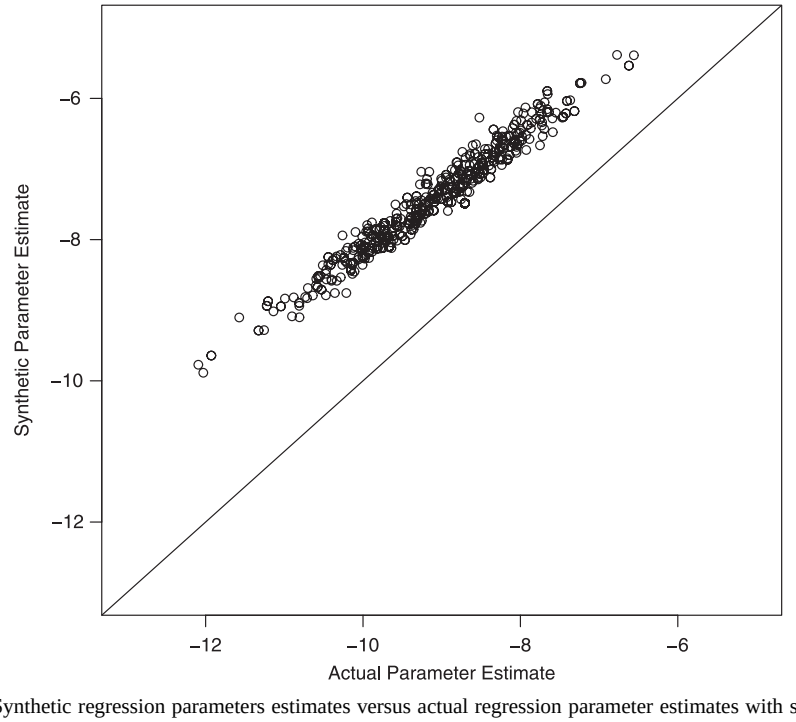

Figure 4. Synthetic regression parameters estimates versus actual regression parameter estimates with synthetic data created using All-Norm, $M = 100$, and Σ_2.

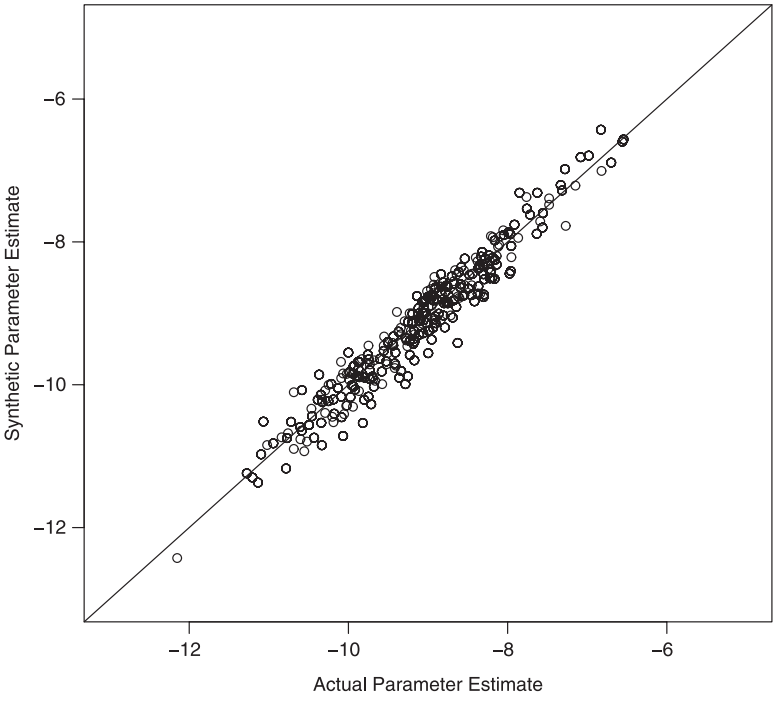

Figure 5. Synthetic regression parameters estimates versus actual regression parameter estimates with synthetic data created using Norm-Log, $M = 100$, and Σ_2.

that PMM had similar measures of bias and MSE compared with All-Norm for estimates of continuous variables. However, for the estimates of the regression coefficient of the indicator, we observed much larger MSE and bias for All-Norm than Norm-PMM, as seen in Tables 1 and 2. With the smaller covariance structure, we cannot really compare the Norm-Log method to the others because the algorithm used to estimate the logistic regression coefficients did not converge, rendering the synthetic data sets created with Norm-Log essentially useless.

When we only use observations with positive variance estimates we see similar results as before, with PMM performing the best of the fully conditional methods. We also see that when $M = 10$ over 90% of variance estimates are positive for PMM compared with only about 75% of positive variance estimates for All-Norm. When $M = 100$, both PMM and All-Norm produced nearly 100% positive variance estimates.

When using the larger covariance structure, Σ_2, we observed a clearer picture of the performance of Norm-Log. As seen in Table 6 for $M = 10$, Norm-Log performed similarly to PMM in terms of bias and MSE for the continuous variables, with Norm-Log slightly better. For the estimate of the regression parameter of the indicator variable with $M = 10$, from Table 5 we see that Norm-Log actually outperforms PMM based on both MSE and coverage. Norm-Log has a coverage of 89.2% while PMM has a coverage of 85.4%. All-Norm lags considerably behind both methods with coverage of only 62.5% for the indicator variable. Figures 3–5 show scatter plots of the synthetic estimates versus the actual estimates for PMM, All-Norm, and Norm-Log, respectively. One can see that PMM and Norm-Log performed similarly. All-Norm still exhibits a significant bias in the estimate of the regression coefficient for the indicator variable, similar to the MN methods SR and AR.

As seen in Tables 7 and 8, the number of positive variance estimates is well over 80% for all of the methods of FCS when $M = 10$, and nearly all variance estimates are positive when

$M = 100$. These results are similar to the number of positive variance estimates with the smaller covariance matrix. Thus, it appears that the amount of negative variance estimates is unaffected by our choice of covariance structures, Σ_1 or Σ_2, and depends almost entirely on our choice of M.

5. Discussion

In this article, the inferences created from fully synthetic data sets for conventional least squares regression where a binary variable is present were examined. In general, the only method that performed well in this study was PMM. However, in the context of synthetic data PMM may be problematic because we are imputing missing values with actual observations from the data and not random draws from a distribution. All imputed values in the synthetic data for each variable actually appeared somewhere in the true data set, which means that sometimes complete true data points are released. This is a good illustration of the potential trade-off between the accuracy of the estimates from the synthetic data and the protection of privacy.

When using SR and AR, estimates for continuous variables were good. However, any encouraging findings for the continuous variables were eliminated by the poor performance of the estimates of the regression coefficient for the indicator variable. These two methods displayed large bias and poor coverage for both values of M and both covariance structures.

All-Norm performed similarly to both of the MVN methods. We observed similar levels of bias in All-Norm and the MVN methods, but All-Norm consistently reduced the MSE of the estimates, if only slightly. Despite this improvement over the MVN methods, All-Norm still fell short of the other two fully conditional methods, PMM and Norm-Log.

With the smaller covariance matrix, Σ_1, and using Norm-Log, the algorithm to estimate the logistic regression parameters failed to converge because of quasi-complete separation. A possible remedy to this situation is to use a penalized likelihood estimation procedure proposed by Firth [30]. When quasi-separation is not present, as in the case with the larger covariance structure, Σ_2, Norm-Log performed very well. In fact, Norm-Log actually performed the best of all the methods based on measures of bias, MSE, and coverage, but, only with the larger covariance structure.

Raghunathan *et al.* [24] showed that when fully synthetic data sets are created using a joint multivariate model, the estimates of the regression coefficients from the synthetic data were very close to the actual values of the parameters. It works well because the joint multivariate model is the true model for the data. Here, the joint multivariate model performed poorly because the model did not truly fit our data. We attempted to use the joint MVN model with binary data because Bernaards *et al.* [25] successfully used this model with rounding techniques. In that article, the highest missingness rate for a single item was only 23.8%. In this article, we were dealing with a missingness rate of 100%, and we saw that the joint multivariate model does not produce good results in this situation.

For each method tested here there is some sort of trades off. The MVN methods, AR and SR, and All-Norm offer the simplest method for creating synthetic data, but they performed the worst when trying to estimate the regression coefficient for the indicator variable. When Norm-Log was used to create synthetic data sets, the regression coefficient was more accurately estimated. However, we can only rely on this method when the data are suitable and the logistic regression algorithm converges. Finally, PMM performs very well under all of the scenarios presented in this article, but at a cost to privacy since we may be releasing complete actual data points. This article is not meant to invalidate the methods used to create synthetic data; however, it does illustrate some of the potential hazards of creating and performing simple analysis on synthetic data sets using these procedures.

References

[1] L. Sweeney, *k-Anonymity: A model for protecting privacy*, Internat. J. Uncertain. 10 (2002), pp. 557–570.
[2] D. An and R.J.A. Little, *Multiple imputation: An alternative to top coding for statistical disclosure control*, J. R. Stat. Soc. Ser. A 170 (2007), pp. 923–940.
[3] T. Dalenius and S.P. Reiss, *Data-swapping: A technique for disclosure control*, J. Stat. Plann. Inference 6 (1982), pp. 73–85.
[4] G. Duncan and R. Pearson, *Enhancing access to microdata while protecting confidentiality: Prospects for the future (with discussion)*, Statist. Sci. 6 (1991), pp. 219–239.
[5] L.H. Cox, *Suppression methodology and statistical disclosure control*, J. Amer. Statist. Assoc. 75 (1980), pp. 377–385.
[6] L.H. Cox, *Matrix masking methods for disclosure limitation in microdata*, Surv. Methodol. 6 (1994), pp. 165–169.
[7] S.E. Fienberg, U.E. Makov, and A.P. Sanil, *A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data*, J. Off. Stat. 13 (1997), pp. 75–89.
[8] G. Duncan and D. Lambert, *The risk of disclosure for microdata*, J. Bus. Econom. Statist. 7 (1989), pp. 207–217.
[9] W.J. Keller and J.G. Bethlehem, *Disclosure protection of microdata: Problems and solutions*, Statist. Neerlandica 46 (1992), pp. 5–19.
[10] R.J.A. Little, *Statistical analysis of masked data (Disc: P455-474) (Corr: 94V10 P469)*, J. Off. Stat. 9 (1993), pp. 407–426.
[11] D.B. Rubin, *Comment on "Statistical disclosure limitation"*, J. Off. Stat. 9 (1993), pp. 461–468.
[12] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.
[13] R.J.A. Little and D.B. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, New York, 1987.
[14] O. Harel and X.H. Zhou, *Multiple imputation: Review and theory, implementation and software*, Stat. Med. 26 (2007), pp. 3057–3077.
[15] A.B. Kennickell, *Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances*, in *Record Linkage Techniques*, N. Alvey and B. Jamerson, eds., National Acadamies Press, Washington, D.C. 1997, pp. 248–267.
[16] J. Abowd and S. Woodcock, *Disclosure limitation in longitudinal linked data*, in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz and J. Theeuwes, eds., Amsterdam, North-Holland, 2001, pp. 215–277.
[17] F. Liu and R.J.A. Little, *Selective multiple imputation of keys for statistical disclosure control in microdata*, in *Proceedings of the Joint Statistical Meet*, 2002, pp. 2133–2138.
[18] J.P. Reiter, *Inference for partially synthetic, public use Microdata sets*, Surv. Methodol. 29 (2003), pp. 181–188.
[19] J.P. Reiter, *Using CART to generate partially synthetic public use microdata*, J. Off. Stat. 21 (2005), pp. 441–462.
[20] J.P. Reiter, *Satisfying disclosure restriction with synthetic data sets*, J. Off. Stat. 18 (2002), pp. 531–543.
[21] J.P. Reiter, *New approaches to data dissemination: A glimpse into the future (?)*, Chance 17 (2004), pp. 11–15.
[22] J.P. Reiter, *Simultaneous use of multiple imputation for missing data and disclosure limitation*, Surv. Methodol. 30 (2004), pp. 235–242.
[23] J.P. Reiter, *Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study*, J. Roy. Statist. Soc. Ser. A 168 (2005), pp. 185–205.
[24] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin, *Multiple imputation for statistical disclosure limitation*, J. Off. Stat. 19 (2003), pp. 1–16.
[25] C.A. Bernaards, T.R. Belin, and J.L. Schafer, *Robustness of a multivariate normal approximation for imputation of incomplete binary data*, Stat. Med. 26 (2007), pp. 1368–1382.
[26] S. Van Buuren, J. Brand, C. Groothuis-Oudshoorn, and D. Rubin, *Fully conditional specification in multivariate imputation*, J. Stat. Comput. Simul. 76 (2006), pp. 1049–1064.
[27] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2007, ISBN 3-900051-07-0. Available at http://www.R-project.org/.
[28] S. Van Buuren and C. Oudshoorn, *mice: Multivariate Imputation by Chained Equations, R package version 1.16*, 2007, Available at http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm/.
[29] S. Geman and D. Geman, *Stochasitc relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984), pp. 721–741.
[30] D. Firth, *Bias reduction of maximum likelihood estimates (Corr: 95V82 P667)*, Biometrika 80 (1993), pp. 27–38.