



A note on imputing squares via polynomial combination approach

Mingyang Cai¹  · Gerko Vink¹

Received: 29 December 2020 / Accepted: 4 January 2022 / Published online: 4 February 2022
© The Author(s) 2022

Abstract

The polynomial combination (PC) method, proposed by Vink and Van Buuren, is a hot-deck multiple imputation method for imputation models containing squared terms. The method yields unbiased regression estimates and preserves the quadratic relationships in the imputed data for both MCAR and MAR mechanisms. However, Vink and Van Buuren never studied the coverage rate of the PC method. This paper investigates the coverage of the nominal 95% confidence intervals for the polynomial combination method and improves the algorithm to avoid the perfect prediction issue. We also compare the original and the improved PC method to the substantive model compatible fully conditional specification method proposed by Bartlett et al. and elucidate the two imputation methods' characters.

Keywords Multiple imputation · Missing data · Quadratic relation · Squared terms

1 Introduction

Squared terms are often included in real-life data models to accommodate some form of nonlinearity. When the analysis model contains the partially observed covariates and corresponding squared terms, some challenges arise:

1. The analysis and imputation models should accommodate squared terms, i.e., the squares themselves should be considered in the imputation procedure of corresponding linear terms.
2. The relation between the square term and its lower-order polynomial should be preserved.
3. The analysis model parameter estimates should be unbiased.

✉ Mingyang Cai
m.cai@uu.nl

Gerko Vink
g.vink@uu.nl

¹ Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

To obtain unbiased estimates, one could impute the squared term as if it were another variable. We will refer to this as *Transform, then Impute* (TTI; Von Hippel 2009). However, this approach distorts the relationship between the original variable and its square. A straightforward process to preserve the quadratic relation during imputation is calculating the squared term only after imputation (*Impute, then Transform*, ITT). However, ITT biases estimates of regression coefficients, as its contribution during imputation is ignored (Von Hippel 2009; Vink and van Buuren 2013). Moreover, both these partial fixes only work when the missingness is completely random.

To solve these issues for a more general class of missingness mechanisms, Vink and van Buuren (2013) propose to impute the combination of the original variable and its square and decompose it into distinct roots. This polynomial combination approach (PC) is built around predictive mean matching, a nonparametric imputation hot-deck technique that does not assume a specific distribution for the data (Rubin 1987; Little 1988). Seaman et al. (2012) demonstrated that predictive mean matching gives biased estimation when the analysis is a linear regression with a quadratic term and the missingness mechanism is missing at random. However, PC yields unbiased estimates for MCAR and MAR missingness mechanisms by applying a reasonable donor selection procedure on the polynomial combination instead.

More recently, Bartlett et al. (2015) proposed a substantive model compatible approach (SMC-FCS), which generalizes the imputation of nonlinear covariates beyond the squared term model. The SMC-FCS technique is efficient but needs the correct data analysis model during imputation to obtain draws of values that conform to this model. It yields unbiased estimates if 1) the substantive model is correctly specified and 2) the normality assumption of missing variables with quadratic effects is tenable because of a restriction of the software (the package `smcfcs` (Bartlett et al. 2021) in R (R Core Team 2021)). When the missing variable with quadratic effects does not follow the normal distribution, one could apply an appropriate transformation to make the normality assumption plausible (e.g., log-normal distribution). More interestingly, Bartlett et al. (2015) suggest that the SMC-FCS estimates can yield unbiased inference, meaning that estimates are both unbiased and properly covered cf. Neyman (1934). Such an investigation into coverage of multiply imputed parameters was not part of the study by Vink and van Buuren (2013).

We now have two techniques that seem promising in imputing squared terms: the polynomial combination method and SMC-FCS. Both approaches have appealing properties, preserve the relationship between the square and its base, and yield unbiased estimates. However, both techniques differ fundamentally in their approach. SMC-FCS is a strictly model-based technique that requires the correct specification of the complete-data model and the substantive model. On the other hand, PC is a hot-deck technique with a data-driven estimation procedure that only requires the specification of the polynomial combination. We highlighted the most used properties and promising methods in Table 1 (Von Hippel 2009; Vink and van Buuren 2013; Bartlett et al. 2015).

The interpretation of Table 1, taking the SMC-FCS approach as an example, should be as follows:

1. SMC-FCS yields unbiased regression estimates β provided that the missing mechanism is *MAR*;

Table 1 Summary of properties of four squared term imputation methods

	TTI	ITT	PC	SMC-FCS
Unbiased estimates of β	MCAR only	–	MCAR & MAR	MCAR & MAR
Quadratic relationship	Not preserved	Preserved	Preserved	Preserved
Coverage rate of β	Poor	Poor	Unknown	Correct
Violation of normality			Robust	Somewhat robust
Model specification			Non-parametric	Parametric

- SMC-FCS does *preserve* the quadratic relationship between the original variable and its square;
- SMC-FCS produces *correct* coverage rates of corresponding regression estimates β ;
- SMC-FCS is *somewhat robust* against the violation of normality assumption of the covariates with quadratic effects;
- SMC-FCS is a *parametric* imputation approach, which means SMC-FCS requires an explicit specified imputation model.

The “-” sign in a cell indicates that the method cannot produce unbiased estimates. Whether the PC method has a correct coverage rate is not thoroughly studied and will be investigated in the following section. There are four blank cells left because the TTI and ITT methods cannot produce unbiased regression estimates or preserve the quadratic relationships; it is redundant to investigate the violation of normality and model specification for them.

In this manuscript, we evaluate the performance of imputing squared terms with SMC-FCS and the PC method to investigate these techniques’ strengths and limitations in different scenarios. In the next section, we briefly discuss the SMC-FCS and PC methodology and propose a minor adjustment to the PC method.

2 Polynomial combination

In this section we detail both the original polynomial combination (OPC) approach proposed by Vink and van Buuren (2013) and a modification that is robust against perfect prediction issues. We refer to the modified polynomial combination approach as MPC.

2.1 Original polynomial combination

Suppose the model of scientific interest is

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon \quad (1)$$

with $\epsilon \sim N(0, \sigma^2)$. We assume that Y is complete and that $X = (X_{obs}, X_{mis})$ is partially missing.

The original polynomial combination method first performs predictive mean matching (PMM; Little 1988) on the combined variable $Z = X\beta_1 + X^2\beta_2$, and then decomposes Z into components X and X^2 . Under the model in Eq. (1), two roots of variable X are:

$$\begin{aligned} X_- &= -\frac{1}{2\beta_2}(\sqrt{4\beta_2 Z + \beta_1^2} + \beta_1) \\ X_+ &= \frac{1}{2\beta_2}(\sqrt{4\beta_2 Z + \beta_1^2} - \beta_1) \end{aligned} \quad (2)$$

where the discriminant $4\beta_2 Z + \beta_1^2$ should be larger than 0. For any imputed Z , we select either $X = X_-$ or $X = X_+$ and square it to derive the square term X^2 .

The choice between the roots X_- and X_+ is made by random sampling, conditional on Y , Z , and their interaction YZ . The binary random variable V is defined as 0 if $X < X_{min}$ and 1 if $X > X_{min}$, where the minimum of the parabola $X_{min} = -\beta_1/2\beta_2$. We model the probability $P(V = 1)$ by logistic regression as

$$\text{logit}P(V = 1) = Y\beta_Y + Z\beta_Z + YZ\beta_{YZ} \quad (3)$$

on the observed data, where $\text{logit}P(V = 1) = \log(P(V = 1)/P(V = 0))$ is the logistic function. Under the assumption of ignorability, we apply the same model to calculate the predicted probability $P(V = 1)$ for Z_{mis} , where Z_{mis} denotes the polynomial combination of X_{mis} and X_{mis}^2 . Finally, a random draw from the binomial distribution is made ($V = 0$ or 1), and the corresponding (negative or positive) root is selected as the imputation.

2.2 Modification of polynomial combination

Since we estimate binary variables V in the OPC imputation procedure, it is necessary to avoid bias due to perfect prediction. When imputers apply the original polynomial combination method, perfect prediction occurs when all the observed binary variables V_{obs} are equal to one (or zero). In this case, the likelihood tends to a limit as one or some regression coefficients tend to infinity, which leads to seriously implausible imputations of the binary variable V (White et al. 2010).

Suppose all observed X are located on the parabolic function's right arm, then the perfect prediction arises. If no corrections are performed, the coefficients of the logistic function $\text{logit}P(V = 1) = Y\beta_Y + Z\beta_Z + YZ\beta_{YZ}$ will have extremely wide and flat posterior distributions, which tends to derive extremely positive or negative estimates of coefficients. Provided all observed X are located on the right arm of the parabolic function, some missing values of X would be addressed incorrectly on the left arm, as shown in Fig. 1a.

A computationally convenient approach to avoid perfect prediction is data augmentation (van Buuren 2018, section 3.6.2). We augment the data with a few extra observations and add a small weight to these observations (White et al. 2010, section 5.2). To improve the polynomial combination method, we calculate any unobserved

Table 2 Augmented data

	V	Y	Z
1	1	$E(Y_{obs}) + SD(Y_{obs})$	$E(Z_{obs})$
2	1	$E(Y_{obs}) - SD(Y_{obs})$	$E(Z_{obs})$
3	0	$E(Y_{obs}) + SD(Y_{obs})$	$E(Z_{obs})$
4	0	$E(Y_{obs}) - SD(Y_{obs})$	$E(Z_{obs})$
5	1	$E(Y_{obs})$	$E(Z_{obs}) + SD(Z_{obs})$
6	1	$E(Y_{obs})$	$E(Z_{obs}) - SD(Z_{obs})$
7	0	$E(Y_{obs})$	$E(Z_{obs}) + SD(Z_{obs})$
8	0	$E(Y_{obs})$	$E(Z_{obs}) - SD(Z_{obs})$

dichotomous outcomes (whether to take the positive or negative distinct real root for X_{mis}) V_{mis} by logistic regression of V given Y, Z, and YZ (i.e., Eq. 3) with the augmented data instead of the observed data. More specifically, based on the observed V, Y and Z, the augmented data adds eight subjects shown in Table 2, with the weight 3/8, to the observed data. When the population estimation of the probability $P(V = 1)$ equals one (or zero), we expect the modified polynomial combination method would provide more plausible imputations, as shown in Fig. 1b.

2.3 SMC-FCS

The substantive model compatible fully conditional specification (SMC-FCS) is a parametric imputation method proposed by Bartlett et al. (2015). In general, the missing predictor is imputed based on other predictors. A rejection sampling (e.g., Metropolis-Hastings algorithm) is used where the acceptance ratio is generated based on the likelihood of the substantive model. Suppose ϕ is a vector containing the coefficients of the model $f(Y|X)$ and $\theta_i, i = 1, \dots, p$ is a vector containing the coefficients of the model $f(X_i|X_{-i})$, where X_{-i} are all the other covariates excluding X_i . The parametric density function of the partially observed variable X_i is proportional to $f(Y|X, \phi)f(X_i|X_{-i}, \theta_i)$, rooted in the Bayesian rule:

$$f(X_i|X_{-i}, Y) = \frac{f(X_i, X_{-i}, Y)}{f(Y, X_{-i})} \propto f(Y|X_i, X_{-i})f(X_i|X_{-i}). \tag{4}$$

Since the density generally does not follow a standard parametric family, the rejection sampling is necessary to draw coefficients from the posterior distributions of ϕ and θ_i . With the assumption of independent priors $f(\phi)$ and $f(\theta_i)$, the posterior distributions of ϕ and θ_i would be:

$$\begin{aligned} \phi &\sim f(Y|X_i, X_{-i}, \phi)f(\phi) \\ \theta_i &\sim f(X_i|X_{-i}, \theta_i)f(\theta_i). \end{aligned} \tag{5}$$

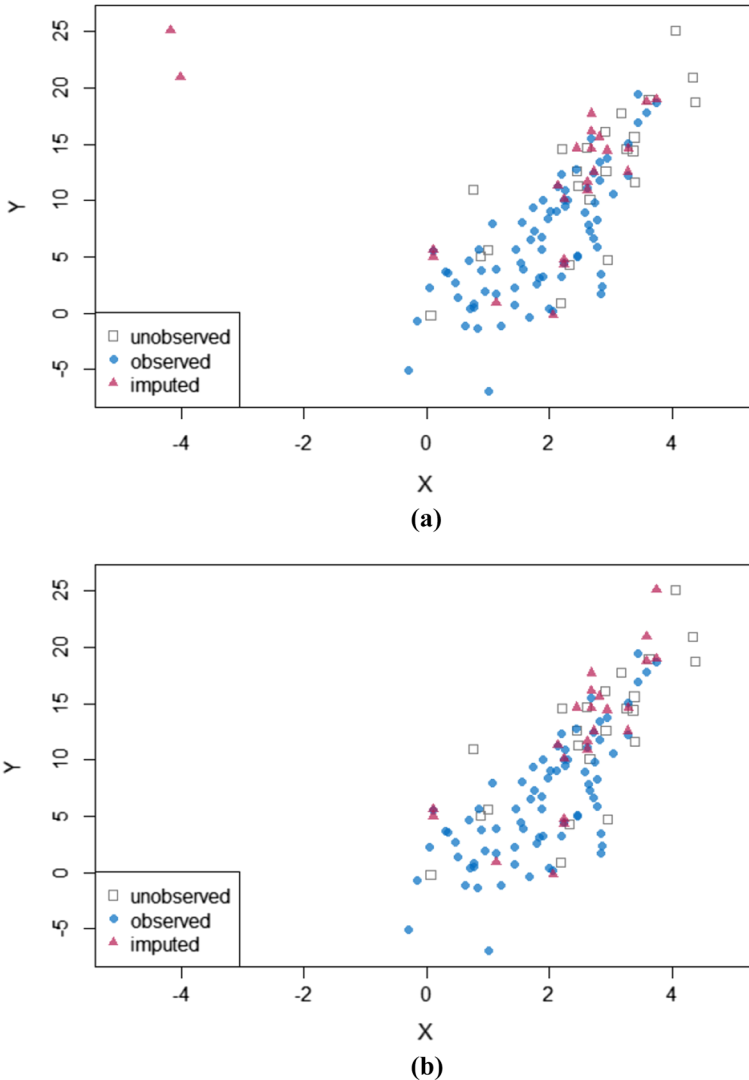


Fig. 1 Imputations (triangles) generated by OPC and MPC. We see that in **a** some imputations fall outside of the range of the observed (circle) and unobserved values (square), due to the OPC algorithm assigning the donor values to the incorrect distinct real root. In **b** the MPC approach assigns the imputations to the distinct real root that corresponds to the observed and unobserved data

The statistical property of this approach is that if the substantive model $f(Y|X)$ is correctly specified, the imputation model will be congenial to the analysis model (Meng 1994). The lack of congeniality can sometimes produce implausible imputations that result in biased inferences in the downstream analysis (Robins and Wang 2000).

3 Evaluation

We evaluated the average biases across all simulations, the coverage of nominal 95% confidence intervals, and the average width of corresponding confidence intervals of the regression weights β_1 and β_2 .

3.1 Simulation setup

The outcome Y was simulated according to the scientific model:

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \epsilon \quad (6)$$

with $\alpha = 0$, $\beta_1 = 1$, $\beta_2 = 1$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$. The value of σ_ϵ varied according to different distributions of X so that the coefficient of determination R^2 was always equal to 0.75.

The predictor X was generated from a normal, a skew-normal, or a normal mixture distribution. The mean of X was either 0 or 1, and the variance was 1 for all three distributions. The abscissa at the parabolic minimum was $X = -1/2$. Hence, when the location of X was 0, there was a strong U-shaped association between Y and X . If X had location 2, the relationship between Y and X would be somewhat linear. For the skew-normal distribution, we set the slant parameter to be 6 when the mean of X equaled 0 and -3 when the mean of X equaled 2. For the normal mixture distribution, X was drawn from $N(-0.875, 0.234)$ and $N(0.875, 0.234)$ with equal probability to have mean 0 and $N(1.125, 0.234)$ and $N(2.875, 0.234)$ with equal probability to have mean 2.

We generated a sample of size $n = 100$ and repeated 1000 simulations for each missingness scenario. For each simulation scenario, 30 percent missingness was induced jointly in X and X^2 for five missingness mechanisms: MCAR, MARleft, MARmid, MARTail, and MARright. Specifically, MCAR denotes that the probability of X being missing is the same for all cases. While with a left-tailed (MARleft), centered (MARmid), both tailed (MARTail) or right-tailed (MARright) missingness mechanism, a higher probability of X being missing is assigned to the units with low, centered, extreme and high values of Y respectively. Let R be the response indicator for X , where R equals 0 if X is missing and 1 otherwise. For MARleft, the missingness probability is defined as $\text{logit}P(R = 0) = -X + \bar{x} + \gamma_l$, where γ_l was chosen to make the probability of missing X equal to 0.3. Similarly, the missingness probability is defined as $\text{logit}P(R = 0) = -|X - \bar{x}| + \gamma_m$, $\text{logit}P(R = 0) = |X - \bar{x}| + \gamma_t$ and $\text{logit}P(R = 0) = X - \bar{x} + \gamma_r$ for MARmid, MARTail and MARright, where γ_m , γ_r and γ_t were chosen to make corresponding probabilities of missing X equal to 0.3 ((van Buuren 2018), section 3.2.4). All missingness was generated with the `ampute` function (Schouten et al. 2018) from the package MICE (van Buuren and Groothuis-Oudshoorn 2011) in R. The `mice.impute.quadratic` function in the package MICE was modified by including data augmentation.

3.2 Simulation results

We compared five approaches: TTI, ITT, OPC, MPC and SMC-FCS and focused on some remarkable findings of OPC, MPC and SMC-FCS. The results of the TTI and ITT simulations reiterated the corresponding conclusions in Table 1. In general, TTI did not preserve the quadratic relation, even though it gave unbiased and confidence-valid estimates in some cases (e.g., with MCAR and standard normal distribution X). Furthermore, ITT had considerable bias under nearly all combinations of missingness mechanisms and distributions of X .

Table 3 shows the average biases, the coverage of the nominal 95% confidence intervals, and the average width of confidence intervals for β_1 and β_2 when $E(X)$ equals 0. The outcome Y follows a U-shape. With MCAR, MARleft and MARmid and when X is distributed as normal, skewed normal or a mixture of two normals, OPC and MPC gave unbiased estimates and correct CI coverage. The CI coverage of SMC-FCS was close to 95%. However, with X skew-normal distributed MCAR and MARmid, SMC-FCS was slightly biased. With MARTail and MARright, SMC-FCS outperformed OPC and MPC when X followed a normal distribution or a normal mixture distribution. OPC and MPC had slight bias and somewhat reduced CI coverage (approximately 85%) with X distributed according to a normal, a skewed normal or a mixture of two normals. SMC-FCS was unbiased and had CI coverage close to 95% with normal and mixture normal. However, with skewed normal X , SMC-FCS was somewhat biased and the CI had slightly lower than nominal coverage.

Table 4 demonstrates the mean biases of β_1 and β_2 , the empirical coverage and the mean width of the corresponding 95% CIs where X is location 2 and scale 1. Almost all observed values of Y are on the right arm of the quadratic function. With normal X , SMC-FCS consistently yielded confidence-valid estimates because of the congeniality of the analysis model and imputation model. However, with MAR (MARleft, MARmid, MARTail and MARright), SMC-FCS gave a slightly biased estimate for β_1 . OPC and MPC gave unbiased results and the CI had approximately 95% coverage under MCAR and MARmid. With MARleft, OPC and MPC were slightly biased for β_1 , but the 95% CI for β_1 and β_2 had the correct coverage. With MARTail and MARright, OPC and MPC were unbiased. The CI of MPC (around 90%) had higher nominal coverage than OPC (around 80%). With X distributed according to a skewed normal, OPC and MPC yielded unbiased estimates under MCAR, MARmid, MARTail and MARright but slightly biased estimates under MARleft. The CI coverage of OPC and MPC was close to 95% under MCAR, MARleft and MARmid. Like the case with normal X , the CI of MPC (around 90%) had better coverage than OPC (around 85%). With X distributed as a mixture of two normals, SMC-FCS gave biased results under all missingness mechanisms and its CI had somewhat reduced coverage under MARright. OPC and MPC were unbiased under MCAR, MARmid and MARTail but biased under MARleft and MARright. MPC had CI coverage close to 95% under all missingness mechanisms. The CI from OPC had correct coverage under MCAR, MARleft and MARmid, but approximately 85% coverage under MARTail and MARright.

We investigated if the biases of SMC-FCS were caused by Monte Carlo error. Figures 2a, b demonstrate that the bias is due to simulation error. The estimates for β_1

Table 3 Simulation results for five missingness mechanisms when imputing a squared term regression where the mean of X equals 0. Shown are absolute bias of the estimate, coverage of the 95% confidence interval for the estimate and the average confidence interval width

Missingness mechanism														
MCAR			MARleft			MARmid			MARtail			MARright		
Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw
Normal														
TTI														
β_1	0.01	0.94	0.52	-0.05	0.92	0.47	0.96	0.51	0.11	0.88	0.63	0.21	0.78	0.74
β_2	0	0.94	0.39	-0.06	0.9	0.34	0.93	0.36	0.1	0.86	0.51	0.2	0.73	0.62
ITT														
β_1	-0.05	0.9	0.9	-0.02	0.98	0.94	0.98	0.86	-0.04	0.99	1.09	-0.04	0.98	1.13
β_2	-0.25	0.88	0.89	-0.25	0.86	0.86	0.9	0.78	-0.34	0.84	1.17	-0.31	0.87	1.16
OPC														
β_1	0.03	0.94	0.56	0.01	0.95	0.46	0.95	0.48	0.08	0.87	0.75	0.09	0.88	0.87
β_2	0.03	0.92	0.41	0	0.94	0.33	0.94	0.34	0.13	0.84	0.6	0.15	0.83	0.72
MPC														
β_1	0.03	0.94	0.56	0.01	0.95	0.46	0.95	0.47	0.1	0.86	0.75	0.1	0.88	0.86
β_2	0.03	0.92	0.41	0	0.94	0.33	0.94	0.34	0.13	0.84	0.6	0.15	0.83	0.72
SMC-FCS														
β_1	0	0.96	0.53	-0.01	0.95	0.47	0.94	0.49	0	0.96	0.6	0	0.95	0.67
β_2	0	0.95	0.39	0	0.95	0.33	0.95	0.34	0.01	0.95	0.51	0.02	0.95	0.57
Skewed-normal														
TTI														
β_1	-0.15	0.95	3.21	-0.25	0.92	2.98	0.15	2.79	-0.43	0.92	4.93	-0.26	0.94	5.75

Table 3 continued

Missingness mechanism												
MCAR												
MARleft			MARmid			MARtail			MARright			
Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	
β_2	0.94	1.48	0.04	0.93	1.28	-0.09	0.94	1.23	0.35	0.88	2.59	0.91
ITT												
β_1	0.94	2.55	0.27	0.93	2.24	0.18	0.96	2.39	0.5	0.9	2.91	0.94
β_2	0.96	1.27	-0.12	0.93	1.01	-0.07	0.96	1.07	-0.24	0.9	1.71	0.95
OPC												
β_1	0.93	2.68	-0.02	0.94	2.49	0.03	0.93	2.4	-0.07	0.86	3.64	0.82
β_2	0.92	1.27	0.01	0.94	1.11	-0.01	0.94	1.08	0.16	0.85	1.98	0.82
MPC												
β_1	0.95	2.69	-0.02	0.94	2.47	0.03	0.93	2.39	-0.02	0.92	3.87	0.88
β_2	0.94	1.29	0.01	0.94	1.11	-0.01	0.93	1.08	0.15	0.91	2.09	0.86
SMC-FCS												
β_1	0.94	2.59	-0.01	0.95	2.39	-0.15	0.96	2.43	-0.41	0.93	3.51	0.91
β_2	0.95	1.21	0	0.94	1.08	0.07	0.96	1.07	0.28	0.91	1.87	0.88
Normal-mixture												
TTI												
β_1	0.96	0.4	-0.05	0.93	0.38	0	0.95	0.4	0.02	0.95	0.43	0.86
β_2	0.96	0.46	-0.05	0.94	0.42	-0.04	0.95	0.44	0.02	0.96	0.52	0.91
ITT												
β_1	0.98	0.62	0.03	0.98	0.59	-0.02	0.98	0.58	-0.07	0.98	0.65	0.98

Table 3 continued

Missingness mechanism															
MCAR															
	MARleft			MARmid			MARtail			MARright					
	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw			
β_2	-0.42	0.59	0.98	-0.36	0.71	0.96	-0.32	0.72	0.86	-0.55	0.35	0.97	-0.52	0.4	0.96
OPC															
β_1	0.02	0.94	0.38	0	0.93	0.36	0.01	0.94	0.37	0.06	0.88	0.44	0.08	0.87	0.48
β_2	0.01	0.94	0.44	0	0.93	0.41	0	0.94	0.41	0.04	0.92	0.53	0.05	0.92	0.58
MPC															
β_1	0.02	0.94	0.38	0	0.93	0.36	0.01	0.94	0.36	0.06	0.89	0.44	0.08	0.87	0.48
β_2	0.01	0.95	0.44	0	0.94	0.41	0	0.94	0.4	0.04	0.92	0.53	0.05	0.93	0.58
SMC-FCS															
β_1	-0.01	0.95	0.4	-0.02	0.95	0.38	0.02	0.95	0.38	-0.05	0.93	0.43	-0.05	0.93	0.48
β_2	-0.02	0.95	0.46	0.01	0.94	0.41	-0.03	0.94	0.42	-0.03	0.94	0.51	-0.06	0.93	0.57

Table 4 Simulation results for five missingness mechanisms when imputing a squared term regression where the mean of X equals 2. Shown are absolute bias of the estimate, coverage of the 95% confidence interval for the estimate and the average confidence interval width

Missingness mechanism												
MCAR												
MARleft			MARmid			MARtail			MARright			
Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	
Normal												
TTT												
β_1	-0.58	0.93	7.17	-1	0.92	8	-0.08	0.95	6.04	-1.32	0.91	10.11
β_2	0.11	0.93	1.66	0.15	0.93	1.73	0	0.95	1.39	0.31	0.89	2.43
ITT												
β_1	0.71	0.94	5.61	0.6	0.94	5.57	0.5	0.96	5.27	0.99	0.92	5.95
β_2	-0.19	0.95	1.41	-0.14	0.95	1.31	-0.13	0.96	1.28	-0.28	0.9	1.58
OPC												
β_1	-0.05	0.92	5.28	-0.15	0.93	5.56	0.01	0.93	4.98	-0.12	0.87	6.04
β_2	0.02	0.93	1.27	0.03	0.94	1.27	0	0.94	1.17	0.05	0.87	1.51
MPC												
β_1	-0.05	0.93	5.17	-0.15	0.94	5.39	0	0.91	4.77	-0.03	0.93	6.26
β_2	0.02	0.93	1.25	0.03	0.94	1.24	0.01	0.92	1.13	0.04	0.91	1.56
SMC-FCS												
β_1	0.03	0.95	5.44	-0.11	0.95	5.78	-0.06	0.95	5.14	-0.22	0.94	6.02
β_2	0	0.95	1.3	0.03	0.94	1.31	0.01	0.95	1.25	0.06	0.94	1.5
Skewed-normal												
TTT												
β_1	-0.15	0.93	2.95	-0.45	0.92	4.14	-0.06	0.95	2.56	-0.32	0.95	4.07
β_2	0.06	0.93	1.27	0.14	0.92	1.59	0.03	0.95	1.11	0.12	0.94	1.73
ITT												
β_1	-0.07	0.94	2.97	-0.07	0.94	2.97	0.06	0.93	1.41	0.06	0.93	1.41

Table 4 continued

Missingness mechanism															
MCAR			MARleft			MARmid			MARtail			MARright			
Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	
β_1	0.46	0.92	2.64	0.29	0.93	2.86	0.35	0.94	2.49	0.62	0.87	2.83	0.73	0.81	2.48
β_2	-0.28	0.88	1.26	-0.13	0.94	1.23	-0.21	0.9	1.16	-0.37	0.8	1.37	-0.48	0.64	1.26
OPC															
β_1	0.01	0.92	2.24	-0.11	0.93	2.74	0.02	0.93	2.19	-0.07	0.88	2.5	0	0.85	2.19
β_2	0	0.93	0.99	0.05	0.93	1.14	-0.01	0.94	0.96	0.04	0.89	1.11	0.02	0.86	1.03
MPC															
β_1	0	0.92	2.18	-0.11	0.93	2.66	0	0.92	2.02	-0.04	0.94	2.59	0.02	0.91	2.28
β_2	0.01	0.93	0.96	0.05	0.93	1.11	0	0.93	0.89	0.04	0.95	1.15	0.02	0.9	1.06
SMC-FCS															
β_1	0.1	0.95	2.46	-0.04	0.95	3.01	0.09	0.95	2.27	0.14	0.95	2.65	0.2	0.93	2.22
β_2	-0.05	0.94	1.08	0.03	0.95	1.22	-0.04	0.94	1	-0.08	0.94	1.2	-0.14	0.91	1.04
Normal-mixture															
TTI															
β_1	-1.5	0.9	10.48	-2.15	0.86	11.93	-0.71	0.93	8.8	-2.35	0.87	13.63	-1.73	0.91	12.47
β_2	0.33	0.9	2.52	0.43	0.87	2.73	0.16	0.94	2.11	0.51	0.88	3.29	0.43	0.9	3.15
ITT															
β_1	1.38	0.88	6.4	0.83	0.92	6.13	1.07	0.92	6.33	1.76	0.8	6.08	2.08	0.78	6.35
β_2	-0.35	0.88	1.62	-0.19	0.94	1.52	-0.25	0.92	1.58	-0.49	0.76	1.57	-0.57	0.72	1.63
OPC															
β_1	-0.05	0.92	6.28	-0.24	0.92	6.51	0.02	0.93	6.05	-0.09	0.88	6.61	0.19	0.85	6.43

Table 4 continued

Missingness mechanism															
MCAR															
	MARleft			MARmid			MARtail			MARright					
	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw	Bias	Cov	Ciw
β_2	0.02	0.92	1.53	0.06	0.92	1.57	0	0.94	1.47	0.04	0.87	1.64	-0.03	0.86	1.62
MPC															
β_1	-0.04	0.93	6.24	-0.23	0.92	6.39	0.02	0.91	5.92	-0.04	0.93	6.77	0.24	0.92	6.69
β_2	0.02	0.93	1.53	0.05	0.93	1.54	0	0.92	1.45	0.03	0.92	1.68	-0.03	0.92	1.69
SMC-FCS															
β_1	0.32	0.95	6.37	-0.32	0.96	6.85	0.25	0.94	6.18	0.5	0.94	6.58	0.95	0.91	6.69
β_2	-0.07	0.94	1.57	0.09	0.95	1.65	-0.04	0.94	1.5	-0.14	0.92	1.66	-0.25	0.9	1.71

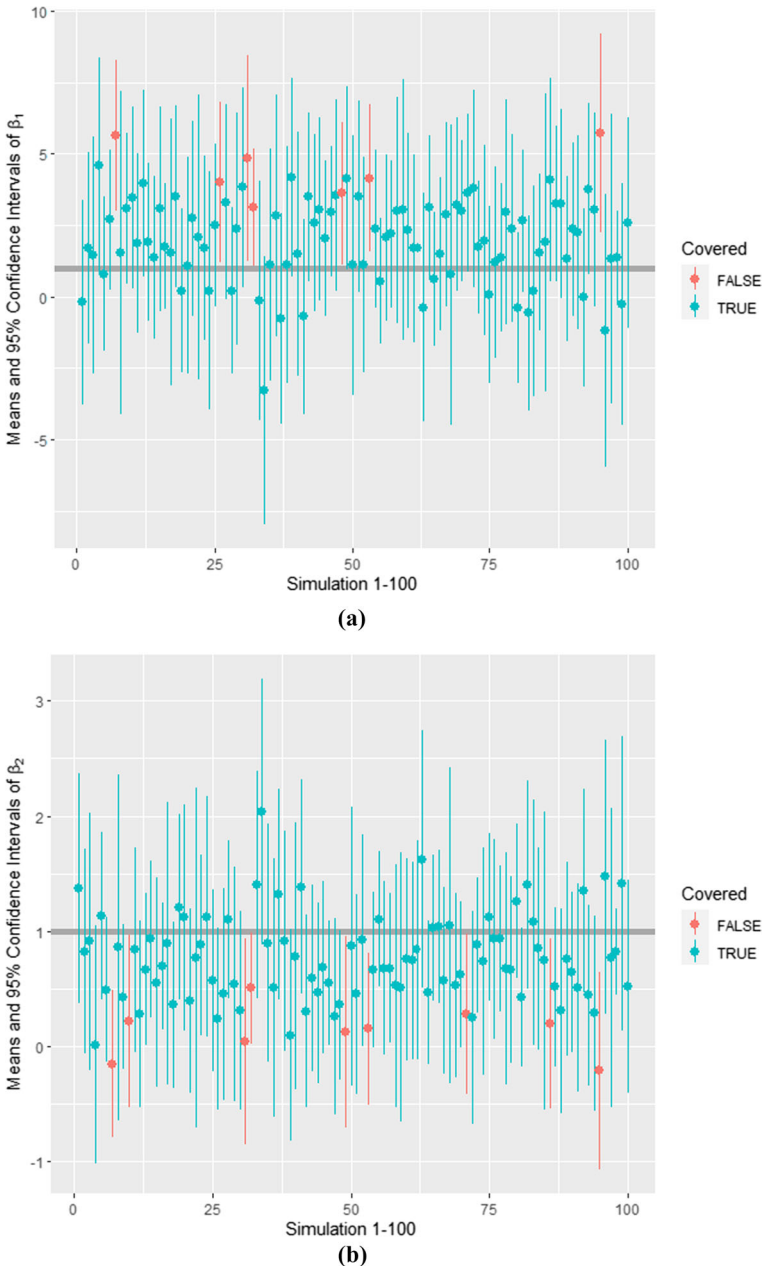


Fig. 2 The plot of means and confidence intervals of β_1 and β_2 from the first 100 simulations. The model of interest is $Y = X + X^2 + \epsilon$, where $X \sim \frac{1}{2}N(1.125, 0.234) + \frac{1}{2}N(2.875, 0.234)$. The missingness mechanism is MARright and the imputation approach is SMC-FCS. The imputations seem to primarily overestimate the true parameter estimate of β_1 in (a) and underestimate the true parameter estimate of β_2 in (b)

and β_2 show primarily overestimation and underestimation, respectively. This implies that, when applying SMC-FCS, the explicit specification of the distribution of the incomplete variable with the quadratic effect may need careful consideration.

4 Conclusion

We evaluate the performance of four imputation approaches for incomplete data problems where the model of scientific interest contains squared terms. We improve the performance of the polynomial combination method by incorporating a data augmentation step, thus realizing more plausible imputations when the missingness covariate relates almost exclusively to one arm of the quadratic curve.

In our simulation studies, although ITT preserves the quadratic relations, it gives biased estimates under almost all combinations of experimental factors. Oppositely, TTI provides unbiased estimates in some cases but fails to keep the quadratic relations in the imputed data. With normally distributed predictors and right-tailed missingness mechanisms, the performance of SMC-FCS is superior to that of MPC, with coverages closer to the nominal level. However, when the normality assumption is violated, the polynomial combination yields less biased estimates. Overall, both the SMC-FCS and polynomial combination methods produce plausible imputations of squared terms and outperform TTI and ITT. Differences between the approaches only become apparent under intense MARtail and MARright scenarios in simulation. However, these two mechanisms are more extreme than we are likely to see in practice since there is a strong relationship between the outcome Y and the probability of the variable X being unobserved in the tail. All in all, when differences in performance are found, such differences are small, and it may be challenging to interpret them as meaningful. This means that, in practice, the choice for an imputation approach could largely be a choice of preference.

If there is a solid, well-known scientific model, we highly recommend using SMC-FCS to sharpen results. The substantive model would then be correctly specified, ensuring that the distribution from which imputations are generated is compatible. SMC-FCS is a reliable model-based method to impute predictors with quadratic effects. It is theoretically well-grounded, and procedures are available for substantive models based on standard regression, discrete outcomes and proportional hazards van Buuren (2018). However, with an increasing number of variables in the dataset, it becomes increasingly challenging to infer the correct substantive model based on the incomplete data a priori. The strategy of applying SMC-FCS in practice is performing model selection once imputed datasets are generated to ensure the accuracy of substantive model specification, which is not a trivial process (Bartlett et al. 2015). Usually, the substantive model is specified according to prior studies or assumptions.

In contrast, we advise using the polynomial combination approach when the scientific model is less specific or when modeling efforts are challenging. It is proven to be a valid data-driven imputation method that is flexible in applying because we only need to specify the quadratic term. This makes it straightforward to implement in any imputation effort. The polynomial combination method is based on predictive mean matching, and the performance of imputation procedures involving PMM are proven

to work well in a wide range of research problems (Vink et al. 2015; Rubin 1986; Little 1988). Therefore, we expect that the polynomial combination approach could be of great practical importance in incomplete data analyses with squared terms.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bartlett J, Keogh R, Bonneville EF (2021) smcfcs: multiple imputation of covariates by substantive model compatible fully conditional specification. URL <https://CRAN.R-project.org/package=smcfcs>, rpackage version 1.6.0
- Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's Disease Neuroimaging Initiative* (2015) Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Method Med Res* 24(4):462–487
- Little RJA (1988) Missing-data adjustments in large surveys. *J Bus Econ Stat* 6(3):287–296
- Meng XL (1994) Multiple-imputation inferences with uncongenial sources of input. *Stat Sci* pp. 538–558
- Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J Royal Stat Soc* 97(4):558–625
- R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>
- Robins JM, Wang N (2000) Inference for imputation estimators. *Biometrika* 87(1):113–124
- Rubin DB (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *J Bus Econ Stat* 4(1):87–94
- Rubin DB (1987) Multiple imputation for nonresponse in surveys. John Wiley and Sons, New York
- Schouten RM, Lugtig P, Vink G (2018) Generating missing values for simulation purposes: a multivariate amputation procedure. *J Stat Comput Simul* 88(15):2909–2930
- Seaman SR, Bartlett JW, White IR (2012) Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol* 12(1):1–13
- Vink G, van Buuren S (2013) Multiple imputation of squared terms. *Sociol Method Res* 42(4):598–607
- van Buuren S (2018) Flexible imputation of missing data, 2nd edn. Chapman and Hall/CRC, London
- van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. *J Stat Softw* 45(3):1–67
- Vink G, Lazendic G, van Buuren S (2015) Partitioned predictive mean matching as a large data multilevel imputation technique. *Psychol Test Assess Model* 57(4):577–594
- Von Hippel P (2009) How to impute interactions, squares, and other transformed variables. *Sociol Methodol* 39(1):265–291
- White IR, Daniel R, Royston P (2010) Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal* 54(10):2267–2275

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.