8-3-2024

Comparing the Performance of Different Missing Data Imputation Approaches in Discrete Outcome Modeling

Md Istiak Jahan
University of Central Florida

Tanmoy Bhowmik
Portland State University, tbhowmik@pdx.edu

Lauren Hoover
University of Central Florida

Naveen Eluru
University of Central Florida

Citation Details

Jahan, M. I., Bhowmik, T., Hoover, L., & Eluru, N. (2024). Comparing the Performance of Different Missing Data Imputation Approaches in Discrete Outcome Modeling. Transportation Research Record: Journal of the Transportation Research Board.

Addressing Uncertainties in Discrete Outcome Modelling Frameworks with Missing Data: A Comparative Study between Single and Multiple Imputation

Md. Istiak Jahan[*]
Graduate Research Assistant
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: md.istiak.jahan@ucf.edu
ORCiD number: 0000-0002-4056-7816

Tanmoy Bhowmik
Assistant Professor
Department of Civil, and Environmental Engineering
Portland State University
Tel: 407-927-6574
Email: tbhowmik@pdx.edu
ORCiD number: 0000-0002-0258-1692

Lauren Hoover
Graduate Research Assistant
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: lauren.hoover@ucf.edu

Naveen Eluru
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

Submitted for publication in Transportation Research Record

[*] Corresponding author

ABSTRACT

While several approaches exist for data imputation, these approaches are not commonly applied in transportation. The current paper is geared towards assisting transportation researchers and practitioners in developing models using datasets with missing data. The study begins with a data simulation exercise evaluating different solutions implemented for missing data. The dimensions considered in our analysis include: the nature of independent variables, different types of missing variables, different shares of missing values, multiple data sample sizes and evaluation of single imputation (SI), multiple imputation (MI) and complete case data (CCD) approach. The comparison is conducted by adopting the appropriate inference process for MI approach with multiple realizations. From the simulation exercise, we find that the MI approach consistently performs better than SI approach. Among various realizations, MI approach with five realizations is selected based on our results. The MI approach with five realizations is compared with the CCD approach under different conditions using model fit measures and parameter marginal effects. In the presence of a small share of missing data, for larger datasets, the results suggest that it might be beneficial to develop a CCD model by dropping observations with missing values as opposed to developing imputation models. However, when the share of missing data warrants variable exclusion, it is important and even necessary that multiple imputation approach be employed for model development. In the second part of the paper, based on our findings, we implemented the MI approach for real empirical datasets with missing values for four discrete outcome variables.

Keywords: Missing value, Data imputation, Multiple imputation, Data simulation

BACKGROUND

Transportation econometric model development is an important tool for researchers and practitioners across domains such as transportation safety, travel behavior, emerging transportation technology adoption, and traffic engineering. These econometric models are employed to determine the relationship between a dependent variable of interest (or multiple dependent variables) and a suite of independent variables. The development of these model systems relies on available data from public and private agencies. A common issue faced in the development of these models is related to missing information for independent variables. In public or private datasets, several reasons influence data unavailability for independent variables. First, in data collection efforts respondents might be unwilling to provide information for certain attributes (such as household income). Second, in several empirical contexts, data might be available only for the chosen alternative and is unavailable for all other alternatives (such as level of service measures for travel mode). Third, in complex data collection efforts such as naturalistic driving data or smartphone data collection, data can be missing due to technological issues (such as lost Cell/GPS connection) or privacy concerns (such as preserving the privacy of respondents). Finally, missing data can also occur due to respondent errors, and data transfer errors.

Several simplifications are applied to develop models in the presence of missing data in practice. First, the records with missing data are excluded from model development exercise. The process might seem innocuous in cases where missing data represent a small share (say <2%). However, in cases where several independent variables are affected these small percentages across the variables could result in substantially larger share of records to be removed. Also, the elimination of records with missing data can possibly result in larger standard errors for parameter estimates (*1*). Second, the variables with missing data are excluded from the analysis. The approach is employed when data missing is a significant share of the observations (such as above 30%) and/or there is reason to consider that missing data is not a truly random occurrence and is closely tied to the dependent variable or other independent variables. In these cases, the analyst is introducing misspecification in the model by eliminating the variable. Finally, the researchers can address the missing data problem by imputing data for these missing variables and then develop econometric models. While there is extensive literature in econometrics proposed and developed by Rubin and colleagues to address missing data related issues, we have found majority of transportation modeling approaches adopting the preceding two methods (*2*–*8*).

To be sure, research efforts have developed frameworks where imputation techniques were considered for model development (*9*–*12*). However, these approaches do not always systematically consider the potential uncertainty associated with imputation in their framework (as noted by Rubin and colleagues). Several approaches consider model development with a single imputation i.e., the missing data is imputed only once for each missing record. The imputation is achieved either employing a simple approach (such as mean or mode imputation) or a complex approach (such as using a regression model approach to generate the missing value). In both approaches a single imputation is considered to represent the missing value for the corresponding record and generate a full dataset. However, this process ignores the potential uncertainty associated with the imputation process (Rubin and his colleagues in several articles discussed this in great detail in (*2*, *4*, *6*–*8*, *13*). According to Rubin 1988 (*6*), approaches that employ single imputation and develop econometric models as if the imputed data creates a complete dataset systematically underestimate uncertainty in the data. Thus, the models developed with single imputation can result in incorrect inferences. Rubin and colleagues proposed techniques to improve inference from missing data using Multiple Imputation (MI) i.e., each missing record is expected to have multiple realizations with each realization resulting in one complete dataset. However, in these approaches, the complexity increases in terms of model inference as each realization results in one model. The analyst will need to employ inference approaches to generate parameter estimates (coefficient and standard error) from all imputed datasets.

The current research effort is geared toward evaluating and offering insights on imputation processes for datasets with missing data within a discrete choice modeling framework. Several research efforts (such as (*14*–*16*)) evaluated various imputation approaches for different types of missing values in discrete outcome models. While these studies focused on different approaches, it is also imperative to evaluate different dimensions of the problem including different types of missing values, share of missing values and nature of the variable(s) with missing values that require data imputation. In this study, we conduct a two-pronged analysis. First, we build on earlier research and evaluate the applicability of the MI approach proposed in literature employing a data simulation exercise. A dataset with a known data generation process (DGP) will allow us to create missing data with certain assumptions and test how different imputation approaches of varying complexity perform in terms of model inference. The dimensions considered in our analysis include: (a) the nature of missing independent variables including continuous and categorical variables, (b) different types of missing variables (Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR)), (c) different shares of missing values, (d) multiple data sample sizes and (e) single imputation (SI) approach, multiple imputation (MI) (with varying number of imputation realizations) and complete case data (CCD) approach (dropping records with missing values). The comparison of these imputation approaches will allow us to identify the complexity appropriate for applied research. Second, based on our findings we implement the superior approach in practice for real empirical datasets. In the empirical analysis we consider various discrete outcome modeling frameworks addressing four transportation safety variables including (a) crash prone segment selection, (b) crash prone intersection selection, (c) crash type and (d) crash severity.

The remainder of the paper is organized as follows. In the next section, a review of literature employing imputation in transportation is described while positioning the current study. The simulation experiment, the data generation and model estimation findings are presented in the subsequent section. The following section provides a description of empirical datasets with missing data that are employed for model development after imputation. Finally, the last section concludes the paper and identifies future directions of research.

LITERATURE REVIEW

Application of MI is very common in econometrics since first proposed by Rubin (*8*). In several studies Rubin and colleagues utilized the MI approach to tackle the issue of survey nonresponses (*2*, *3*, *5–7*, *17*). In recent years, application of MI for missing values is common in different domains of research, such as political science (*18*), medical science (*19*) and social economics (*20*).

Several research studies have examined data imputation approaches in transportation literature. The research has mainly proceeded in two directions. The first stream of research is geared towards comparing the performance of different imputation techniques in generating missing data. The approaches tested include mean/mode imputation or interpolation (*10*) least squares-based methods (*21*), latent variable approach (*9*, *22*), Fuzzy C-means imputation (*23*), principal component analysis (*24*), K-means clustering (*25*), K-nearest neighborhood approach (*26*), tensor-based traffic volume imputation (*27*), graph aggregate generative adversarial network approach (*28*), deep learning based algorithms (*29*), support vector machine (*11*, *21*), hot deck approach (*30*), multivariate imputation using chained equations (*31*), decision tree methodology (*32*), joint multivariate approaches (*33*), convolution neural networks (*21*) and inverse probability weighting (*34*).

Of particular relevance to the current study, the second stream of research studies focus on using imputed data for model development. The dimensions investigated include travel mode choice behavior (*35*), active transportation (*36*), transportation safety (*9*, *37*), bike-sharing system (*38*), and parking facility (*12*). In these studies, the number of datasets generated for imputation range from 1 (Single Imputation) to 100 (Multiple Imputation). The studies cited above that considered more than 1 dataset did employ relevant techniques for generating inferences using parameters from all datasets. However, several studies in transportation have avoided considering MI approaches because of the inherent complexity. For example, Budhwani et al. (*11*) stated in their study that MI is computationally burdensome with challenges in data generation and inference process.

Current Study

The current study is motivated toward clarifying the value of MI approach for missing data in model estimation and employing it for empirical application. The paper employs a two-pronged approach to address these objectives. First, the paper employs a data simulation experiment to compare how model parameter retrieval varies with SI and MI approaches. Using a multinomial logit model dependent variable, the experimental design explores the influence of missing values in the independent variables along these dimensions: (a) type of missing variables including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) values, (b) for continuous and categorical variables, (c) different shares of missing values (10%, 20% and 30%), (d) multiple sample sizes (500, 1000 and 2000). The experiment compared the performance of three approaches –SI, MI and CCD (removing missing value records). The data simulation results are compared using (a) model fit measures (log likelihood improvement) and (b) true parameter retrieval ability as determined from differences in marginal effects relative to the true model.

Second, drawing on the conclusions of the experimental design, the research study employed the appropriate method for empirical datasets. The case study consists of four datasets including (a) estimation of crash prone segment (MNL model), (b) estimation of crash prone intersection (MNL model), (c) estimation of different crash types (MNL model) and (d) estimation of crash severity (ordered logit model). Different categories of independent variables including - roadway and traffic characteristics, crash characteristics, vehicle characteristics, environmental characteristics and driver's demographic characteristics are employed in these modeling frameworks.

DATA SIMULATION AND EXPERIMENT

Data Simulation Approach

The dataset used for the simulation experiment is synthesized using a MNL model with a discrete variable of three alternatives and four exogenous variables. The simulation approach is described below:

Let, the number of alternatives in the simulated data is m $(m = 1, 2, 3)$ can be represented as a, b, and c; x_v are the exogenous variables where $v = 1, 2, 3, 4$; x_1 is a continuous generic variable, x_2 *and* x_3 are continuous alternative specific variables, and x_4 is a categorical dummy variable. Considering alternative – a as the base, the choice of alternative was synthesized using the following utility equations:

$$U_a = \beta_2 x_{2a} + \beta_3 x_{3a} + \varepsilon_1 \tag{1}$$
$$U_b = \beta_{0b} + \beta_{1b} x_1 + \beta_2 x_{2b} + \beta_3 x_{3b} + \beta_{4b} x_4 + \varepsilon_2 \tag{2}$$
$$U_c = \beta_{0c} + \beta_{1c} x_1 + \beta_2 x_{2c} + \beta_3 x_{3c} + \beta_{4c} x_4 + \varepsilon_3 \tag{3}$$

where, ε_m represents the error in the utility equation which is assumed to be independent and identically Gumbel-distributed across the dataset. The alternative with the maximum utility was considered as the chosen alternative in the simulated data.

The variable x_3 and x_4, representing a continuous and categorical variable, are selected as the variables of interest to evaluate data imputation. Therefore, three different Standard Deviation (SD)–mean ratios (0.25, 0.75, and 1.5) and two different means (0.4, 0.6) are considered while simulating variable x_3 and x_4 respectively. For each SD–mean ratio (or mean) 30 datasets of 500, 1000, 2000, 5000, and 10,000 samples were simulated. The simulation results across larger sample sizes are found to be consistent. Therefore, to conserve space, the results of the 500, 1000 and 2000-observation samples are presented[1]. The performance of the simulation exercise was conducted based on the parameter retrieval examined using: (1) absolute parameter bias and (2) asymptotic standard error (see (*39*) for similar analysis). The results of our simulation exercise are presented in Table 1. In the table, the SD-mean ratio of x_3 is considered as 1.5 and mean of x_4 is considered as 0.4. The retrieved mean parameter is calculated as the mean of estimated parameter of 30 samples. The absolute percentage bias was computed as

[1] The readers should recognize that estimating discrete outcome models implicitly assumes that model parameters converge asymptotically. However, depending on the characteristics of the dependent and independent variables, the sample size requirements for asymptotic convergence and parameter stability could vary substantially. Interested readers can explore earlier work on sample size requirements for their specific dataset following guidelines from earlier research (see (*44–47*)).

$\left(\left|\frac{Retreived\ mean\ parameter - True\ parameter}{True\ parameter}\right| * 100\right)$. The asymptotic standard error was computed as the standard error of parameter values across the samples. The values presented in Table 1 clearly illustrate that the simulation exercise retrieves the parameters with small bias and very small standard errors. The simulation experiment was extended for random parameters MNL model by introducing a random parameter for β_2 that follows a standard normal error term.

TABLE 1 Summary of data simulation

Variables	**Assumed mean**	**Assumed standard deviation**	**Assumed parameter**	**Retrieved mean parameter**	**Absolute percentage bias**	**Asymptotic standard error**
Intercept: b	NA	NA	-1.200	-1.192	0.635	0.070
Intercept: c	NA	NA	1.000	1.069	6.923	0.049
x_1	1.50	0.50	1.000	1.005	0.504	0.039
			0.500	0.476	4.863	0.024
x_{2a}	1.20	0.50	-0.500	-0.506	1.195	0.010
x_{2b}	0.70	0.20				
x_{2c}	1.60	1.20				
x_{3a}	1.50	2.25	-1.300	-1.327	2.100	0.008
x_{3b}	2.50	3.75				
x_{3c}	1.00	1.50				
x_4	0.4	NA	-1.000	-1.031	3.084	0.050
			-1.800	-1.841	2.300	0.041

Comparison Between Single and Multiple Imputation

The data simulated is employed to compare the performance of SI and MI. We examine the percentage error in true parameter retrieval across the two approaches. From the full datasets of different sample sizes (500, 1000 and 2000), we create missing variables for continuous variable (x_3) and categorical variable (x_4) randomly, completely randomly and in non-random fashion[2] (see (*40*) for different types of missing values) at different percentages (10, 20 and 30). Then, within the datasets with missing records, we impute the missing values for continuous and categorical variables based on equations 4 and 5 below:

$$\widetilde{x_v} = \mu_v + \sigma_v * x_{norm} \tag{4}$$

$$\widetilde{x_v} = \begin{cases} 1 & if\ x_{unif} \le \mu_v \\ 0 & if\ x_{unif} > \mu_v \end{cases} \tag{5}$$

where, $\widetilde{x_v}$ is the imputed data, μ_v and σ_v are the mean and standard deviation of the non-missing cases of the variable x_v; x_{norm} is a random variable that follows a standard normal distribution $N(0,1)$ and x_{unif} is a uniformly distributed random number in a range [0,1].

[2] In the simulation experiment, missing completely at random (MCAR) were created by removing specific percentages of values from the variable of interest randomly. Missing at random (MAR) data were created in two distinct approaches for continuous and categorical variables. For the cases of missing in continuous variable, MAR data were created by removing the specific percentages from an ascendingly ordered variable. On the other hand, for the cases of missing in categorical variable, MAR data were created by removing specific percentages of values from the largest category of the variable of interest. Missing not at random (MNAR) were created by removing the specific percentages of values of the variable of interest that correspond to the largest category of the dependent variable.

For SI, the parameter inference is straight forward as parameters from one dataset are randomly selected. For MI, the parameters retrieved from different number of simulated datasets (5, 10, 15 and 30) are compared. MI estimates need to be updated from the parameters from all datasets as follows.

$$\overline{\alpha_v} = \frac{\sum_{r=1}^{R} \alpha_{vr}}{R} \qquad (6)$$

$$\mu_v = \overline{\gamma_v} + \{(R+1)/R\} * \delta_v \qquad (7)$$

where, $\overline{\alpha_v}$ is the inference imputation estimate for variable v; α_{vr} is the estimate of variable v at r^{th} imputation; μ_v is the associated variance-covariance; $\overline{\gamma_v}$ is the within-imputation variability, which is equal to $(\sum_{r=1}^{R} \gamma_{vr})/R$; and $\delta_v$is the between-imputation variability, which is equal to $(\sum_{r=1}^{R}(\alpha_{vr} - \overline{\alpha_v})(\alpha_{vr} - \overline{\alpha_v})')/(R-1)$ (see (*3*) for more description). Now, if β_v is the true parameter of variable v estimated by using the simulated dataset before creation of missing cases, the computation of error ω in parameter retrieval will be as follows:

$$\omega = \left|\frac{\overline{\alpha_v} - \beta_v}{\beta_v}\right| * 100\% \qquad (8)$$

The parameter retrieval errors of MCAR, MAR and MNAR imputation at different SD-mean ratio for continuous variables and different means for categorical variables offer a similar trend. Therefore, in the interest of space, we only present the error percentages for two cases with variable x_3 with a SD-mean ratio of 0.75 and variable x_4 with a mean 0.6 in Table 2. From Table 2, we can see that as the share of missing values increases the parameter retrieval worsens as expected. Among the various samples presented we can also see that the mean error across parameters is the largest for SI (with 1 dataset) and becomes quite stable for 5 and above. Further, a comparison between the single and multiple imputations across different sample sizes is presented in Figure 1. It is noticeable in the figure that, in the case of datasets with 1000 or 2000 records, multiple imputation performs slightly better in parameter retrieval than the single imputation. However, for a very smaller dataset (N = 500), multiple imputation significantly outperforms single imputation. The result supports earlier literature and suggests the adoption of MI with 5 datasets as a reasonable solution for modeling exercises.

TABLE 2 Comparison between single and multiple imputation in retrieving true parameter

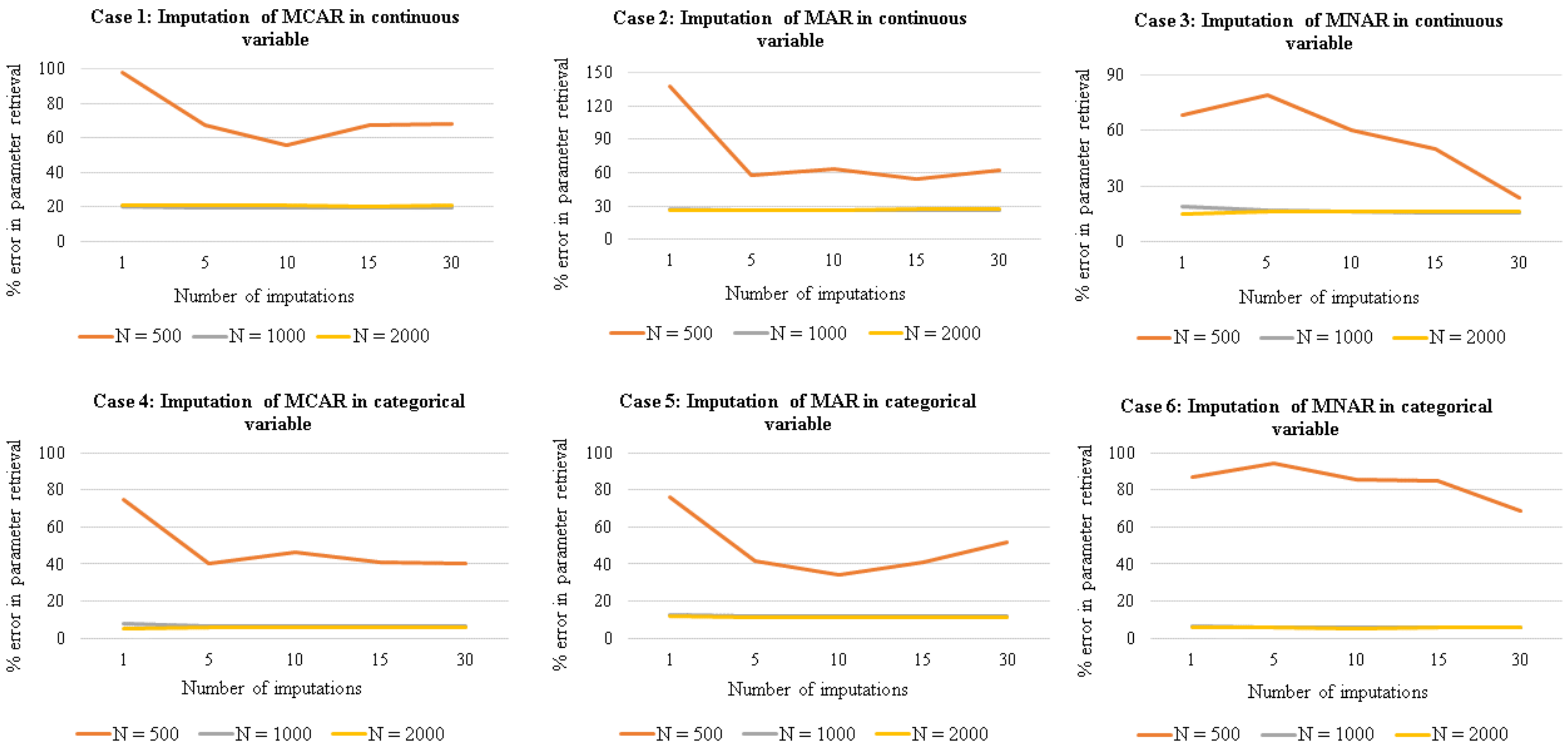
Variable characteristics	**No of repetitions**	**Measure**	**Parameter value**						**Mean error**
			$\boldsymbol{\beta_{1b}}$	$\boldsymbol{\beta_{1c}}$	$\boldsymbol{\beta_{2}}$	$\boldsymbol{\beta_{3}}$	$\boldsymbol{\beta_{4b}}$	$\boldsymbol{\beta_{4c}}$	
True value			*0.99*	*0.48*	*-0.50*	*-1.30*	*-1.00*	*-1.77*	
Case 1: MCAR in continuous variable									
Standard deviation/mean = 0.75 and **10%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.85	0.42	-0.45	-1.03	-0.88	-1.59	13.23
		ω	13.84	12.48	10.06	20.93	12.02	10.04	
	5	$\bar{\alpha}$	0.86	0.43	-0.45	-1.04	-0.90	-1.60	12.31
		ω	13.41	10.73	9.82	20.28	9.97	9.66	
	10	$\bar{\alpha}$	0.86	0.43	-0.45	-1.04	-0.90	-1.60	12.16
		ω	13.37	10.34	9.80	20.13	9.72	9.59	
	15	$\bar{\alpha}$	0.86	0.43	-0.45	-1.04	-0.90	-1.60	12.11
		ω	13.09	10.32	9.81	19.94	9.84	9.66	
	30	$\bar{\alpha}$	0.86	0.43	-0.45	-1.04	-0.90	-1.60	12.08
		ω	13.01	10.30	9.82	19.92	9.74	9.69	
Standard deviation/mean = 0.75 and **20%** missing	1 (SI)	$\bar{\alpha}$	0.75	0.38	-0.41	-0.85	-0.83	-1.49	21.13
		ω	23.85	19.60	16.42	34.57	16.48	15.82	
	5	$\bar{\alpha}$	0.77	0.39	-0.41	-0.83	-0.83	-1.47	21.21
		ω	21.59	18.40	17.10	36.20	16.90	17.09	
	10	$\bar{\alpha}$	0.78	0.39	-0.41	-0.84	-0.83	-1.47	20.86

value in variable x_3		ω	21.28	18.03	17.00	35.59	16.43	16.80	
	15	$\bar{\alpha}$	0.78	0.40	-0.41	-0.84	-0.83	-1.48	20.78
		ω	21.18	17.82	16.93	35.31	16.70	16.74	
	30	$\bar{\alpha}$	0.78	0.39	-0.41	-0.84	-0.83	-1.48	20.85
		ω	21.47	18.09	16.89	35.21	16.74	16.71	
Standard deviation/mean = 0.75 and **30%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.70	0.37	-0.39	-0.71	-0.78	-1.40	26.98
		ω	28.83	23.68	21.00	45.56	21.71	21.11	
	5	$\bar{\alpha}$	0.71	0.37	-0.39	-0.68	-0.77	-1.39	27.53
		ω	28.43	23.01	21.66	47.98	22.36	21.70	
	10	$\bar{\alpha}$	0.71	0.37	-0.39	-0.68	-0.78	-1.39	27.30
		ω	28.07	22.57	21.87	47.65	22.18	21.47	
	15	$\bar{\alpha}$	0.71	0.37	-0.39	-0.69	-0.78	-1.39	27.27
		ω	28.17	22.39	21.83	47.41	22.27	21.54	
	30	$\bar{\alpha}$	0.71	0.37	-0.39	-0.69	-0.78	-1.39	27.29
		ω	28.32	22.72	21.86	47.27	22.09	21.49	
Case 2: MAR in continuous variable									
Standard deviation/mean = 0.75 and **10%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.79	0.38	-0.42	-0.96	-0.80	-1.47	20.16
		ω	20.00	21.11	16.00	26.24	20.60	17.03	
	5	$\bar{\alpha}$	0.81	0.40	-0.42	-1.00	-0.81	-1.49	18.11
		ω	17.89	17.73	14.57	23.07	19.58	15.84	
	10	$\bar{\alpha}$	0.81	0.40	-0.42	-1.00	-0.80	-1.49	18.12
		ω	17.86	17.55	14.59	23.04	19.81	15.87	
	15	$\bar{\alpha}$	0.81	0.40	-0.42	-1.00	-0.80	-1.49	18.24
		ω	17.92	17.96	14.62	23.09	19.92	15.92	
	30	$\bar{\alpha}$	0.81	0.40	-0.42	-1.00	-0.80	-1.49	18.24
		ω	17.98	17.93	14.65	22.94	19.98	15.95	
Standard deviation/mean = 0.75 and **20%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.70	0.35	-0.39	-0.81	-0.76	-1.38	26.75
		ω	28.80	26.94	21.12	37.49	24.14	21.99	
	5	$\bar{\alpha}$	0.71	0.36	-0.39	-0.80	-0.76	-1.38	26.66
		ω	27.92	25.60	21.19	38.36	24.73	22.15	
	10	$\bar{\alpha}$	0.72	0.36	-0.39	-0.81	-0.76	-1.38	26.66
		ω	27.66	25.64	21.31	38.22	24.95	22.20	
	15	$\bar{\alpha}$	0.71	0.36	-0.39	-0.80	-0.75	-1.38	26.83
		ω	28.15	25.79	21.41	38.28	25.10	22.24	
	30	$\bar{\alpha}$	0.71	0.36	-0.39	-0.81	-0.75	-1.38	26.80
		ω	27.93	25.86	21.39	38.11	25.27	22.26	
Standard deviation/mean = 0.75 and **30%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.66	0.34	-0.37	-0.67	-0.73	-1.32	31.63
		ω	33.35	28.87	25.39	48.61	27.71	25.85	
	5	$\bar{\alpha}$	0.67	0.35	-0.37	-0.66	-0.72	-1.32	31.42
		ω	32.55	27.97	25.10	49.10	28.11	25.68	
	10	$\bar{\alpha}$	0.67	0.34	-0.37	-0.66	-0.72	-1.32	31.65
		ω	32.63	28.28	25.19	49.42	28.56	25.80	
	15	$\bar{\alpha}$	0.67	0.34	-0.37	-0.66	-0.72	-1.32	31.78
		ω	32.72	28.54	25.33	49.67	28.66	25.79	
	30	$\bar{\alpha}$	0.67	0.34	-0.37	-0.66	-0.72	-1.32	31.82
		ω	32.72	28.54	25.38	49.55	28.89	25.84	
Case 3: MNAR in continuous variable									
Standard deviation/mean = 0.75 and **10%** missing	1 (SI)	$\bar{\alpha}$	0.88	0.44	-0.46	-1.11	-0.86	-1.63	10.72
		ω	11.05	8.87	6.77	14.59	14.81	8.21	
	5	$\bar{\alpha}$	0.89	0.45	-0.46	-1.14	-0.91	-1.65	8.67
		ω	9.55	7.46	6.57	12.65	9.04	6.73	
	10	$\bar{\alpha}$	0.90	0.45	-0.46	-1.15	-0.92	-1.66	8.22

value in variable x_3		ω	9.20	7.47	6.26	11.87	8.16	6.35	
	15	$\bar{\alpha}$	0.90	0.45	-0.46	-1.15	-0.92	-1.66	8.14
		ω	9.19	7.64	6.19	11.64	7.99	6.21	
	30	$\bar{\alpha}$	0.90	0.45	-0.46	-1.15	-0.92	-1.66	8.23
		ω	9.16	7.66	6.30	11.74	8.16	6.34	
Standard deviation/mean = 0.75 and **20%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.82	0.41	-0.44	-1.02	-0.84	-1.58	15.11
		ω	17.04	14.34	10.73	21.49	16.02	11.03	
	5	$\bar{\alpha}$	0.81	0.41	-0.44	-0.99	-0.83	-1.56	16.41
		ω	17.96	14.64	12.05	24.22	17.35	12.24	
	10	$\bar{\alpha}$	0.81	0.41	-0.44	-1.00	-0.83	-1.56	16.15
		ω	17.84	14.33	11.93	23.62	17.08	12.12	
	15	$\bar{\alpha}$	0.81	0.41	-0.44	-1.00	-0.83	-1.56	16.07
		ω	17.80	14.45	11.93	23.23	16.93	12.11	
	30	$\bar{\alpha}$	0.81	0.41	-0.44	-1.00	-0.83	-1.56	15.91
		ω	17.63	14.24	11.73	22.97	16.84	12.08	
Standard deviation/mean = 0.75 and **30%** missing value in variable x_3	1 (SI)	$\bar{\alpha}$	0.73	0.39	-0.42	-0.90	-0.78	-1.50	21.45
		ω	25.52	19.73	14.81	31.32	22.10	15.21	
	5	$\bar{\alpha}$	0.74	0.39	-0.42	-0.85	-0.78	-1.49	22.21
		ω	25.45	18.52	16.11	34.96	22.08	16.16	
	10	$\bar{\alpha}$	0.74	0.39	-0.41	-0.86	-0.78	-1.49	22.04
		ω	25.20	18.53	16.23	34.28	21.92	16.10	
	15	$\bar{\alpha}$	0.74	0.39	-0.42	-0.86	-0.78	-1.49	21.96
		ω	24.83	18.58	16.16	34.01	22.01	16.20	
	30	$\bar{\alpha}$	0.74	0.39	-0.42	-0.86	-0.78	-1.49	21.93
		ω	24.89	18.35	16.08	33.83	22.25	16.21	
Case 4: MCAR in categorical variable									
Mean = 0.60 and **10%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-1.01	-1.65	3.56
		ω	1.59	4.28	1.53	0.84	6.51	6.59	
	5	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.98	-1.64	3.26
		ω	1.17	3.21	1.49	0.86	5.64	7.19	
	10	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.98	-1.65	3.21
		ω	1.21	3.34	1.43	0.86	5.35	7.06	
	15	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.64	3.27
		ω	1.27	3.49	1.44	0.86	5.44	7.13	
	30	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.65	3.19
		ω	1.23	3.29	1.43	0.84	5.31	7.03	
Mean = 0.60 and **20%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.99	-1.51	5.60
		ω	2.38	5.85	2.66	1.66	6.54	14.52	
	5	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.50	5.94
		ω	2.22	5.37	2.50	1.74	8.74	15.04	
	10	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.51	5.80
		ω	2.09	5.39	2.43	1.73	8.32	14.87	
	15	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.51	5.76
		ω	2.09	5.37	2.39	1.70	8.26	14.76	
	30	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.51	5.79
		ω	2.11	5.31	2.36	1.69	8.51	14.74	
Mean = 0.60 and **30%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.92	-1.36	8.45
		ω	2.97	6.76	3.84	2.58	11.23	23.34	
	5	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.85	-1.35	8.73
		ω	2.62	5.47	3.59	2.63	14.48	23.59	

	10	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.86	-1.36	8.65
		ω	2.60	5.59	3.55	2.61	14.15	23.40	
	15	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.85	-1.35	8.73
		ω	2.64	5.48	3.57	2.59	14.59	23.50	
	30	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.86	-1.36	8.63
		ω	2.59	5.52	3.55	2.57	14.22	23.34	
Case 5: MAR in categorical variable									
Mean = 0.60 and **10%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.97	0.47	-0.49	-1.29	-0.83	-1.51	6.69
		ω	1.63	3.36	1.85	1.42	16.89	14.98	
	5	$\bar{\alpha}$	0.97	0.47	-0.49	-1.29	-0.84	-1.53	6.07
		ω	1.35	2.95	1.53	1.30	15.82	13.48	
	10	$\bar{\alpha}$	0.97	0.47	-0.49	-1.29	-0.84	-1.53	5.95
		ω	1.25	2.72	1.52	1.26	15.55	13.37	
	15	$\bar{\alpha}$	0.97	0.47	-0.49	-1.29	-0.85	-1.54	5.82
		ω	1.24	2.64	1.49	1.22	15.25	13.05	
	30	$\bar{\alpha}$	0.97	0.47	-0.49	-1.29	-0.85	-1.54	5.85
		ω	1.24	3.01	1.48	1.20	15.21	12.96	
Mean = 0.60 and **20%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.69	-1.30	11.93
		ω	2.77	5.88	3.00	2.46	30.66	26.81	
	5	$\bar{\alpha}$	0.97	0.47	-0.48	-1.27	-0.71	-1.32	11.24
		ω	2.21	4.92	2.84	2.31	29.85	25.28	
	10	$\bar{\alpha}$	0.97	0.47	-0.48	-1.27	-0.70	-1.32	11.23
		ω	2.02	4.55	2.87	2.30	30.34	25.29	
	15	$\bar{\alpha}$	0.97	0.47	-0.48	-1.27	-0.70	-1.33	11.14
		ω	2.03	4.51	2.88	2.26	30.07	25.12	
	30	$\bar{\alpha}$	0.97	0.47	-0.48	-1.28	-0.70	-1.33	11.14
		ω	2.10	4.75	2.83	2.22	30.14	24.82	
Mean = 0.60 and **30%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.95	0.46	-0.48	-1.26	-0.58	-1.10	16.48
		ω	3.66	7.48	4.11	3.47	42.45	37.70	
	5	$\bar{\alpha}$	0.96	0.46	-0.48	-1.26	-0.59	-1.13	15.85
		ω	3.16	6.69	4.06	3.33	41.79	36.08	
	10	$\bar{\alpha}$	0.96	0.46	-0.48	-1.26	-0.59	-1.13	15.87
		ω	2.99	6.51	4.09	3.34	42.13	36.13	
	15	$\bar{\alpha}$	0.96	0.46	-0.48	-1.26	-0.59	-1.13	15.73
		ω	3.05	6.27	4.15	3.31	41.66	35.94	
	30	$\bar{\alpha}$	0.96	0.46	-0.48	-1.26	-0.59	-1.13	15.83
		ω	3.12	6.67	4.15	3.31	41.82	35.92	
Case 6: MNAR in categorical variable									
Mean = 0.60 and **10%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.98	-1.67	2.70
		ω	1.29	3.27	1.06	0.93	3.71	5.94	
	5	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.64	2.91
		ω	1.15	3.11	1.20	1.04	3.50	7.46	
	10	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.64	2.91
		ω	1.15	3.24	1.25	1.03	3.35	7.47	
	15	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.63	3.02
		ω	1.19	3.29	1.34	1.08	3.32	7.91	
	30	$\bar{\alpha}$	0.98	0.47	-0.49	-1.29	-0.97	-1.63	3.07
		ω	1.21	3.42	1.40	1.09	3.34	7.96	
Mean = 0.60 and **20%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.94	-1.51	5.83
		ω	2.45	5.63	2.49	2.08	7.68	14.67	
	5	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.94	-1.48	5.91
		ω	2.30	5.12	2.58	2.17	6.57	16.70	

	10	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.48	5.81
		ω	2.23	5.00	2.56	2.13	6.57	16.36	
	15	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.48	5.91
		ω	2.25	4.99	2.66	2.16	6.71	16.69	
	30	$\bar{\alpha}$	0.97	0.46	-0.48	-1.28	-0.93	-1.47	5.96
		ω	2.26	5.02	2.72	2.19	6.57	16.97	
Mean = 0.60 and **30%** missing value in variable x_4	1 (SI)	$\bar{\alpha}$	0.96	0.46	-0.48	-1.27	-0.91	-1.33	8.73
		ω	3.35	7.35	3.58	2.98	10.40	24.71	
	5	$\bar{\alpha}$	0.96	0.45	-0.48	-1.26	-0.91	-1.30	8.90
		ω	3.41	7.16	3.70	3.07	9.14	26.90	
	10	$\bar{\alpha}$	0.96	0.45	-0.48	-1.26	-0.91	-1.31	8.76
		ω	3.29	7.04	3.65	3.04	9.28	26.27	
	15	$\bar{\alpha}$	0.96	0.45	-0.48	-1.26	-0.91	-1.30	8.89
		ω	3.30	7.06	3.80	3.08	9.35	26.77	
	30	$\bar{\alpha}$	0.96	0.46	-0.48	-1.26	-0.91	-1.30	8.84
		ω	3.26	7.04	3.87	3.09	9.00	26.80	


Case 1: Imputation of MCAR in continuous variable

Case 2: Imputation of MAR in continuous variable

Case 3: Imputation of MNAR in continuous variable

Case 4: Imputation of MCAR in categorical variable

Case 5: Imputation of MAR in categorical variable

Case 6: Imputation of MNAR in categorical variable

Figure 1 Comparison between single and multiple imputations across sample sizes

Comparison between imputed data and complete case data model

After establishing that MI with 5 imputations offers improved results relative to SI approach, we focus on the comparison of the MI with CCD. In multiple empirical contexts, researchers develop a model with CCD by either dropping all records with missing values (row elimination) or dropping the variable completely (column elimination). The most common approach to missing data employs either a row elimination or column elimination approach. In contexts with a small share of missing values (<2%), it is common to employ row elimination while column elimination is employed when data for a variable is missing for a large share (>20% missing). The current exercise is designed to examine when MI approaches are more useful compared to using CCD approaches. The MI and CCD model performances are compared based on log-likelihood improvement and parameter retrieval error. The log-likelihood (LL) improvement is computed as follows:

$$LL\ improvement = LL_{Q3} - LL_{Q2} \tag{9}$$

where, LL_{Q2} and LL_{Q3} are the log-likelihood of the CCD model and imputed data model respectively. The reader would note that comparing MI and CCD approach with row elimination offers a challenge due to the difference in the number of records across the dataset. To address this, we evaluate the model fit using CCD parameters on the full dataset. As the CCD model is developed on a subset of records (after removing missing records), the LL from this model is not directly comparable to an imputation model generated on the full sample. So, we use the CCD model and predict the model outcome for all records in the full dataset to generate an equivalent log-likelihood. This eliminates the difference in number of records and provides a way to see how effective CCD is on the full dataset. The CCD case with column elimination does not require any adaptation.

The parameter retrieval of the CCD and imputed data models are calculated as follows:

$$\varphi_{Q2} = abs\left(\frac{(\beta_{v,Q2} - \beta_{v,Q1})}{\beta_{v,Q1}}\right) * 100 \tag{10}$$

$$\varphi_{Q3} = abs\left(\frac{(\beta_{v,Q3} - \beta_{v,Q1})}{\beta_{v,Q1}}\right) * 100 \tag{11}$$

where, φ_{Q2} and φ_{Q3} are the errors in parameter retrieval by the CCD and imputed data model respectively, $\beta_{v,Q1}$, $\beta_{v,Q2}$, and $\beta_{v,Q3}$ are the vectors of parameters of v variables for original data, CCD, and imputed data model respectively. The simulation analysis and experiment were conducted using RStudio software.

Average log-likelihood (LL) improvements are shown in Figure 2 and 3, and percentage error in parameter retrieval is presented in Table 3. In generating the results, we tested several possible variations of the independent variables (SD/mean or mean). However, as all of the different variations offered similar results, we present results for only one variation. Table 3 presents the results for SD/mean = 0.75 for continuous variable and mean = 0.6 for categorical variable. The results for missing values created in a random manner (MAR) in both continuous and categorical data are shown in the table.

Several observations can be drawn from Figure 2 and 3 and Table 3. For continuous variables, we observe that CCD model with row elimination provides better model fit than imputed data model. On the other hand, CCD model with column elimination consistently underperforms the imputed model. For categorical variables, it can be observed in Figure 3 that for the MCAR and MNAR cases, with the increase in missing share the imputed model performs better than the CCD (by row elimination) model. In terms of column elimination, across all scenarios, imputed data models outperform CCD models. The model fit improvement of the imputed data model over CCD model across different sample sizes are shown in Figure 4. It can be observed that the difference of log-likelihood between the two models are found to increase with an increase in sample size.

Further, to assess the variation of imputation bias with varying sample sizes, the parameter retrieval efficiency of the imputed data model was compared with that of the CCD model considering different sample sizes and the outcomes are shown in Figure 5. It is noticeable that, in the cases of categorical variable, across all three types of missing values, multiple imputation provides less biased parameters. However, in the case of continuous variable, data imputation performs better for the case of MCAR in a smaller dataset (N=500) and MNAR in a smaller dataset with large share of missing values.

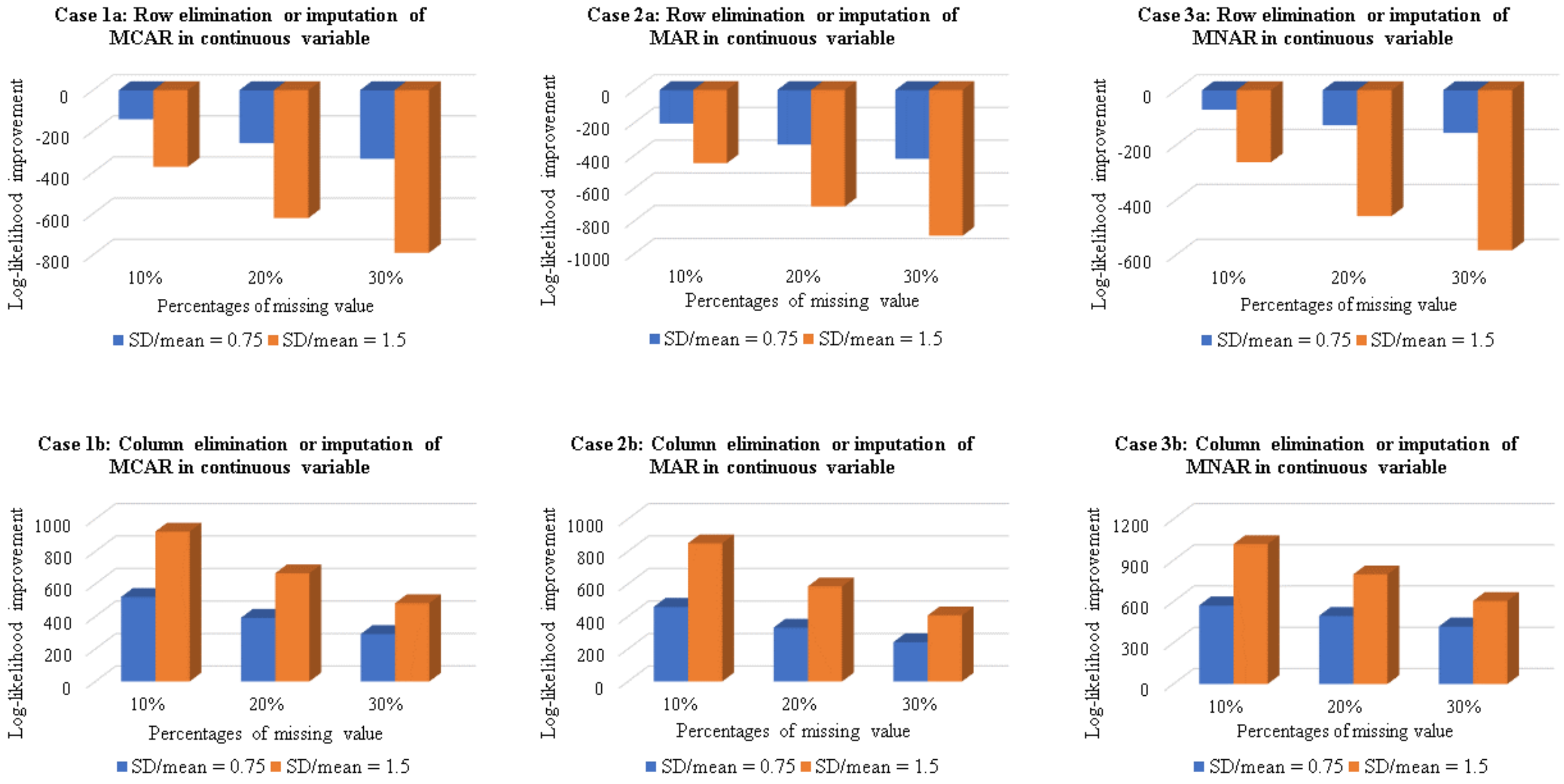
In addition to comparing model fit and parameter retrieval efficiency measures, it is also useful to examine how marginal effects of the variables vary across the models from different datasets. A comparison between the efficiency of the CCD model and the imputed data model in retrieving the marginal effects of the original data model are summarized in Figure 6. The results offer interesting insights. For continuous variables, we observe that complete case data models perform better than imputed data models for the MAR scenarios. However, with increasing sample size, the difference between complete case data and imputed data models becomes small. In the MCAR context, imputed data model outperforms CCD model for smaller dataset with small share of missing values while in the MNAR scenario, the differences between CCD and imputed models are small. For categorical variables, we observe that imputed data models consistently outperform CCD models. For larger samples, both approaches offer very similar errors. These results offer important implications for empirical research. From our analysis, we notice that when the sample sizes are 1000 and above, CCD approach performs slightly better than the multiple imputation approach for continuous variables at all three missing percentages. For categorical variables, multiple imputation offers better results relative to CCD approach for all three missing percentages. However, the differences become smaller for datasets with more than 1000 records. Hence, for large datasets (>1000 records), it might be beneficial to simply develop a CCD model with row elimination as opposed to developing imputation models. However, when the share of missing data warrants column elimination, it is important and even necessary that multiple imputation approach be employed for model development. Finally, the parameter retrieval efficiency of the imputed data model and the CCD (both by row and column elimination) model in the random parameters MNL framework is evaluated in our study. The percentages error in parameter retrieval of the fixed parameters MNL model and random parameters MNL model are presented in Table 4 and 5 for all the scenarios. The new parameter in the random parameters MNL model is considered based on one additional mixing parameter (standard deviation of β_2). The results indicate a consistency in findings identified in the previous analysis across all scenarios considered in our experiment. The behavior of the imputed data model and the CCD model in both modeling frameworks is very similar. It can be observed that, for both cases, the CCD model with row elimination provides better performance than the imputed data model while the CCD model with column elimination consistently underperforms the imputed model. This implies that the conclusions drawn from our simulation experiment with fixed parameters MNL model are reliable for random parameter variants of the frameworks.

TABLE 3 Comparison between imputed data and complete case data model in retrieving true parameter

β	**Percentage error in coefficient retrieval at different percentage of missing values**				
	True parameter	**Imputed data model parameter**	**CCD model parameter**	**Error in imputed model**	**Error in CCD model**
Row elimination or imputation for MAR in continuous variable					
SD/mean = 0.75 and 10% missing value					
β_{1b}	0.99	0.81	0.98	17.89	5.59
β_{1c}	0.48	0.40	0.48	17.73	10.08
β_2	-0.50	-0.42	-0.49	14.57	2.86
β_3	-1.30	-1.00	-1.30	23.07	1.13
β_{4b}	-1.00	-0.81	-0.98	19.58	5.39
β_{4c}	-1.77	-1.49	-1.77	15.84	2.59
Mean error				**18.11**	**4.60**

SD/mean = 0.75 and 20% missing value					
β_{1b}	0.99	0.71	0.99	27.92	8.67
β_{1c}	0.48	0.36	0.48	25.60	13.57
β_2	-0.50	-0.39	-0.49	21.19	3.67
β_3	-1.30	-0.80	-1.30	38.36	1.77
β_{4b}	-1.00	-0.76	-0.97	24.73	8.22
β_{4c}	-1.77	-1.38	-1.76	22.15	3.80
Mean error				**26.66**	**6.62**
SD/mean = 0.75 and 30% missing value					
β_{1b}	0.99	0.67	0.99	32.55	12.43
β_{1c}	0.48	0.35	0.48	27.97	23.99
β_2	-0.50	-0.37	-0.50	25.10	4.53
β_3	-1.30	-0.66	-1.30	49.10	2.12
β_{4b}	-1.00	-0.72	-0.99	28.11	13.65
β_{4c}	-1.77	-1.32	-1.78	25.68	7.27
Mean error				**31.42**	**10.66**
Column elimination or imputation for MAR in continuous variable					
SD/mean = 0.75 and 10% missing value					
β_{1b}	0.99	0.81	0.56	17.89	43.35
β_{1c}	0.48	0.40	0.31	17.73	34.08
β_2	-0.50	-0.42	-0.33	14.57	33.57
β_3	-1.30	-1.00	0.00	23.07	100.00
β_{4b}	-1.00	-0.81	-0.68	19.58	32.25
β_{4c}	-1.77	-1.49	-1.22	15.84	31.49
Mean error				**18.11**	**45.79**
SD/mean = 0.75 and 20% missing value					
β_{1b}	0.99	0.71	0.56	27.92	43.35
β_{1c}	0.48	0.36	0.31	25.60	34.08
β_2	-0.50	-0.39	-0.33	21.19	33.57
β_3	-1.30	-0.80	0.00	38.36	100.00
β_{4b}	-1.00	-0.76	-0.68	24.73	32.25
β_{4c}	-1.77	-1.38	-1.22	22.15	31.49
Mean error				**26.66**	**45.79**
SD/mean = 0.75 and 30% missing value					
β_{1b}	0.99	0.67	0.56	32.55	43.35
β_{1c}	0.48	0.35	0.31	27.97	34.08
β_2	-0.50	-0.37	-0.33	25.10	33.57
β_3	-1.30	-0.66	0.00	49.10	100.00
β_{4b}	-1.00	-0.72	-0.68	28.11	32.25
β_{4c}	-1.77	-1.32	-1.22	25.68	31.49
Mean error				**31.42**	**45.79**
Row elimination or imputation for MAR in categorical variable					
Mean = 0.6 and 10% missing value					
β_{1b}	0.99	0.97	0.99	1.35	5.77
β_{1c}	0.48	0.47	0.47	2.95	9.00
β_2	-0.50	-0.49	-0.49	1.53	2.29
β_3	-1.30	-1.29	-1.30	1.30	0.65

β_{4b}	-1.00	-0.84	-1.00	15.82	4.51
β_{4c}	-1.77	-1.53	-1.76	13.48	1.41
Men error				**6.07**	**3.94**
Mean = 0.6 and 20% missing value					
β_{1b}	0.99	0.97	0.99	2.21	6.28
β_{1c}	0.48	0.47	0.47	4.92	11.37
β_2	-0.50	-0.48	-0.49	2.84	3.31
β_3	-1.30	-1.27	-1.31	2.31	1.28
β_{4b}	-1.00	-0.71	-0.99	29.85	7.56
β_{4c}	-1.77	-1.32	-1.76	25.28	2.92
Mean error				**11.24**	**5.45**
Mean = 0.6 and 30% missing value					
β_{1b}	0.99	0.96	0.97	3.16	9.80
β_{1c}	0.48	0.46	0.47	6.69	17.32
β_2	-0.50	-0.48	-0.49	4.06	5.01
β_3	-1.30	-1.26	-1.31	3.33	1.88
β_{4b}	-1.00	-0.59	-0.99	41.79	11.77
β_{4c}	-1.77	-1.13	-1.76	36.08	3.31
Mean error				**15.85**	**8.18**
Column elimination or imputation for MAR in categorical variable					
Mean = 0.6 and 10% missing value					
β_{1b}	0.99	0.97	0.93	1.35	5.91
β_{1c}	0.48	0.47	0.44	2.95	11.18
β_2	-0.50	-0.49	-0.46	1.53	7.97
β_3	-1.30	-1.29	-1.22	1.30	6.22
β_{4b}	-1.00	-0.84	0.00	15.82	100.00
β_{4c}	-1.77	-1.53	0.00	13.48	100.00
Mean error				**6.07**	**38.55**
Mean = 0.6 and 20% missing value					
β_{1b}	0.99	0.97	0.93	2.21	5.91
β_{1c}	0.48	0.47	0.44	4.92	11.18
β_2	-0.50	-0.48	-0.46	2.84	7.97
β_3	-1.30	-1.27	-1.22	2.31	6.22
β_{4b}	-1.00	-0.71	0.00	29.85	100.00
β_{4c}	-1.77	-1.32	0.00	25.28	100.00
Mean error				**11.24**	**38.55**
Mean = 0.6 and 30% missing value					
β_{1b}	0.99	0.96	0.93	3.16	5.91
β_{1c}	0.48	0.46	0.44	6.69	11.18
β_2	-0.50	-0.48	-0.46	4.06	7.97
β_3	-1.30	-1.26	-1.22	3.33	6.22
β_{4b}	-1.00	-0.59	0.00	41.79	100.00
β_{4c}	-1.77	-1.13	0.00	36.08	100.00
Mean error				**15.85**	**38.55**



Case 1a: Row elimination or imputation of MCAR in continuous variable

Case 2a: Row elimination or imputation of MAR in continuous variable

Case 3a: Row elimination or imputation of MNAR in continuous variable

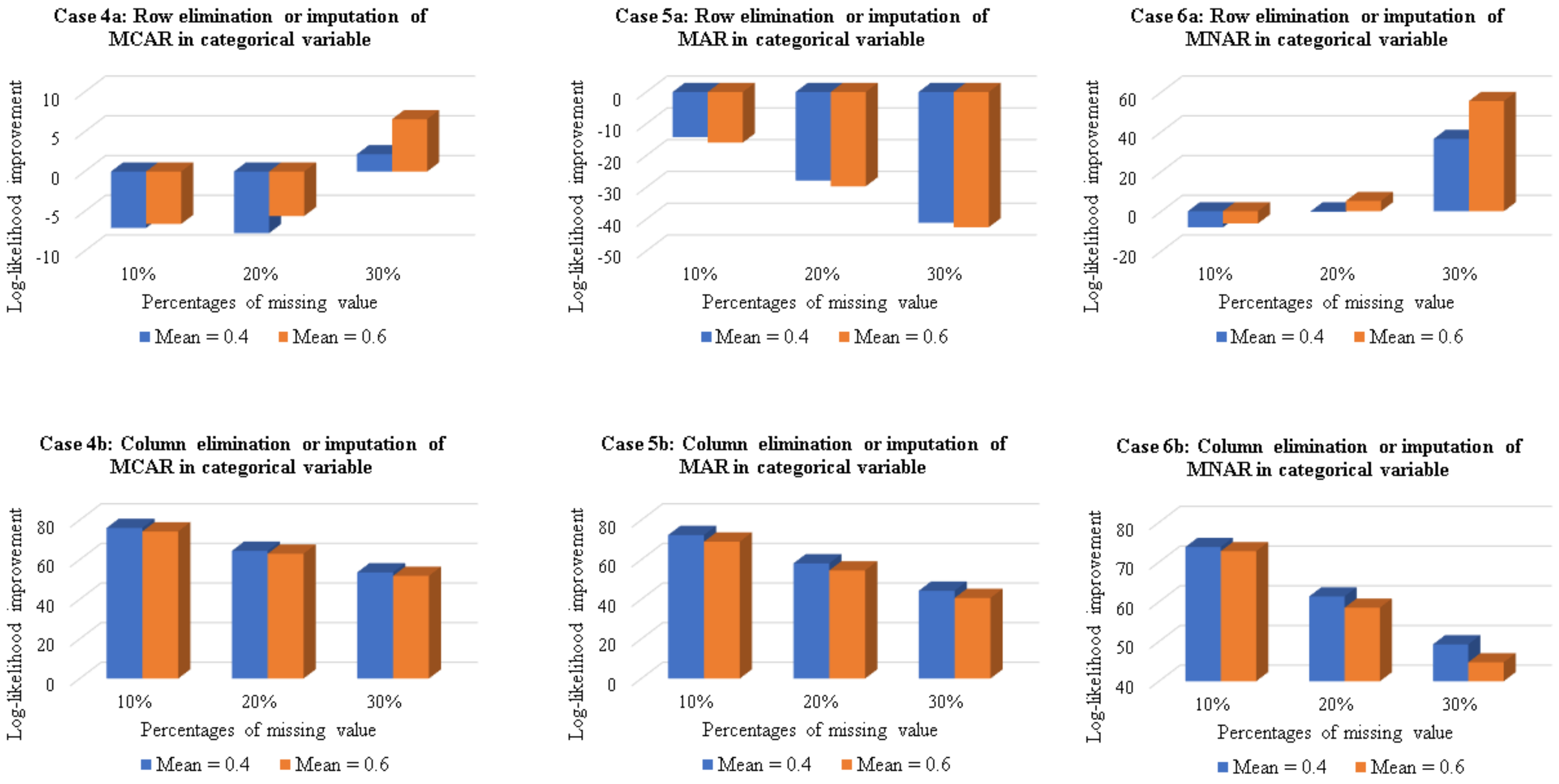
Case 1b: Column elimination or imputation of MCAR in continuous variable

Case 2b: Column elimination or imputation of MAR in continuous variable

Case 3b: Column elimination or imputation of MNAR in continuous variable

Log-likelihood improvement
1200
900
600
300
0
10%
20%
30%
Percentages of missing value
🟦 SD/mean = 0.75 🟧 SD/mean = 1.5

Figure 2 Log-likelihood improvement by data imputation for missing values in continuous variables



Case 4a: Row elimination or imputation of MCAR in categorical variable

Case 5a: Row elimination or imputation of MAR in categorical variable

Case 6a: Row elimination or imputation of MNAR in categorical variable

Case 4b: Column elimination or imputation of MCAR in categorical variable

Case 5b: Column elimination or imputation of MAR in categorical variable

Case 6b: Column elimination or imputation of MNAR in categorical variable

Figure 3 Log-likelihood improvement by data imputation for missing values in categorical variables

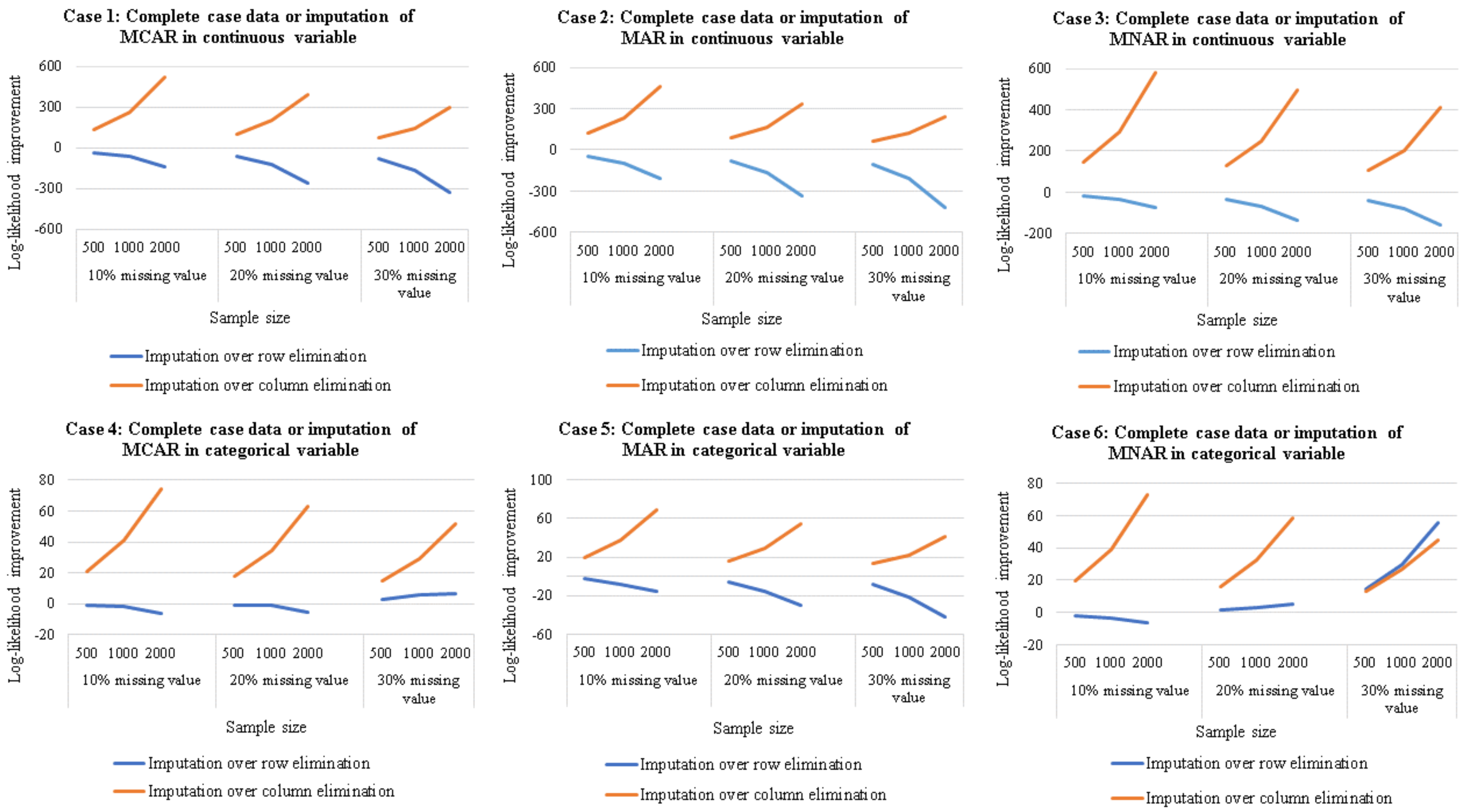


Figure 4 Comparison of log-likelihood improvements across sample sizes

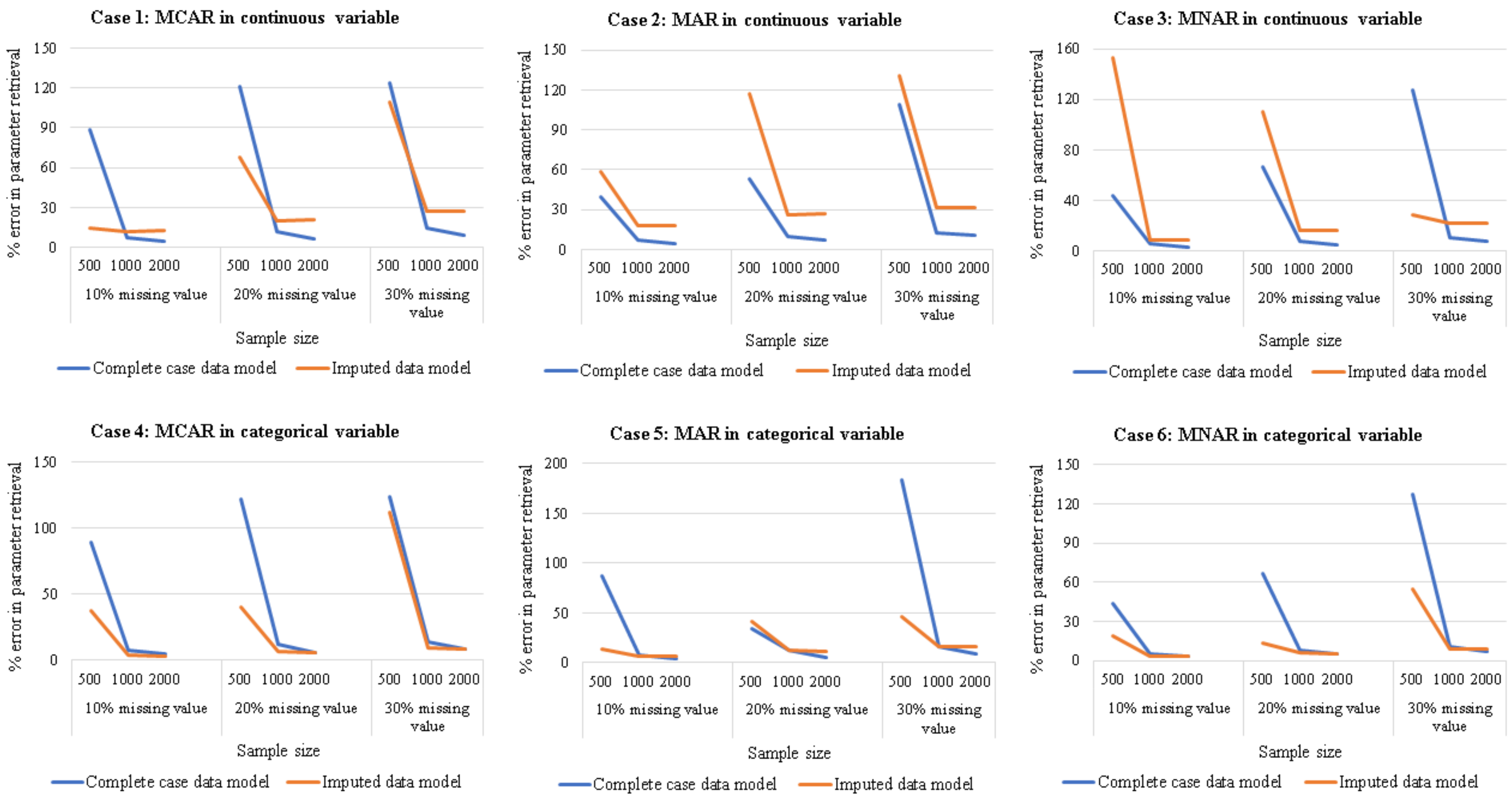


Figure 5 Comparison of the parameter retrieval efficiency between complete case data and imputed data model

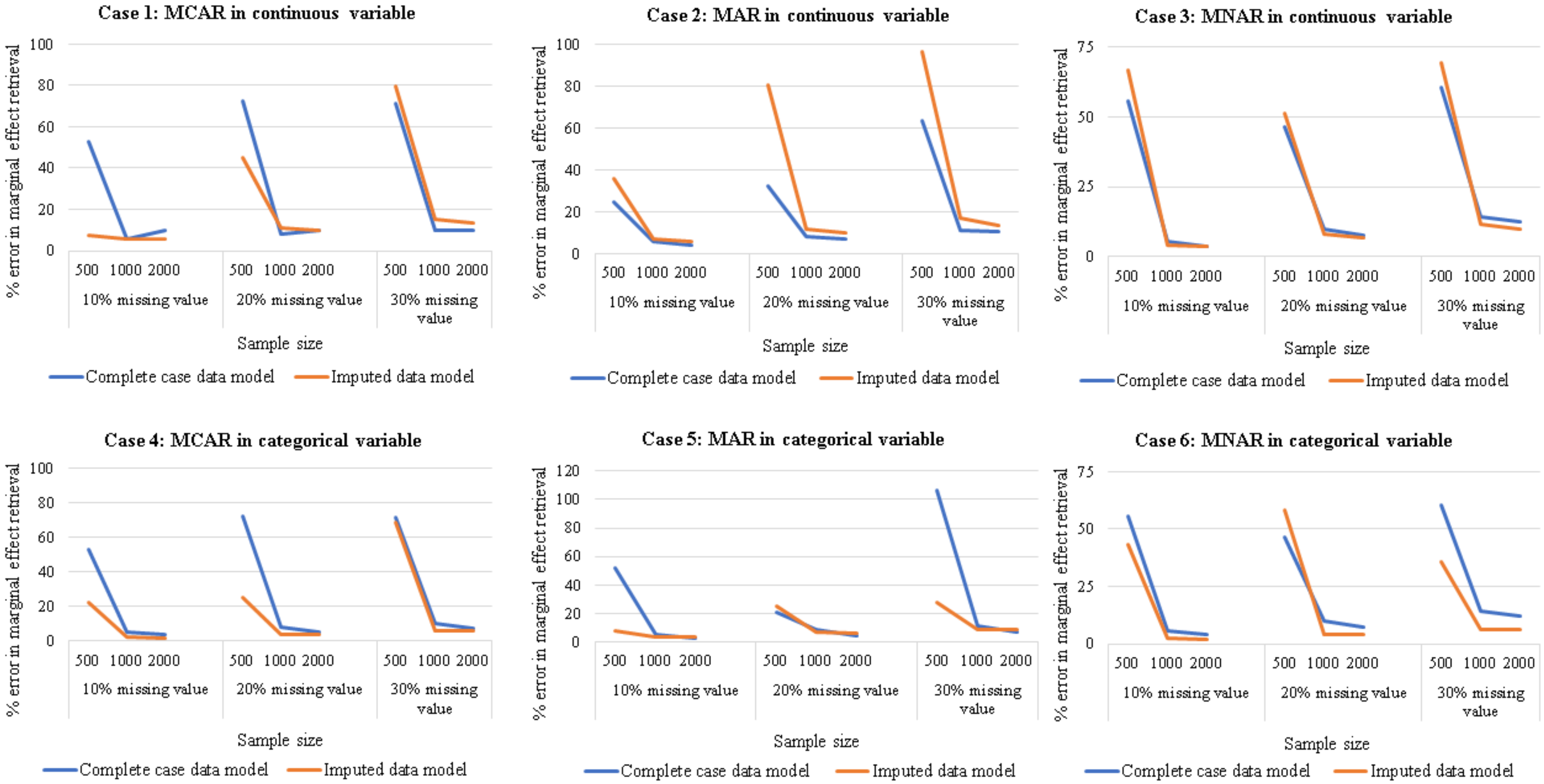


Figure 6 Comparison of the marginal effect retrieval efficiency between complete case data and imputed data model

TABLE 4 Parameter retrieval efficiency of fixed parameters MNL model (Sample size = 2000)

β	**Percentage error in coefficient retrieval at different percentage of missing values**				
	True parameter	**Imputed data model parameter**	**CCD model parameter**	**Error in imputed model**	**Error in CCD model**
Case 1a: Row elimination or imputation of MCAR in continuous variable					
β_{1b}	0.99	0.77	1.04	21.59	8.88
β_{1c}	0.48	0.39	0.50	18.40	7.76
β_{2}	-0.50	-0.41	-0.52	17.10	5.27
β_{3}	-1.30	-0.83	-1.34	36.20	3.22
β_{4b}	-1.00	-0.83	-1.05	16.90	8.27
β_{4c}	-1.77	-1.47	-1.83	17.09	4.09
Mean error				**21.21**	**6.25**
Case 1b: Column elimination or imputation of MCAR in continuous variable					
β_{1b}	0.99	0.77	0.56	21.59	43.35
β_{1c}	0.48	0.39	0.31	18.40	34.08
β_{2}	-0.50	-0.41	-0.33	17.10	33.57
β_{3}	-1.30	-0.83	0.00	36.20	100.00
β_{4b}	-1.00	-0.83	-0.68	16.90	32.25
β_{4c}	-1.77	-1.47	-1.22	17.09	31.49
Mean error				**21.21**	**45.79**
Case 2a: Row elimination or imputation of MAR in continuous variable					
β_{1b}	0.99	0.71	0.99	27.92	8.67
β_{1c}	0.48	0.36	0.48	25.60	13.57
β_{2}	-0.50	-0.39	-0.49	21.19	3.67
β_{3}	-1.30	-0.80	-1.30	38.36	1.77
β_{4b}	-1.00	-0.76	-0.97	24.73	8.22
β_{4c}	-1.77	-1.38	-1.76	22.15	3.80
Mean error				**26.66**	**6.62**
Case 2b: Column elimination or imputation of MAR in continuous variable					
β_{1b}	0.99	0.71	0.56	27.92	43.35
β_{1c}	0.48	0.36	0.31	25.60	34.08
β_{2}	-0.50	-0.39	-0.33	21.19	33.57
β_{3}	-1.30	-0.80	0.00	38.36	100.00
β_{4b}	-1.00	-0.76	-0.68	24.73	32.25
β_{4c}	-1.77	-1.38	-1.22	22.15	31.49
Mean error				**26.66**	**45.79**
Case 3a: Row elimination or imputation of MNAR in continuous variable					
β_{1b}	0.99	0.82	0.99	17.38	4.63
β_{1c}	0.48	0.41	0.47	14.42	14.70
β_{2}	-0.50	-0.43	-0.50	12.63	3.22
β_{3}	-1.30	-0.98	-1.31	24.53	1.39
β_{4b}	-1.00	-0.83	-0.99	17.10	3.77
β_{4c}	-1.77	-1.56	-1.78	12.23	2.56
Mean error				**16.38**	**5.04**

Case 3b: Column elimination or imputation of MNAR in continuous variable					
β_{1b}	0.99	0.82	0.56	17.38	43.35
β_{1c}	0.48	0.41	0.31	14.42	34.08
β_2	-0.50	-0.43	-0.33	12.63	33.57
β_3	-1.30	-0.98	0.00	24.53	100.00
β_{4b}	-1.00	-0.83	-0.68	17.10	32.25
β_{4c}	-1.77	-1.56	-1.22	12.23	31.49
Mean error				**16.38**	**45.79**

TABLE 5 Parameter retrieval efficiency of random parameters MNL model (Sample size = 2000)

β	**Percentage error in coefficient retrieval at different percentage of missing values**				
	True parameter	**Imputed data model parameter**	**CCD model parameter**	**Error in imputed model**	**Error in CCD model**
Case 1a: Row elimination or imputation of MCAR in continuous variable					
β_{1b}	0.94	0.76	0.98	18.94	14.34
β_{1c}	0.49	0.41	0.48	19.33	10.57
β_2	1.50	1.21	1.55	19.53	3.14
SD of β_2	0.80	0.56	0.80	31.97	7.39
β_3	-1.31	-0.84	-1.34	35.71	2.50
β_{4b}	-0.98	-0.79	-1.02	18.77	10.70
β_{4c}	-1.82	-1.50	-1.86	17.75	4.17
Mean error				**23.14**	**7.54**
Case 1b: Column elimination or imputation of MCAR in continuous variable					
β_{1b}	0.94	0.76	0.58	18.94	37.28
β_{1c}	0.49	0.41	0.34	19.33	34.30
β_2	1.50	1.21	1.02	19.53	32.05
SD of β_2	0.80	0.56	0.49	31.97	39.42
β_3	-1.31	-0.84	0.00	35.71	100.00
β_{4b}	-0.98	-0.79	-0.63	18.77	35.34
β_{4c}	-1.82	-1.50	-1.25	17.75	31.25
Mean error				**23.14**	**44.24**
Case 2a: Row elimination or imputation of MAR in continuous variable					
β_{1b}	0.94	0.70	0.94	26.07	10.41
β_{1c}	0.49	0.38	0.49	24.78	12.54
β_2	1.50	1.15	1.50	23.23	2.96
SD of β_2	0.80	0.51	0.79	37.80	8.85
β_3	-1.31	-0.80	-1.30	38.45	2.54
β_{4b}	-0.98	-0.72	-0.98	26.35	16.29
β_{4c}	-1.82	-1.41	-1.83	22.59	5.74
Mean error				**28.47**	**8.48**
Case 2b: Column elimination or imputation of MAR in continuous variable					
β_{1b}	0.94	0.70	0.58	26.07	37.31
β_{1c}	0.49	0.38	0.34	24.78	34.29

β_2	1.50	1.15	1.02	23.23	31.94
SD of β_2	0.80	0.51	0.47	37.80	42.73
β_3	-1.31	-0.80	0.00	38.45	100.00
β_{4b}	-0.98	-0.72	-0.63	26.35	35.28
β_{4c}	-1.82	-1.41	-1.25	22.59	31.19
Mean error				**28.47**	**44.68**
Case 3a: Row elimination or imputation of MNAR in continuous variable					
β_{1b}	0.94	0.81	0.94	14.04	3.42
β_{1c}	0.49	0.44	0.47	12.28	11.18
β_2	1.50	1.39	1.51	7.35	2.05
SD of β_2	0.80	0.78	0.84	5.12	7.40
β_3	-1.31	-1.07	-1.32	18.05	1.40
β_{4b}	-0.98	-0.86	-0.97	13.04	6.07
β_{4c}	-1.82	-1.65	-1.82	9.23	2.93
Mean error				**11.30**	**4.92**
Case 3b: Column elimination or imputation of MNAR in continuous variable					
β_{1b}	0.94	0.81	0.58	14.04	37.27
β_{1c}	0.49	0.44	0.34	12.28	34.32
β_2	1.50	1.39	1.02	7.35	32.09
SD of β_2	0.80	0.78	0.49	5.12	40.45
β_3	-1.31	-1.07	0.00	18.05	100.00
β_{4b}	-0.98	-0.86	-0.63	13.04	35.35
β_{4c}	-1.82	-1.65	-1.25	9.23	31.27
Mean error				**11.30**	**44.39**

EMPIRICAL APPLICATION OF MULTIPLE IMPUTATION

The research team processed data compiled by the Strategic Highway Research Program-2 (SHRP2) including Naturalistic driving data, and Roadway Information Database (RID). The RID crash data from 2011 through 2013 was used for crash prone segment and intersection selection model. The data was obtained from 6 US cities – Bloomington, State College, Tampa Bay, Buffalo, Durham, and Seattle. The dataset contains the record of 857 segment crashes with 73,383 segments, and 129 intersection crashes with 10,782 intersections. Further, the Crash Report Sampling System (CRSS) database was employed in our analysis for crash type and severity model. The database contains a record of 197,092 crashes with 362,596 vehicles and 361,792 drivers from the entire USA.

From the data sources, the research team recognized the presence of a large set of variables with missing records. Given the unique mechanism employed to collect the data, the research team wanted to maximize the consideration of as many records as possible. This was the main motivation for our research into MI approaches. In this section, we document our efforts in estimation models for different components of SHRP2 and CRSS data with missing records. Drawing on the insights from our data simulation, we employ MI approaches for developing models for (a) crash prone segment selection, (b) crash prone intersection selection, (c) crash type and (d) crash severity.

Descriptive statistics of the variables utilized in the models are presented in Figure 7 and 8, and the variables with missing cases are listed in Figure 9. The missing values are generated following the Equation 4 and 5 for continuous and categorical variable respectively. As established in the simulation experiment, a multiple imputation approach of five repetitions is utilized for data imputation and subsequent inference. Most of the variables with missing values are imputed considering their statistical parameters (mean and standard deviation). However, in case of a few variables, the imputation is performed considering their distribution under different categories of another control variable(s). A list of control variables is provided in Figure 9. None refers to no control variable for the variable. The data imputation procedures are customized for the different levels of the control variables i.e., each attribute level is associated with a different rule for data imputation as explained in Figure 10 and 11.

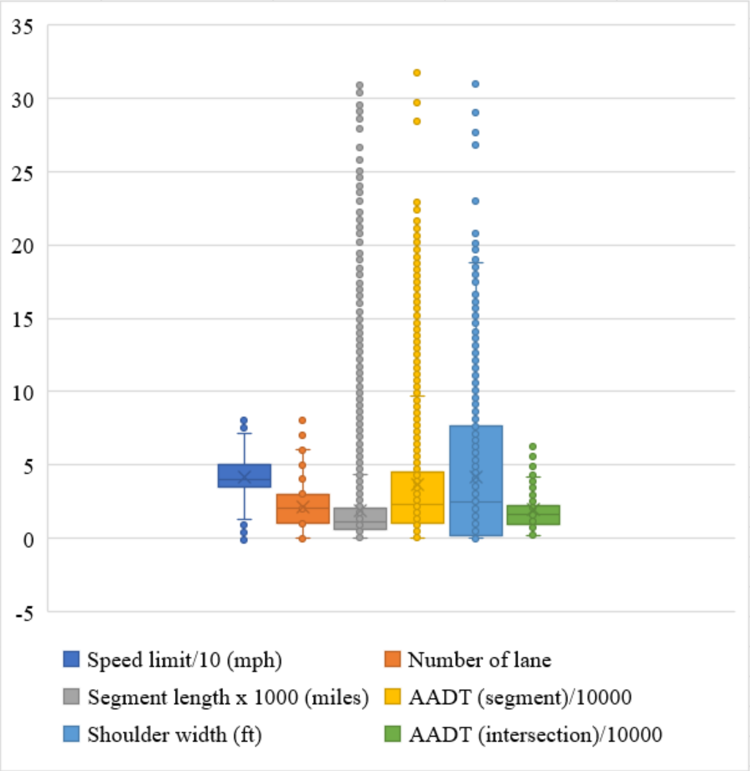


Figure 7 Distribution of continuous variables

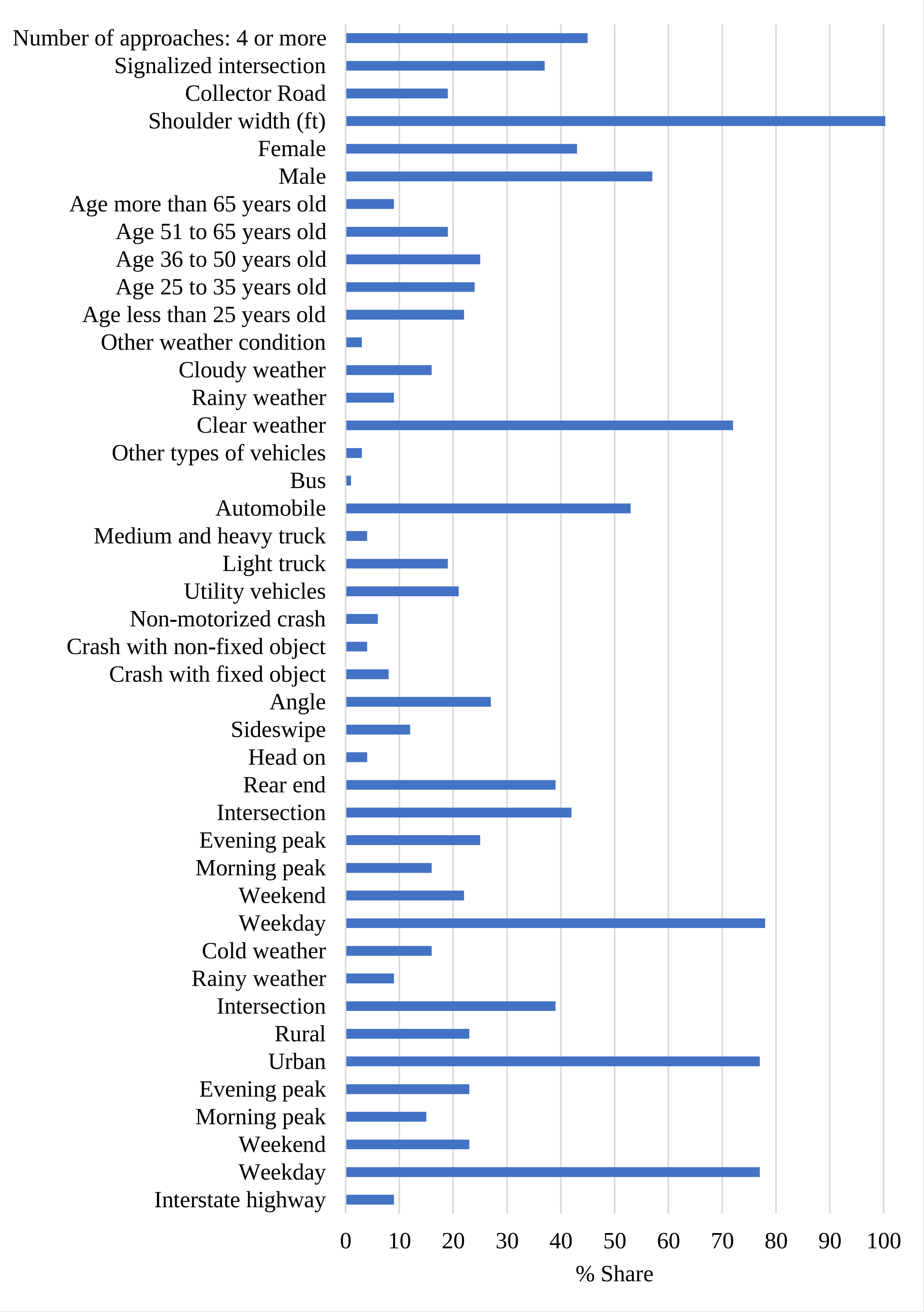


Figure 8 Distribution of categorical dummy variables

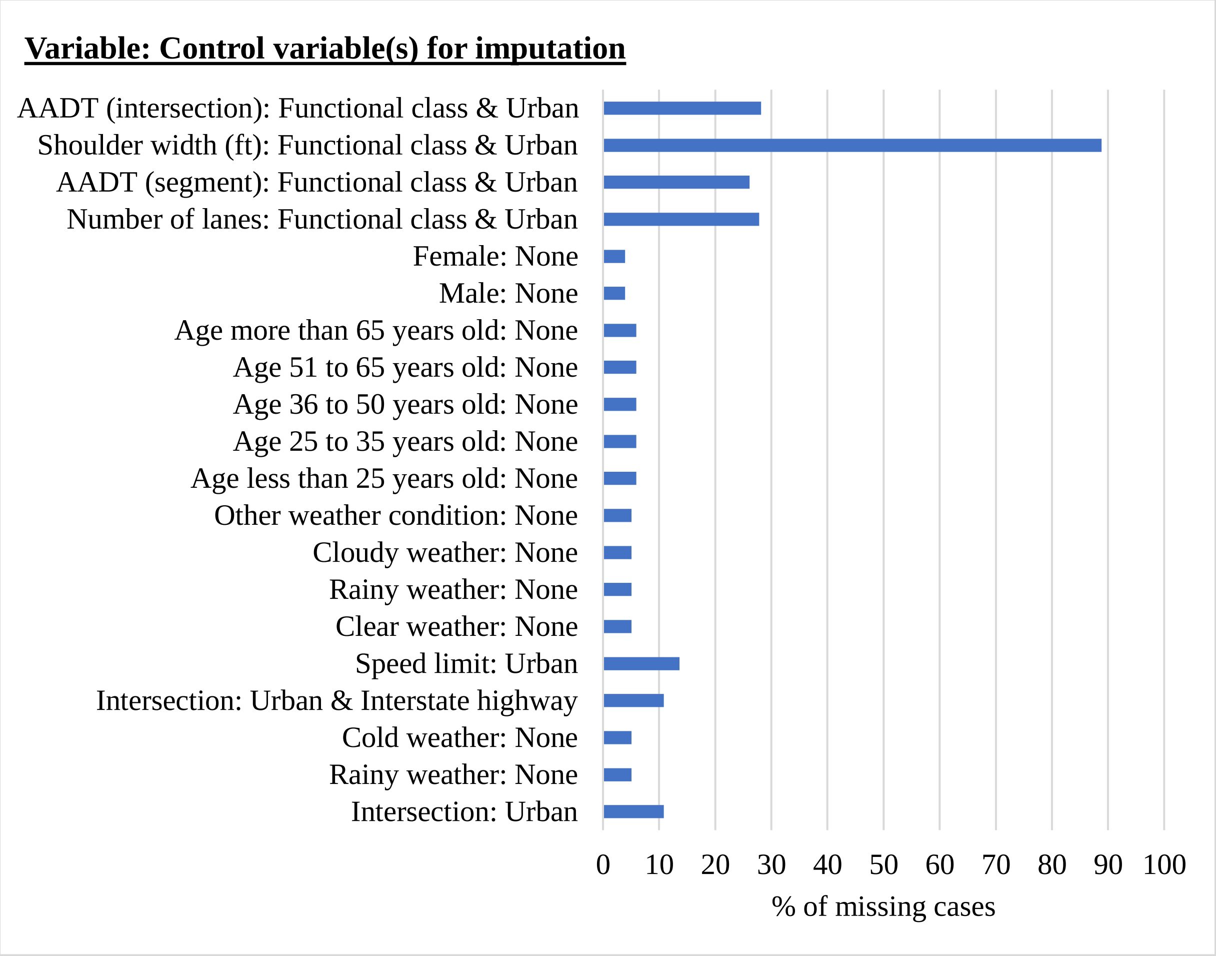
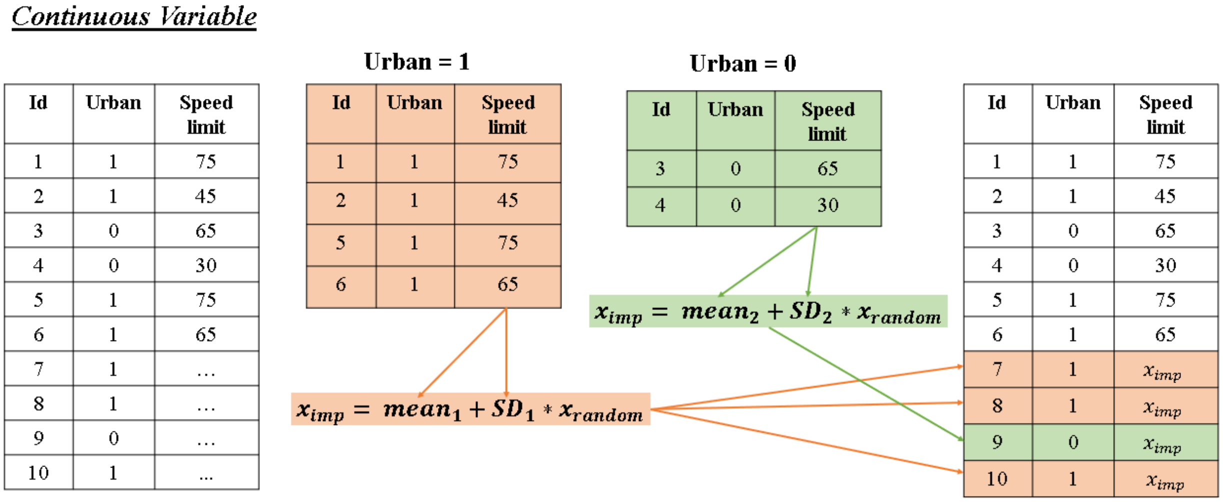


Figure 9 Variables with missing cases along with the corresponding control variable(s)



Id	Urban	Speed limit
1	1	75
2	1	45
3	0	65
4	0	30
5	1	75
6	1	65
7	1	…
8	1	…
9	0	…
10	1	...

Id	Urban	Speed limit
1	1	75
2	1	45
5	1	75
6	1	65

Id	Urban	Speed limit
3	0	65
4	0	30

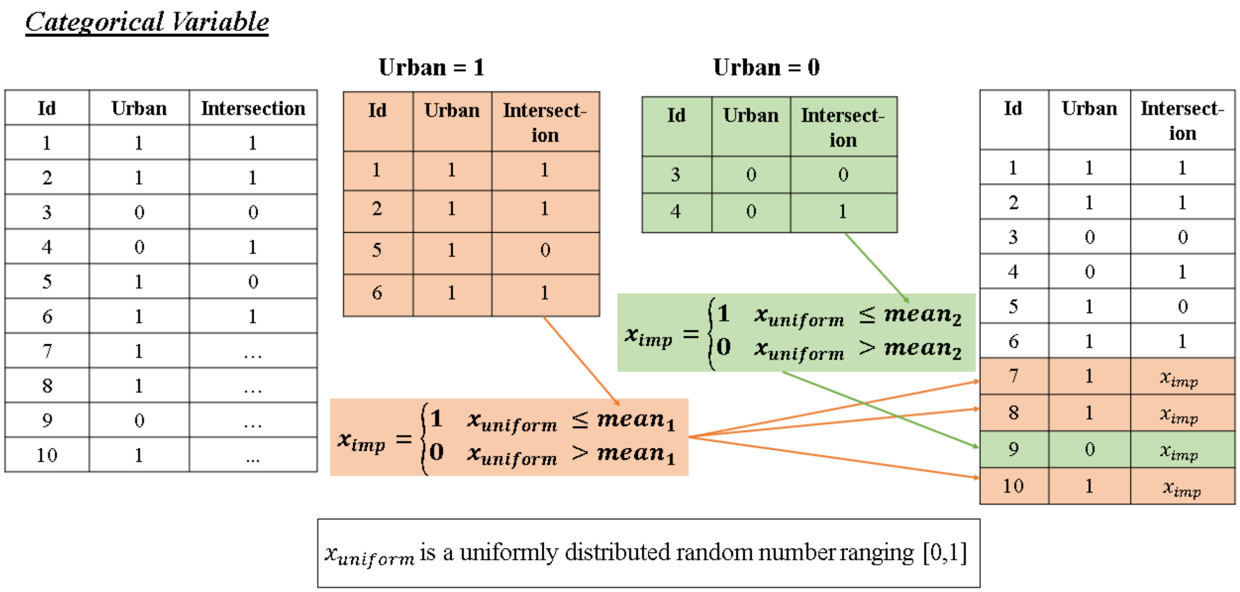
$x_{imp} = mean_2 + SD_2 * x_{random}$

$x_{imp} = mean_1 + SD_1 * x_{random}$

Id	Urban	Speed limit
1	1	75
2	1	45
3	0	65
4	0	30
5	1	75
6	1	65
7	1	x_{imp}
8	1	x_{imp}
9	0	x_{imp}
10	1	x_{imp}

x_{random} is random number of
a) standard normal distribution when existing cases of speed limit is normally distributed or,
b) uniform distribution ranging [0,1] when existing cases of speed limit is uniformly distributed

Figure 10 Data imputation procedure for continuous variable



Categorical Variable

Id	Urban	Intersection
1	1	1
2	1	1
3	0	0
4	0	1
5	1	0
6	1	1
7	1	…
8	1	…
9	0	…
10	1	…

Urban = 1

Id	Urban	Intersect-ion
1	1	1
2	1	1
5	1	0
6	1	1

$$x_{imp} = \begin{cases} \mathbf{1} & x_{uniform} \leq mean_1 \\ \mathbf{0} & x_{uniform} > mean_1 \end{cases}$$

Urban = 0

Id	Urban	Intersect-ion
3	0	0
4	0	1

$$x_{imp} = \begin{cases} \mathbf{1} & x_{uniform} \leq mean_2 \\ \mathbf{0} & x_{uniform} > mean_2 \end{cases}$$

Id	Urban	Intersect-ion
1	1	1
2	1	1
3	0	0
4	0	1
5	1	0
6	1	1
7	1	x_{imp}
8	1	x_{imp}
9	0	x_{imp}
10	1	x_{imp}

$x_{uniform}$ is a uniformly distributed random number ranging [0,1]

Figure 11 Data imputation procedure for categorical variable

MODEL ESTIMATION RESULT

Model Fit Measure

The four variables of interest were: (1) crash prone segment, (2) crash prone intersection, (3) crash type, and (4) crash severity. Among the four variables of interest, we intended to compare the MI with 5 realizations to the relevant CCD model. However, due to the presence of a higher percentage of missing data, and the presence of small number of relevant independent variables no model considering CCD is developed for crash prone segment and intersection selection. So, we estimated the following models: (1) crash prone segment selection model using imputed data, (2) crash prone intersection selection model using imputed data, (3) crash type model using (a) CCD by column elimination and (b) imputed data, and (4) crash severity model using (a) CCD by column elimination, and (b) imputed data. The Bayesian Information Criterion (BIC) values of the crash prone segment selection, crash prone intersection selection, crash type, and crash severity model employing imputed (CCD) data are respectively as follows: 4,821(N/A), 584(N/A), 82,196(85,369), and 94,991(95,332). The comparison of BIC values indicates that at any discrete choice modeling framework addition of variables through multiple imputation outperforms CCD in terms of model fit.

In the interest of space, we do not describe the results of all models and focus on crash type and severity models (see Hoover et al. (*41*) for preliminary model discussion).

Model Results

Crash Type Model

The alternatives for the crash type dependent variable include: rearend-, head on-, angle-, sideswipe-, fixed object-, non-fixed object- and non-motorized crash. Prior to the estimation of crash type MNL model, three independent variables - Intersection location, urban area, rainy weather and cold weather variables – were imputed 5 times to obtain 5 dataset realizations. The model parameters from the 5 estimations are consolidated to generate inferences and results are presented in Table 6. Interested readers can find the crash type model results using the complete case data approach in the appendix (Table A1).

Roadway and Traffic Characteristics

Among the highway and traffic characteristics 6 variables were found to have significant impact in the model. The results regarding interstate highways reveal that interstate highways are more likely to experience sideswipe crashes and less likely to experience head-on, angle and non-motorized crashes. Crash type was also found to be affected by different days of week and hours of day. It is noticeable that, on the weekdays the likelihood of rear-end and sideswipe crashes occurring is higher than other crash types. Further, in both morning- and evening- peak period the probability of the occurrence of rear-end crash is the highest. The reader would note that, the impact of time on the crash type indicates that increased traffic volume on the roadways increases the likelihood of occurring rear-end crashes. Intersection is found to offer a positive impact on head-on and angle crashes. The result regarding head-on crashes seems to be counter-intuitive but a possible reason could be the stoppings of the left-turning vehicles at the left-most lane(s), that might collide with the through traffic from the opposite direction (see (*39*) for similar result). Again, rural areas are found to experience less rear-end and sideswipe crashes. However, there are no significant differences among the probabilities of head-on, angle, fixed object, non-fixed object and non-motorized crashes in rural areas.

Environmental Characteristics

Among the variables regarding environmental characteristics only weather condition was found to offer a significant impact. It is noticeable that, on rainy days the likelihood of experiencing crashes with fixed object is higher than any other crashes, whereas the probability of occurring crashes with non-fixed object is the lowest. On the other hand, cold weather conditions increase the likelihood of experiencing rear-end crashes than other crash types. It is also noticeable that there is no difference among the impacts of cold weather on head-on, angle, sideswipe, fixed object, non-fixed object, and non-motorized crashes. The reader would note that MI approach allowed us to represent the impact of the intersection and weather variables.

TABLE 6 Estimation of crash type model (Imputed data model)

Parameter	**Rearend**	**Head on**	**Angle**	**Sideswipe**	**Fixed object**	**Non-fixed object**	**Non-motorized**
	Estimate (t- value)	**Estimate (t- value)**	**Estimate (t- value)**	**Estimate (t- value)**	**Estimate (t- value)**	**Estimate (t- value)**	**Estimate (t- value)**
Intercept	--	-2.10 (-23.91)	-0.80 (-16.66)	-0.78 (-20.27)	0.01 (0.34)	-0.63 (-10.98)	-0.54 (-10.55)
Roadway and Traffic Characteristics							
Interstate highway (base: No)							
Yes	--	-1.98 (-7.00)	-1.79 (-14.51)	0.14 (2.31)	--	-0.40 (-4.85)	-1.79 (-14.63)
Days of week (base: weekend)							
Weekday	--	-0.35 (-4.03)	-0.16 (-3.78)	--	-0.63 (-13.81)	-0.57 (-9.53)	-0.24 (-4.68)
Hours of day (base: Off-peak)							
Morning peak	--	-0.21 (-1.92)	-0.20 (-3.81)	-0.20 (-3.23)	-0.47 (-7.59)	-0.54 (-6.32)	-0.43 (-6.53)
Evening peak	--	-0.26 (-2.85)	-0.18 (-4.06)	-0.31 (-5.76)	-0.82 (-14.59)	-0.59 (-8.29)	-0.38 (-7.10)
Location (base: Urban)							
Rural	-0.60 (-16.59)	--	--	-0.54 (-10.03)	--	--	--
Intersection (base: No)							

Yes	--	0.37 (4.71)	1.20 (29.25)	-0.52 (-9.40)	-1.24 (-21.70)	-1.41 (-17.88)	-0.22 (-4.75)
Environmental Characteristics							
Weather condition (base: Clear weather and others)							
Rainy weather	--	--	--	-0.22 (-2.79)	0.48 (7.72)	-0.30 (-2.77)	-0.19 (-2.38)
Cold weather	0.13 (3.32)	--	--	--	--	--	--

Crash Severity Model

The outcome levels considered in the ordered logit crash severity model include: no apparent injury, possible injury, suspected minor injury and suspected serious injury including fatality. Several variables were identified to have large shares of missing values for the crash severity dataset. Prior to estimation, the missing records were imputed and five realizations were generated. The model results were consolidated to obtain the crash severity estimation results are presented in Table 7. Interested readers can view the crash severity model results for complete case data approach in the Appendix (Table A2).

Roadway and Traffic Characteristics

Several highway and traffic characteristics related attributes were found to be significant in our model. Among them, weekdays offered a negative association with crash severity. Again, injury risk is also lower during morning and evening peak period. The result implies that, traffic congestion and lower traffic speed during weekdays and peak periods tend to lower the injury severity (see (*42*) for similar result). The model result for intersection crashes reveals a higher injury risk propensity to the drivers. Further, the positive association of speed limit reflects that injury severity is higher for crashes occurring on high-speed facilities (see (*43*) for similar result).

Crash Characteristics

The model output regarding the manner of collisions reveals that there is no significant difference among the injury severity of rear-end crashes, non-motorized crashes, and crashes with non-fixed objects. The results indicate that head-on crashes, angle crashes, and crashes with fixed objects are found to be more severe than rear-end crashes, non-motorized crashes, and crashes with non-fixed objects (see (*43*) for similar result). Sideswipe collisions are associated with the least severity.

Vehicle Characteristics

Only one variable – vehicle type – was found to offer significant impact in the model. Drivers in utility vehicles and all types of trucks (light, medium, and heavy) are observed to experience less severe crashes than the drivers in automobiles, buses, and other vehicles. The model result follows the trend described in earlier literature (*43*).

Environmental Characteristics

Several environmental characteristics were tried in this study, however, only weather conditions were noticed to have significant impact on crash severity. It was found that crashes in clear weather conditions are more severe than crashes in rainy, cloudy, and other weather conditions. The output reflects that higher speed of vehicles in clear weather and/or cautious driving in inclement weather are possible explanations for this result.

Demographic Characteristics

Among several demographic characteristics, drivers' age and gender provided significant impact in the model. The result for age indicates that young drivers (age < 25 years old) are less likely to sustain severe injuries compared to drivers from other age categories (see (*43*)). On the other hand, male drivers are likely to sustain severe crashes than female drivers (the finding is documented in earlier literature (*42*)).

TABLE 7 Estimation of crash severity model (Imputed data model)

Parameters	**Coefficient**	**T-value**
Threshold		
a	0.46	9.53
b	1.62	32.92
c	2.62	52.45
Roadway and Traffic Characteristics		
Days of week (base: weekend)		
Weekday	-0.17	-7.51
Hours of day (base: Off-peak)		
Morning peak	-0.12	-4.42
Evening peak	-0.11	-4.70
Intersection (base: No)		
Yes	0.04	1.72
Speed limit	0.01	14.00
Crash Characteristics		
Manner of collisions (base: Rear end, crash with non-fixed object, non-motorized crash)		
Head on	1.32	24.94
Sideswipe	-0.85	-24.04
Crash with fixed object	0.48	11.26
Angle	0.42	17.99
Vehicle Characteristics		
Vehicle type (base: automobile, bus, and others)		
Utility vehicles	-0.10	-4.05
Light truck	-0.14	-5.48
Medium and heavy truck	-0.58	-9.61
Environmental Characteristics		
Weather condition (base: Rainy, cloudy and others)		
Clear weather	0.18	7.56
Demographic Characteristics		
Age (base: Others)		
Less than 25 years old	-0.14	-5.79
Gender (base: Female)		
Male	0.11	5.55

CONCLUSIONS

The development of transportation econometric model relies on available data from public and private agencies. In these datasets, several reasons influence data unavailability for independent variables. The elimination of records with missing data can possibly result in larger standard errors for parameter estimates of the variables. Further the analyst may introduce a misspecification in the model by eliminating the variable with missing data from the analysis. Therefore, the researchers can address the missing data problem by imputing data for these missing variables and then develop econometric models. While approaches for imputation are documented in econometric literature, their application in transportation research is limited.

The current study is motivated toward clarifying the value of Multiple Imputation approach for missing data in model estimation and employing it for empirical application. The paper employs a data simulation experiment comparing the performance of–single imputation, MI with different realizations and complete case data (CCD) approach (removing missing value records). The data simulation results are compared using (a) model fit measures (log likelihood improvement) and (b) the true parameter retrieval ability. From our analysis, we conclude that MI approach with 5 realizations outperforms the SI approach. Further, in comparing the MI approach with CCD approach, we notice that when the sample sizes are 1000 and above, CCD approach performs slightly better than the multiple imputation approach for continuous variables at all three missing percentages. For categorical variables, multiple imputation offers better results relative to CCD approach for all three missing percentages. However, the differences become smaller for datasets with more than 1000 records. Hence, for large datasets (>1000 records), in the presence of a small share of missing data, it might be beneficial to simply develop a CCD model by dropping observations with missing values as opposed to developing imputation models. However, when the share of missing data warrants variable exclusion, it is important and even necessary that multiple imputation approach be employed for model development.

Drawing on the conclusions of the experimental design, the research study employed MI for empirical datasets. The case study consisted of four datasets including (a) estimation of crash prone segment (MNL model), (b) estimation of crash prone intersection (MNL model), (c) estimation of different crash types (MNL model) and (d) estimation of crash severity (ordered logit model). The comparison of BIC values indicates that for any discrete choice modeling framework addition of variables through multiple imputation outperforms CCD in terms of model fit.

To be sure, the research conducted is not without limitations. First, in our study our emphasis has been on evaluating the impact of imputation on model development and not on the actual data imputation procedures. It would be interesting to conduct the analysis with advanced imputation approaches to see if the conclusions from our work are reproduced. The consideration of different imputation approaches and alternative formulations (such as latent variables approach) will require substantial investigation and need to be examined in future research efforts to build on the current study (see (9, 13) for details on these approaches). Second, in our analysis we did not consider the potential for missing records in the dependent variable and/or independent variables. The simultaneous presence of missing records in dependent and independent variables can increase the complexity of the analysis conducted and is an avenue for future research. Third, the current research examined the presence of one missing variable – the analysis can be extended to multiple missing variables. Fourth, while it was encouraging to see that imputation approaches provide similar quality of results in fixed parameters and random parameters MNL models, it is important to note that we did not test the impact of mixing on the variable with missing data. We believe this is a complex issue beyond the scope of our paper. Finally, the analysis can also be extended to other econometric models such as generalized linear models, Bayesian models, and structural equation models.

ACKNOWLEDGEMENTS

This study was supported by the Federal Highway Administration through the grant FHWA-PROJ-19-0014. However, the authors are solely responsible for the facts and accuracy of the information presented in the paper. The authors extend their sincere appreciation to the anonymous reviewers for their invaluable insights and constructive suggestions.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Naveen Eluru; data collection: Md. Istiak Jahan, Lauren Hoover; analysis and interpretation of results: Md. Istiak Jahan, Tanmoy Bhowmik, Naveen Eluru; draft manuscript preparation: Md. Istiak Jahan, Naveen Eluru. All authors reviewed the results and approved the final version of the manuscript.

Appendix

TABLE A1 Estimation of crash type model (Complete case data model)

Parameters	Rear end	Head on	Angular	Sideswipe	Fixed object	Non-fixed objects	Non-motorized
	Estimate (t- value)	Estimate (t- value)	Estimate (t- value)	Estimate (t- value)	Estimate (t- value)	Estimate (t- value)	Estimate (t- value)
Intercept	- -	-1.50 (-13.59)	0.71 (13.73)	-0.79 (-12.77)	1.25 (21.51)	0.96 (14.10)	-0.39 (-5.40)
Roadway and Traffic Characteristics							
Number of lanes	- -	-0.11 (-4.23)	-0.24 (-18.98)	-0.03 (-1.83)	-0.44 (-27.67)	-0.55 (-28.08)	-0.28 (-13.66)
Interstate highway	- -	-2.14 (-8.46)	-2.39 (-21.49)	0.32 (5.89)	- -	-0.24 (-3.34)	-2.16 (-11.65)
Weekdays	0.15 (4.00)	-0.15 (-1.78)	- -	- -	-0.55 (-13.01)	-0.40 (-7.75)	- -
Morning peak	- -	- -	- -	- -	-0.23 (-4.23)	-0.50 (-6.91)	-0.21 (-2.61)
Evening peak	- -	-0.25 (-2.90)	-0.20 (-4.85)	-0.29 (-5.67)	-0.83 (-15.82)	-0.71 (-11.65)	-0.38 (-5.67)

TABLE A2 Estimation of crash severity model (Complete case data model)

Parameters	Estimates	T-Value
Thresholds		
a	0.95	23.88
b	1.90	46.98
c	2.92	67.85
Demographic characteristics		
Age (Base: 25 years and more)		
Less than 25 years	-0.30	-11.15
Roadway and Traffic Characteristics		
Base: Other roadways		
Interstate highways	0.24	6.30
Number of lanes	0.03	4.34
Days of week (base: weekend)		
Weekday	-0.08	-2.90
Hours of day (base: off-peak)		
Morning peak	-0.05	-1.56
Evening peak	-0.15	-5.65
Crash characteristics		
Manner of collisions (base: Rear end)		
Head on	1.67	31.26

Side swipe	-0.39	-9.43
Angular crash	0.59	22.10
Crash with fixed objects	1.55	39.10
Crash with non-fixed objects	0.85	16.99
Non-motorized crash	-2.58	-13.87
Vehicle characteristics		
Vehicle type (Base: Automobiles, motorcycle and bus		
Utility vehicles	-0.31	-11.23
Light truck	-0.52	-16.71
Medium and heavy truck	-1.73	-20.38

REFERENCES

1. Bhat, C. R. Imputing a Continuous Income Variable from Grouped and Missing Income Observations. *Economics Letters*, Vol. 46, No. 4, 1994, pp. 311–319. https://doi.org/10.1016/0165-1765(94)90151-1.
2. Rubin, D. B., and N. Schenker. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, Vol. 81, No. 394, 1986, pp. 366–374.
3. Rubin, D. B. Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, Vol. 91, No. 434, 1996, pp. 473–489. https://doi.org/10.1080/01621459.1996.10476908.
4. Rubin, D. B. Inference and Missing Data. *Biometrika*, Vol. 63, No. 3, 1976, pp. 581–592. https://doi.org/doi.org/10.1093/biomet/63.3.581.
5. Rubin, D. B. The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys. *American Statistician*, Vol. 58, No. 4, 2004, pp. 298–302. https://doi.org/10.1198/000313004X6355.
6. Rubin, D. B. An Overview of Multiple Imputation. *Proceedings of the survey research methods section of the American statistical association*, 1988, pp. 79–84.
7. Campion, W. M., and D. B. Rubin. Multiple Imputation for Nonresponse in Surveys. *Journal of Marketing Research*, Vol. 26, No. 4, 1989, p. 485. https://doi.org/10.2307/3172772.
8. Rubin, D. B. Multiple Imputations in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. *American Statistical Association*, Vol. 1, 1978, pp. 20–34.
9. Afghari, A. P., S. Washington, C. Prato, and M. M. Haque. Contrasting Case-Wise Deletion with Multiple Imputation and Latent Variable Approaches to Dealing with Missing Observations in Count Regression Models. *Analytic Methods in Accident Research*, Vol. 24, 2019, p. 100104. https://doi.org/10.1016/j.amar.2019.100104.
10. Al-Deek, H. M., and C. V. S. R. Chandra. New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse. *Transportation Research Record*, No. 1867, 2004, pp. 116–126. https://doi.org/10.3141/1867-14.
11. Budhwani, A., T. Lin, D. Feng, and C. Bachmann. Assessing and Comparing Data Imputation Techniques for Item Nonresponse in Household Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, 2022, p. 036119812211048. https://doi.org/10.1177/03611981221104802.
12. Gopalakrishnan, R., C. A. Guevara, and M. Ben-Akiva. Combining Multiple Imputation

and Control Function Methods to Deal with Missing Data and Endogeneity in Discrete-Choice Models. *Transportation Research Part B: Methodological*, Vol. 142, 2020, pp. 45–57. https://doi.org/10.1016/j.trb.2020.10.002.

13. Murray, J. S. Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, Vol. 33, No. 2, 2018, pp. 142–159. https://doi.org/10.1214/18-STS644.
14. Enders, C. K. *Applied Missing Data Analysis*. Guilford Publications, Inc., New York, 2010.
15. Newman, J., M. E. Ferguson, and L. A. Garrow. Estimating Discrete Choice Models with Incomplete Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2302, No. 14, 2012, pp. 130–137. https://doi.org/10.3141/2302-14.
16. Zhao, Y., J. Pawlak, and J. W. Polak. Inverse Discrete Choice Modelling: Theoretical and Practical Considerations for Imputing Respondent Attributes from the Patterns of Observed Choices. *Transportation Planning and Technology*, Vol. 41, No. 1, 2018, pp. 58–79. https://doi.org/10.1080/03081060.2018.1402745.
17. Li, K. H., T. E. Raghunathan, and D. B. Rubin. Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. *Journal of the American Statistical Association*, Vol. 86, No. 416, 1991, pp. 1065–1073. https://doi.org/doi.org/10.1080/01621459.1991.10475152.
18. Tang, J., D. S. Hillygus, and J. P. Reiter. Using Auxiliary Marginal Distributions in Imputations for Nonresponse While Accounting for Survey Weights, with Application to Estimating Voter Turnout. *arXiv*, 2022, pp. 1–21.
19. Miller, M., A. Michaels-obregón, K. O. Rocha, and R. Wong. Imputation of Non-Response in Height and Weight in the Mexican Health and Aging Study. *Real Datos Espacio*, Vol. 13, No. 2, 2023, pp. 78–93.
20. Si, Y., S. Heeringa, D. Johnson, R. J. A. Little, W. Liu, F. Pfeffer, and T. Raghunathan. Multiple Imputation with Massive Data: An Application to the Panel Study of Income Dynamics. *Journal of Survey Statistics and Methodology*, Vol. 11, No. 1, 2023, pp. 260–283. https://doi.org/10.1093/jssam/smab038.
21. Li, L., J. Zhang, Y. Wang, and B. Ran. Multiple Imputation for Incomplete Traffic Accident Data Using Chained Equations. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, Vol. 2018-March, 2018, pp. 1–5. https://doi.org/10.1109/ITSC.2017.8317639.
22. Sanko, N., S. Hess, J. Dumont, and A. Daly. Contrasting Imputation with a Latent Variable Approach to Dealing with Missing Income in Choice Models. *Journal of Choice Modelling*, Vol. 12, 2014, pp. 47–57. https://doi.org/10.1016/j.jocm.2014.10.001.
23. Tang, J., Y. Wang, S. Zhang, H. Wang, F. Liu, and S. Yu. On Missing Traffic Data Imputation Based on Fuzzy C-Means Method by Considering Spatial-Temporal Correlation. *Transportation Research Record*, Vol. 2528, 2015, pp. 86–95. https://doi.org/10.3141/2528-10.
24. Choi, Y. Y., H. Shon, Y. J. Byon, D. K. Kim, and S. Kang. Enhanced Application of Principal Component Analysis in Machine Learning for Imputation of Missing Traffic Data. *Applied Sciences (Switzerland)*, Vol. 9, No. 10, 2019, pp. 1–15. https://doi.org/10.3390/app9102149.
25. Ku, W. C., G. R. Jagadeesh, A. Prakash, and T. Srikanthan. A Clustering-Based Approach for Data-Driven Imputation of Missing Traffic Data. *Proceedings - 2016 IEEE Forum on Integrated and Sustainable Transportation Systems, FISTS 2016*, 2016, pp. 16–21. https://doi.org/10.1109/FISTS.2016.7552320.

26. Tak, S., S. Woo, and H. Yeo. Data-Driven Imputation Method for Traffic Data in Sectional Units of Road Links. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 6, 2016, pp. 1762–1771. https://doi.org/10.1109/TITS.2016.2530312.
27. Tan, H., Z. Yang, G. Feng, W. Wang, and B. Ran. Correlation Analysis for Tensor-Based Traffic Data Imputation Method. *Procedia - Social and Behavioral Sciences*, Vol. 96, No. Cictp, 2013, pp. 2611–2620. https://doi.org/10.1016/j.sbspro.2013.08.292.
28. Xu, D., H. Peng, C. Wei, X. Shang, and H. Li. Traffic State Data Imputation: An Efficient Generating Method Based on the Graph Aggregator. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 8, 2022, pp. 13084–13093. https://doi.org/10.1109/TITS.2021.3119638.
29. Duan, Y., Y. Lv, W. Kang, and Y. Zhao. A Deep Learning Based Approach for Traffic Data Imputation. *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 2014, pp. 912–917. https://doi.org/10.1109/ITSC.2014.6957805.
30. Ouimet, M. C., B. G. Simons-Morton, P. L. Zador, N. D. Lerner, M. Freedman, G. D. Duncan, and J. Wang. Using the U.S. National Household Travel Survey to Estimate the Impact of Passenger Characteristics on Young Drivers’ Relative Risk of Fatal Crash Involvement. *Accident Analysis and Prevention*, Vol. 42, No. 2, 2010, pp. 689–694. https://doi.org/10.1016/j.aap.2009.10.017.
31. Liu, J., A. J. Khattak, S. H. Richards, and S. Nambisan. What Are the Differences in Driver Injury Outcomes at Highway-Rail Grade Crossings? Untangling the Role of Pre-Crash Behaviors. *Accident Analysis and Prevention*, Vol. 85, 2015, pp. 157–169. https://doi.org/10.1016/j.aap.2015.09.004.
32. Deb, R., and A. W. C. Liew. Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data. *Information Sciences*, Vol. 339, 2016, pp. 274–289. https://doi.org/10.1016/j.ins.2016.01.018.
33. Farhan, J., and T. F. Fwa. Improved Imputation of Missing Pavement Performance Data Using Auxiliary Variables. *Journal of Transportation Engineering*, Vol. 141, No. 1, 2015, pp. 1–8. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000725.
34. Ye, F., and Y. Wang. Performance Evaluation of Various Missing Data Treatments in Crash Severity Modeling. *Transportation Research Record*, Vol. 2672, No. 38, 2018, pp. 149–159. https://doi.org/10.1177/0361198118798485.
35. Abe, T., S. Seino, T. Hata, M. Yamashita, N. Ohmori, A. Kitamura, S. Shinkai, and Y. Fujiwara. Transportation Modes and Social Participation in Older Drivers and Non-Drivers: Results from Urbanised Japanese Cities. *Journal of Transport Geography*, Vol. 109, No. April 2022, 2023, p. 103598. https://doi.org/10.1016/j.jtrangeo.2023.103598.
36. Blanchette, S., R. Larouche, M. S. Tremblay, G. Faulkner, N. A. Riazi, and F. Trudeau. Associations Between School Environments, Policies and Practices and Children’s Physical Activity and Active Transportation. *Journal of School Health*, Vol. 92, No. 1, 2022, pp. 31–41. https://doi.org/10.1111/josh.13102.
37. Qi, H., X. Zhao, Y. Yao, H. Yang, S. Chai, and X. Chen. BGCP-Based Traffic Data Imputation and Accident Detection Applications for the National Trunk Highway. *Accident Analysis and Prevention*, Vol. 186, 2023. https://doi.org/10.1016/j.aap.2023.107051.
38. Mohiuddin, H., D. T. Fitch-Polse, and S. L. Handy. Does Bike-Share Enhance Transport Equity? Evidence from the Sacramento, California Region. *Journal of Transport Geography,* Vol. 109, No. July 2022, 2023, p. 103588. https://doi.org/10.1016/j.jtrangeo.2023.103588.

39. Bhowmik, T., S. Yasmin, and N. Eluru. A New Econometric Approach for Modeling Several Count Variables: A Case Study of Crash Frequency Analysis by Crash Type and Severity. *Transportation Research Part B: Methodological*, Vol. 153, 2021, pp. 172–203. https://doi.org/10.1016/J.TRB.2021.09.008.

40. Enders, C. K. *Applied Missing Data Analysis*. Guilford press, 2010.

41. Hoover, L., M. I. Jahan, T. Bhowmik, S. D. Tirtha, K. C. Konduri, J. Ivan, K. Wang, S. Zhao, J. Auld, and N. Eluru. Implementation of a Realistic Artificial Data Generator for Crash Data Generation. *Accident Analysis & Prevention*, Vol. 200, 2024, pp. 1–13.

42. Yasmin, S., N. Eluru, C. R. Bhat, and R. Tay. A Latent Segmentation Based Generalized Ordered Logit Model to Examine Factors Influencing Driver Injury Severity. *Analytic Methods in Accident Research*, Vol. 1, 2014, pp. 23–38. https://doi.org/10.1016/j.amar.2013.10.002.

43. Yasmin, S., and N. Eluru. Evaluating Alternate Discrete Outcome Frameworks for Modeling Crash Injury Severity. *Accident Analysis and Prevention*, Vol. 59, 2013, pp. 506–521. https://doi.org/10.1016/J.AAP.2013.06.040.

44. Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, Vol. 49, No. 12, 1996, pp. 1373–1379. https://doi.org/10.1016/j.amepre.2003.12.002.

45. de Jong, V. M. T., M. J. C. Eijkemans, B. van Calster, D. Timmerman, K. G. M. Moons, E. W. Steyerberg, and M. van Smeden. Sample Size Considerations and Predictive Performance of Multinomial Logistic Prediction Models. *Statistics in Medicine*, Vol. 38, No. 9, 2019, pp. 1601–1619. https://doi.org/10.1002/sim.8063.

46. Riley, R. D., K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell, K. G. M. Moons, and G. S. Collins. Minimum Sample Size for Developing a Multivariable Prediction Model: PART II - Binary and Time-to-Event Outcomes. *Statistics in Medicine*, Vol. 38, No. 7, 2019, pp. 1276–1296. https://doi.org/10.1002/sim.7992.

47. Riley, R. D., K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell, K. G. M. Moons, and G. S. Collins. Minimum Sample Size for Developing a Multivariable Prediction Model: Part I – Continuous Outcomes. *Statistics in Medicine*, Vol. 38, No. 7, 2019, pp. 1262–1275. https://doi.org/10.1002/sim.7993.