Thailand Statistician
January 2022; 20(1): 1-15
http://statassoc.or.th
Contributed paper

Imputation Methods for Multiple Regression with Missing Heteroscedastic Data

Muhammad Asif [a][b][c] and Klairung Samart*[a]
[a] Division of Computational Science, Faculty of Science,
Prince of Songkla University, Songkhla, Thailand
[b] Statistics and Applications Research Unit, Faculty of Science,
Prince of Songkla University, Songkhla, Thailand
[c] Department of Economics, Lasbela University of Agriculture,
Water and Marine Sciences, Uthal, Balochistan, Pakistan
*Corresponding author; e-mail: klairung.s@psu.ac.th

Received: 1 May 2020
Revised: 11 October 2020
Accepted: 29 December 2020

Abstract

The purpose of this research is to compare the efficiency of different imputation methods for multiple regression analysis of heteroscedastic data with missing at random dependent variable. The missing data imputation methods used in this study are mean imputation, hot deck imputation, k-nearest neighbors imputation (KNN), stochastic regression imputation, along with three proposed composite methods, namely hot deck and KNN imputation with equivalent weight (HKEW), hot deck and stochastic regression imputation with equivalent weight (HSEW), and mean and stochastic regression imputation with equivalent weight (MSEW). The comparison between the seven methods was conducted through the simulation study varied by the sample sizes and the missing percentages. The criteria for comparing the efficiency of estimators are bias and mean squared error (MSE). The results show that the stochastic regression imputation performed well in terms of bias in all situations. In terms of MSE, the mean imputation performed well when the sample size is small to medium, whereas the MSEW imputation performed well when the sample size is large and the missing percentage is high (30-40%).

Keywords: Missing data, imputation, equivalent weight, bias, mean squared error.

1. Introduction

Multiple regression analysis is a statistical method used to determine the relationship between a dependent variable and independent variables. Missing data on a dependent variable is common in research studies. It may occur by refusals, miscommunication, withdrawing, lack of information, privacy, loss of questionnaires, irrelevant questions or other reasons. Missing data pattern is an important factor to be identified for the choice of methods to handle missing data. Little and Rubin (1987) introduced three missing data mechanisms namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Missing data in which missingness does not depend on observed and unobserved values is known as MCAR. This type of missingness enlarges standard errors due to the reduced sample size, but does not cause bias (Jakobsen et al.,

2017). More often the missingness depends on observed values but not on unobserved values. This type of missingness is said to be MAR. If the mechanism depends on the missing data itself, then it is classified as MNAR (Sterne et al., 2009; Jakobsen et al., 2017).

There are several methods for handling missing data (Little and Rubin, 1987; Donders et al., 2006; Buhi et al., 2008; Enders, 2010; Lamjaisue et al., 2017). The simple and popular methods such as the deletion method, the overall mean imputation, and missing-indicator method yield biased estimates (Donders et al., 2006; Groenwold et al., 2012). For the deletion method, all observations with missing values in at least one variable are eliminated (Munguía and Armando, 2014). The deletion technique is very easy to apply and does not require much knowledge about statistics. However, it decreases the sample size and therefore weakens the statistical power as well as gives biased parameter estimates, especially when the missingness mechanism is not MCAR (Buhi et al., 2008). Alternatively, imputation of missing data has been a good choice in recent years (Jerez et al., 2010). The idea of this method is to substitute each missing value with a suitable value and then continue the analysis as there were no missing values (Dettori et al., 2018). The purpose of imputation is not only to replace all missing values; but to preserve the characteristics of their distribution and relationships among different variables (Munguía and Armando, 2014).

Imputation methods such as mean imputation, hot deck imputation, regression imputation, stochastic imputation, multiple imputation, and k-nearest neighbors imputation are mainly discussed in literature. It was found that machine learning methods such as k-nearest neighbors imputation outperformed the classical methods in some studies (Jerez et al., 2010; Lamjaisue et al., 2017). The polytomous regression and hot deck imputations were also found to be more effective than other simple methods (Elliott and Hawthorne, 2005; Munguía and Armando, 2014).

In this work, we introduce three composite imputation methods: hot deck and k-nearest neighbor with equivalent weight (HKEW), hot deck and stochastic regression with equivalent weight (HSEW), and mean and stochastic regression imputation with equivalent weight (MSEW). We compare these three methods with four classical and popular imputation methods: mean imputation, hot deck imputation, k-nearest neighbors imputation, and stochastic regression imputation. The multiple regression analysis is performed under heteroscedasticity which usually occurs in reality.

This paper is organized as follows. Section 2 presents the four imputation methods used in this work and introduces the three composite imputation methods. Then, Section 3 is devoted to the performance of the imputation methods via simulation study in different scenarios. Section 4 reveals the results of application to real life data. Conclusions, including discussion of the results are set out in Section 5.

2. Methodology

Let Y_i be a random variable where $i = 1, 2, ..., n$ with m missing values and $X_{i1}, X_{i2}, ..., X_{iq}$ be independent variables. The imputation methods used in this work are as follows.

2.1. Mean imputation

In mean imputation method, missing values of a variable are replaced by the mean of other observed values in the variable (Saunders et al., 2006). Therefore, this method is limited to the numerical data. Although by using this method, the sample size is maintained and the use is uncomplicated, the variance will be downwardly biased irrespective of underlying missing data mechanism (Buhi et al., 2008; Enders, 2010; Dettori et al., 2018). The mean imputed value is given by

$$\bar{y}^* = \sum_{i=1}^{n-m} \left(\frac{y_i}{n-m} \right) \tag{1}$$

2.2. Hot deck imputation

Hot deck imputation method replaces missing values of one or more variables by observed values that are similar with respect to observed characteristics (Andridge and Little, 2010; Beretta and Santaniello, 2016). This method is poplular since it does not rely on model fitting for the variable to be imputed, and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model (Andridge and Little, 2010). Another attractive property of the hot deck is that the imputed values are plausible since values come from observed responses (Andridge and Little, 2010). As a result, this technique is commonly used by other government statistics agencies and survey organizations including the U.S. and the British census, the current population survey, the Canadian census of Construction, the U.S. Annual survey of Manufacturers, and the U.S. National Medical Care utilization and Expenditure survey (Myers, 2011).

2.3. Stochastic regression imputation

In regression imputation, the imputed value is estimated from a regression equation obtained from the observed values (Lodder, 2014). To impute the missing values for the Y variable, a regression equation is constructed using the observed data y^* on x^*. Then we will use this equation to predict the missing values on y. The regression equation is given by

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + ... + \hat{\beta}_q x_q^*$$

Obviously, it assumes that the imputed values fall on a regression line implying that a correlation between the predictors and the missing outcome variable is 1 which is impossible in reality. One way to overcome this problem is to add a residual term to the imputed value. This is called a stochastic regression imputation.

The residual that is added to the predicted value is drawn from a normal distribution with a mean of zero and a variance equal to the residual variance from the regression of the predictor on the outcome. This is to preserve the variability in the data. Also, parameter estimates are unbiased with MAR data (Enders, 2010).

2.4. K-nearest neighbors imputation

K-nearest neighbor (KNN) has been widely applied in literature (Jerez et al., 2010; Beretta and Santaniello, 2016; Lamjaisue et al., 2017). It classifies the data into groups and then replaces the missing values with the corresponding value from the nearest neighbors (Jerez et al., 2010). The number of neighbors k is chosen based on a distance measure which varies according to the type of data such as Euclidean distance for continuous data and Hamming distance for categorical data. Then, their average is used as an imputation estimate.

In this context, the nearest neighbors are the closest values using the Euclidean distance. The formula for Euclidean distance is given by

$$D_{ij} = \sqrt{\sum_{p=1}^{q} (x_{ip} - x_{jp})^2}$$

where x_{ip} denotes the value of an independent variable p of the missing case, $i = 1, 2, ..., m$. x_{jp} denotes the value of an independent variable p of the observed case, $j = m + 1, ..., n$. q denotes the numbers of independent variables. The number of nearest neighbors k is the square root of the total number of observed observations in the data which is always an odd number (Lamjaisue et al., 2017).

2.5. Composite method with equivalent weight

In general, a composite method with equivalent weight is a combination of two or more methods which can be defined as follows.

$$\tilde{y_i} = W_M(\hat{y}_{i1} + \hat{y}_{i2} + ... + \hat{y}_{iM}); \quad i = 1, 2, ..., m$$

where $\hat{y}_{ij}$ is the imputation i^{th} from method j^{th} and the equivalent weight $W_M = \dfrac{1}{M}$ where M is the number of combination methods. Here we develop three composite methods which is a combination of two frequently used methods in the literature as follows.

2.5.1 Hot deck and KNN with equivalent weight (HKEW)

HKEW is a combination between hot deck and KNN imputations. Let $\hat{y_i}$ denote the hot deck imputed value and y_i^* denote the KNN imputed value. Then, the HKEW imputed value is given by

$$\tilde{y_i} = \frac{1}{2}(\hat{y_i} + y_i^*); \quad i = 1, 2, ..., m$$

2.5.2 Hot deck and stochastic regression with equivalent weight (HSEW)

HSEW is a combination of hot deck and stochastic regression imputations which is another frequently used method in the literature. The HSEW imputed value is given by

$$\breve{y_i} = \frac{1}{2}(\hat{y_i} + \hat{y_i^*}); \quad i = 1, 2, ..., m.$$

2.5.3 Mean and stochastic regression with equivalent weight (MSEW)

The third method that we develop in this work is MSEW which is a combination of mean and stochastic regression imputations as they performed well in terms of bias and MSE as single imputations. The MSEW imputed value is given by

$$\breve{y_i} = \frac{1}{2}(\bar{y_i}^* + \hat{y_i^*}); \quad i = 1, 2, ..., m.$$

3. Simulation Studies

This section contains results from a simulation study illustrating the comparative performances of the imputation methods described in the previous section. The assessment of the six imputation methods was based on bias and mean squared error (MSE) of the regression coefficient estimators.

In our simulation study, a random sample of size n (=20, 30, 50, and 100) was generated and the values of independent variables were independently drawn from a uniform distribution $X_p \sim U(0, 1)$ where p represents the p^{th} independent variable used in the model. In our case we have two independent variables i.e X_1 and X_2. Therefore $p = 1, 2$. The corresponding values of Y are then given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

where $i = 1, 2, ..., m, m + 1, ..., n$ and the true value of the coefficients $\beta_0 = \beta_1 = \beta_2 = 1$. The random error term ϵ_i were set to be randomly generated from a normal distribution with zero mean and non-constant variance $x_{i1} + x_{i2}$.

For each sample size, the missingness was generated on the dependent variable Y_i using the MAR mechanism. The proportions of missingness were $10\%, 20\%, 30\%$ and 40%. Then, the seven imputation methods were applied to impute the missing values of the dependent variable Y_i and the regression coefficient estimates were obtained. The simulation process was replicated N = 1,000

times. The bias and MSE of the regression coefficient estimators of the seven imputation methods were then computed. All simulations were accomplished by using R software (R Core Team, 2018).

Tables 1-4 show the biases and MSEs of the regression coefficient estimators obtained by different imputation methods described in Section 2 and varied in different sample sizes when the proportions of missingness were $10\%, 20\%, 30\%$ and 40%, respectively. Figures 1 and 2 display the MSEs of $\hat{\beta}_1$ and $\hat{\beta}_2$ in different sample sizes, imputation methods, and missingness percentages, respectively.

The results set out in Tables 1 and 2 reveal that the estimators obtained by the stochastic regression method perform well in almost all situations in terms of bias. However, regarding the MSE which can also be seen in Figures 1 and 2, it is found that the mean imputation method outperforms other methods whereas the stochastic regression method gives higher MSE as the missingess percentage gets higher.

Table 1 Biases and MSEs obtained from different imputation methods with 10% missing values

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.100	0.297	-0.126	0.770	-0.139	0.751
	Hot deck	0.108	0.334	-0.137	0.857	-0.145	0.860
	Stochastic regression	0.018	0.370	-0.015	1.099	-0.027	0.986
	KNN	0.047	0.338	-0.057	0.875	-0.061	0.884
	HKEW	0.077	0.325	-0.097	0.830	-0.103	0.841
	HSEW	0.063	0.333	-0.076	0.911	-0.086	0.868
	MSEW	0.059	0.321	-0.071	0.897	-0.083	0.835
30	Mean	0.079	0.180	-0.110	0.498	-0.100	0.437
	Hot deck	0.079	0.203	-0.116	0.579	-0.098	0.495
	Stochastic regression	0.006	0.222	-0.007	0.680	-0.003	0.563
	KNN	0.019	0.211	-0.041	0.575	-0.014	0.524
	HKEW	0.049	0.199	-0.079	0.556	-0.056	0.488
	HSEW	0.042	0.201	-0.062	0.595	-0.051	0.498
	MSEW	0.042	0.193	-0.059	0.568	-0.052	0.482
50	Mean	0.086	0.114	-0.089	0.278	-0.134	0.306
	Hot deck	0.087	0.126	-0.090	0.301	-0.136	0.336
	Stochastic regression	0.006	0.143	0.024	0.364	-0.032	0.390
	KNN	0.026	0.131	-0.004	0.326	-0.056	0.362
	HKEW	0.056	0.123	-0.047	0.298	-0.094	0.336
	HSEW	0.046	0.126	-0.033	0.311	-0.084	0.344
	MSEW	0.046	0.123	-0.033	0.308	-0.083	0.336
100	Mean	0.070	0.053	-0.095	0.142	-0.089	0.144
	Hot deck	0.073	0.059	-0.094	0.160	-0.090	0.159
	Stochastic regression	-0.006	0.062	0.012	0.184	0.013	0.181
	KNN	0.003	0.063	-0.005	0.171	0.003	0.177
	HKEW	0.038	0.057	-0.050	0.157	-0.044	0.160
	HSEW	0.033	0.056	-0.041	0.160	-0.038	0.160
	MSEW	0.033	0.054	-0.041	0.155	-0.038	0.155

Table 2 Biases and MSEs obtained from different imputation methods with 20% missing values

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.138	0.263	-0.250	0.690	-0.169	0.678
	Hot deck	0.118	0.311	-0.227	0.881	-0.140	0.787
	Stochastic regression	-0.023	0.460	0.014	1.292	0.033	1.243
	KNN	0.045	0.324	-0.144	0.846	-0.035	0.908
	HKEW	0.081	0.295	-0.186	0.795	-0.088	0.783
	HSEW	0.047	0.341	-0.106	0.952	-0.053	0.896
	MSEW	0.057	0.330	-0.118	0.896	-0.068	0.885
30	Mean	0.160	0.186	-0.256	0.490	-0.195	0.418
	Hot deck	0.171	0.227	-0.249	0.574	-0.212	0.566
	Stochastic regression	-0.001	0.253	0.003	0.845	0.011	0.665
	KNN	0.054	0.230	-0.114	0.621	-0.046	0.584
	HKEW	0.112	0.210	-0.181	0.548	-0.129	0.524
	HSEW	0.083	0.210	-0.123	0.615	-0.101	0.540
	MSEW	0.078	0.199	-0.127	0.604	-0.092	0.497
50	Mean	0.158	0.114	-0.237	0.316	-0.199	0.260
	Hot deck	0.157	0.134	-0.251	0.403	-0.192	0.305
	Stochastic regression	0.000	0.139	-0.006	0.466	0.007	0.382
	KNN	0.040	0.136	-0.078	0.388	-0.038	0.357
	HKEW	0.099	0.122	-0.165	0.362	-0.115	0.298
	HSEW	0.078	0.118	-0.128	0.381	-0.092	0.295
	MSEW	0.079	0.113	-0.121	0.355	-0.096	0.291
100	Mean	0.162	0.068	-0.258	0.193	-0.196	0.153
	Hot deck	0.160	0.078	-0.258	0.223	-0.190	0.182
	Stochastic regression	0.004	0.070	-0.016	0.242	0.006	0.208
	KNN	0.033	0.072	-0.075	0.213	-0.026	0.197
	HKEW	0.096	0.067	-0.166	0.196	-0.108	0.170
	HSEW	0.082	0.062	-0.137	0.198	-0.092	0.168
	MSEW	0.083	0.059	-0.137	0.191	-0.095	0.160

Table 3 Biases and MSEs obtained from different imputation methods with 30% missing values

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.236	0.244	-0.405	0.709	-0.311	0.573
	Hot deck	0.244	0.319	-0.409	0.955	-0.317	0.744
	Stochastic regression	-0.027	0.464	0.070	1.677	-0.008	1.231
	KNN	0.133	0.299	-0.286	0.869	-0.120	0.804
	HKEW	0.178	0.277	-0.348	0.812	-0.218	0.689
	HSEW	0.108	0.310	-0.170	1.025	-0.162	0.803
	MSEW	0.104	0.296	-0.168	0.972	-0.159	0.781
30	Mean	0.245	0.180	-0.398	0.514	-0.312	0.439
	Hot deck	0.248	0.236	-0.383	0.688	-0.321	0.593
	Stochastic regression	-0.007	0.284	0.017	1.189	-0.005	0.813
	KNN	0.083	0.217	-0.203	0.660	-0.079	0.662
	HKEW	0.166	0.198	-0.293	0.596	-0.200	0.553
	HSEW	0.120	0.209	-0.183	0.745	-0.163	0.591
	MSEW	0.119	0.191	-0.191	0.699	-0.158	0.544
50	Mean	0.240	0.128	-0.392	0.367	-0.306	0.269
	Hot deck	0.239	0.150	-0.382	0.463	-0.308	0.350
	Stochastic regression	-0.009	0.154	0.022	0.641	-0.001	0.424
	KNN	0.071	0.141	-0.176	0.434	-0.067	0.377
	HKEW	0.155	0.129	-0.279	0.397	-0.187	0.313
	HSEW	0.115	0.122	-0.180	0.438	-0.155	0.314
	MSEW	0.115	0.113	-0.185	0.408	-0.154	0.292
100	Mean	0.243	0.094	-0.400	0.264	-0.308	0.188
	Hot deck	0.244	0.109	-0.399	0.311	-0.311	0.232
	Stochastic regression	-0.003	0.079	-0.001	0.290	-0.002	0.221
	KNN	0.061	0.086	-0.129	0.259	-0.067	0.235
	HKEW	0.152	0.080	-0.264	0.240	-0.189	0.195
	HSEW	0.120	0.069	-0.200	0.226	-0.156	0.177
	MSEW	0.120	0.065	-0.201	0.214	-0.155	0.166

Table 4 Biases and MSEs obtained from different imputation methods with 40% missing values

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.345	0.253	-0.602	0.728	-0.388	0.545
	Hot deck	0.355	0.331	-0.611	1.082	-0.369	0.722
	Stochastic regression	0.023	0.515	0.016	2.462	-0.027	1.420
	KNN	0.202	0.269	-0.514	0.887	-0.138	0.780
	HKEW	0.278	0.264	-0.562	0.867	-0.267	0.649
	HSEW	0.189	0.310	-0.298	1.237	-0.211	0.834
	MSEW	0.184	0.293	-0.293	1.147	-0.207	0.803
30	Mean	0.332	0.199	-0.588	0.584	-0.425	0.446
	Hot deck	0.326	0.252	-0.594	0.822	-0.413	0.580
	Stochastic regression	-0.011	0.315	0.066	1.589	-0.039	0.922
	KNN	0.160	0.210	-0.434	0.698	-0.151	0.659
	HKEW	0.243	0.199	-0.514	0.667	-0.282	0.530
	HSEW	0.158	0.205	-0.264	0.834	-0.226	0.596
	MSEW	0.161	0.187	-0.261	0.758	-0.232	0.559
50	Mean	0.348	0.174	-0.644	0.563	-0.389	0.298
	Hot deck	0.355	0.214	-0.657	0.689	-0.396	0.395
	Stochastic regression	0.012	0.185	-0.053	1.011	0.023	0.492
	KNN	0.153	0.154	-0.458	0.577	-0.078	0.402
	HKEW	0.254	0.154	-0.557	0.562	-0.237	0.322
	HSEW	0.184	0.139	-0.355	0.588	-0.187	0.329
	MSEW	0.180	0.131	-0.348	0.566	-0.183	0.308
100	Mean	0.340	0.139	-0.604	0.441	-0.416	0.245
	Hot deck	0.340	0.153	-0.602	0.504	-0.415	0.293
	Stochastic regression	0.008	0.082	0.016	0.454	-0.029	0.229
	KNN	0.103	0.092	-0.296	0.345	-0.078	0.245
	HKEW	0.222	0.095	-0.449	0.365	-0.246	0.208
	HSEW	0.174	0.076	-0.293	0.311	-0.222	0.192
	MSEW	0.174	0.074	-0.294	0.295	-0.222	0.180

The results displayed in Tables 3 and 4 show that the estimators obtained by the stochastic regression method still perform well in all situations in terms of bias. However, concerning MSE (see Figures 1 and 2), the mean imputation method outperforms other methods when the sample sizes are 20, 30, and 50. When the sample size is 100, the estimator obtained by MSEW gives the smallest MSE.

Overall, in terms of bias, when data are heteroscedastic, the proposed composite methods HKEW, HSEW and MSEW always perform better than the single methods: mean and hot deck method. Regarding the MSE, for small sample size, the HKEW method always gives smaller MSE than that of hot deck, stochastic regression, and KNN methods. For a large sample size, all composite methods always result in smaller MSE than that of the previously mentioned methods. Particularly, when both sample size and missing percentage are large, the MSEW yields smallest MSE compared to the other methods.

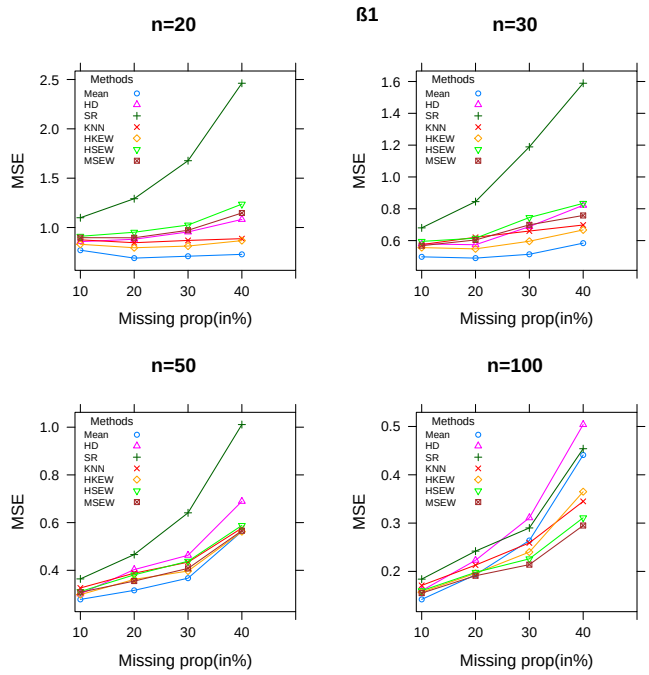

Figure 1 MSEs of $\hat{\beta}_1$ in different sample sizes, imputation methods, and missingness percentages

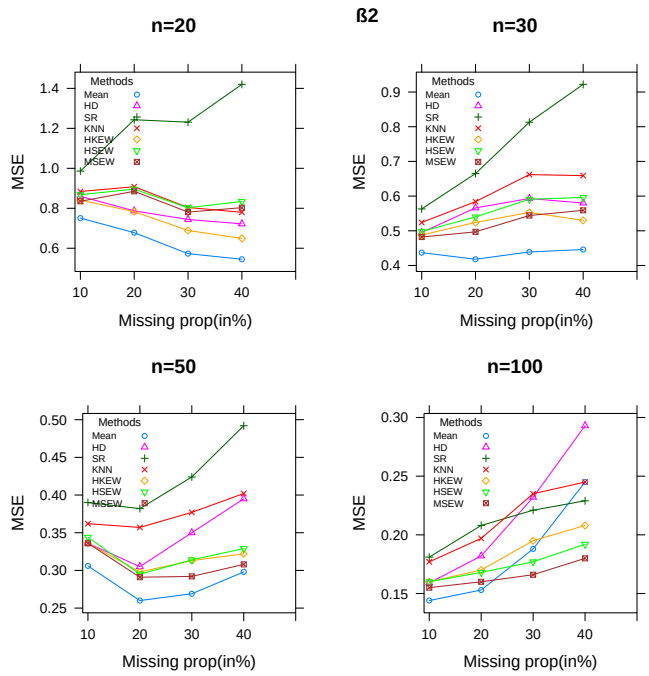

Figure 2 MSEs of $\hat{\beta}_2$ in different sample sizes, imputation methods, and missingness percentages

4. Application to Real Life Data

In this section, we apply all the studied imputation methods to real dataset, Wine Dateset. The dataset was extracted from UCI Machine Learning Repository (Lichman , 2013). To form the relationship between Nonflavanoid phenols, Color intensity and OD280/OD315 of diluted wines, multiple regression analysis was used. The results are shown in Tables 5-8 and Figures 3-4. The model validation of the error assumption for homoscedasticity was conducted using Breusch-Pagan Test. The results revealed that this data is heteroscedastic.

The results in Tables 5-8 and Figures 3-4 reveal that when the sample size and missing percentage are large the composite methods perform well in terms of both bias and MSEs compared with the mean and hot deck imputation methods. These results are in agreement with those in the simulation study. However, when the sample size is not large and missing percentage is lower, there is no exact method that performs well in all situations.

Table 5 Biases and MSEs obtained from different imputation methods with 10% missing values for wine dataset

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.155	0.028	-0.022	0.000	0.041	0.005
	Hot deck	0.036	0.004	0.131	0.015	-0.137	0.053
	Stochastic regression	0.207	0.050	0.116	0.012	-0.020	0.001
	KNN	0.372	0.160	0.093	0.008	0.120	0.040
	HKEW	0.217	0.055	0.112	0.011	-0.009	0.000
	HSEW	0.135	0.021	0.124	0.014	-0.078	0.017
	MSEW	0.181	0.038	0.047	0.002	0.010	0.003
30	Mean	-0.512	0.302	0.083	0.006	-0.329	0.304
	Hot deck	-0.976	1.101	0.212	0.040	-0.755	1.608
	Stochastic regression	-0.558	0.359	0.173	0.026	-0.420	0.498
	KNN	-0.563	0.366	0.295	0.077	-0.509	0.729
	HKEW	-0.769	0.684	0.253	0.057	-0.632	1.125
	HSEW	-0.767	0.679	0.192	0.033	-0.588	0.974
	MSEW	-0.535	0.330	0.128	0.014	-0.374	0.395
50	Mean	-0.205	0.048	-0.038	0.001	-0.105	0.031
	Hot deck	-0.235	0.064	-0.039	0.001	-0.149	0.063
	Stochastic regression	0.157	0.028	0.015	0.000	0.108	0.033
	KNN	0.025	0.001	0.119	0.013	-0.047	0.006
	HKEW	-0.105	0.013	0.040	0.001	-0.098	0.027
	HSEW	-0.039	0.002	-0.012	0.000	0.021	0.001
	MSEW	-0.024	0.001	-0.012	0.000	0.001	0.000
100	Mean	-0.292	0.099	-0.284	0.071	0.014	0.001
	Hot deck	-0.292	0.099	-0.308	0.084	0.031	0.003
	Stochastic regression	-0.070	0.006	-0.125	0.014	0.058	0.010
	KNN	-0.049	0.003	-0.100	0.009	0.059	0.006
	HKEW	-0.171	0.034	-0.204	0.037	0.045	0.006
	HSEW	-0.182	0.038	-0.217	0.042	0.045	0.006
	MSEW	-0.181	0.038	-0.204	0.037	0.036	0.004

Table 6 Biases and MSEs obtained from different imputation methods with 20% missing values for wine dataset

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.624	0.450	-0.255	0.058	0.509	0.730
	Hot deck	0.232	0.062	-0.107	0.010	0.182	0.094
	Stochastic regression	1.499	2.597	-0.469	0.194	1.290	4.686
	KNN	0.672	0.521	-0.037	0.001	0.414	0.482
	HKEW	0.452	0.236	-0.072	0.004	0.298	0.250
	HSEW	0.866	0.866	-0.288	0.073	0.736	1.526
	MSEW	1.062	1.302	-0.362	0.116	0.899	2.279
30	Mean	-0.760	0.667	0.064	0.004	-0.493	0.684
	Hot deck	-1.048	1.269	0.206	0.038	-0.833	1.955
	Stochastic regression	-0.292	0.099	0.173	0.026	-0.259	0.188
	KNN	-0.582	0.391	0.294	0.076	-0.498	0.700
	HKEW	-0.815	0.768	0.250	0.055	-0.666	1.248
	HSEW	-0.670	0.519	0.189	0.032	-0.546	0.839
	MSEW	-0.526	0.320	0.118	0.012	-0.376	0.397
50	Mean	-0.671	0.521	-0.376	0.125	-0.186	0.097
	Hot deck	-0.658	0.501	-0.382	0.129	-0.189	0.101
	Stochastic regression	0.103	0.012	0.116	0.012	-0.013	0.000
	KNN	-0.169	0.033	0.014	0.000	-0.100	0.028
	HKEW	-0.414	0.198	-0.184	0.030	-0.144	0.059
	HSEW	-0.278	0.089	-0.133	0.016	-0.101	0.029
	MSEW	-0.284	0.093	-0.130	0.015	-0.099	0.028
100	Mean	-0.349	0.140	-0.410	0.148	0.021	0.001
	Hot deck	-0.361	0.150	-0.449	0.178	0.047	0.006
	Stochastic regression	0.013	0.000	-0.221	0.043	0.150	0.063
	KNN	0.038	0.002	-0.097	0.008	0.088	0.022
	HKEW	-0.162	0.030	-0.273	0.066	0.068	0.013
	HSEW	-0.174	0.035	-0.335	0.099	0.099	0.027
	MSEW	-0.168	0.033	-0.315	0.088	0.086	0.021

Table 7 Biases and MSEs obtained from different imputation methods with 30% missing values for wine dataset

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	0.163	0.031	-0.284	0.072	0.171	0.083
	Hot deck	-0.011	0.001	-0.140	0.017	0.042	0.005
	Stochastic regression	0.805	0.749	-0.174	0.027	0.599	1.012
	KNN	0.505	0.294	0.028	0.001	0.177	0.088
	HKEW	0.247	0.070	-0.056	0.003	0.109	0.034
	HSEW	0.397	0.182	-0.157	0.022	0.321	0.290
	MSEW	0.484	0.271	-0.229	0.046	0.385	0.418
30	Mean	-1.117	1.441	-0.396	0.139	-0.366	0.376
	Hot deck	-1.139	1.498	0.061	0.003	-0.773	1.684
	Stochastic regression	0.753	0.656	1.236	1.353	-0.355	0.354
	KNN	-0.606	0.425	0.271	0.065	-0.494	0.689
	HKEW	-0.873	0.880	0.166	0.024	-0.634	1.132
	HSEW	-0.193	0.043	0.648	0.372	-0.564	0.896
	MSEW	-0.180	0.038	0.420	0.156	-0.360	0.365
50	Mean	-0.711	0.585	-0.645	0.368	-0.092	0.024
	Hot deck	-0.594	0.408	-0.635	0.357	-0.035	0.004
	Stochastic regression	0.294	0.100	-0.005	0.000	0.170	0.082
	KNN	-0.070	0.006	-0.212	0.040	0.075	0.017
	HKEW	-0.332	0.127	-0.423	0.159	0.021	0.001
	HSEW	0.150	0.026	-0.320	0.095	0.067	0.013
	MSEW	-0.209	0.050	-0.325	0.093	0.039	0.004
100	Mean	-0.692	0.553	-0.683	0.413	-0.057	0.009
	Hot deck	-0.557	0.385	-0.620	0.340	-0.021	0.001
	Stochastic regression	0.091	0.010	-0.247	0.054	0.168	0.080
	KNN	-0.067	0.005	-0.333	0.098	0.154	0.066
	HKEW	-0.322	0.120	-0.476	0.201	0.066	0.012
	HSEW	-0.243	0.068	-0.433	0.166	0.074	0.015
	MSEW	-0.300	0.104	-0.465	0.192	0.055	0.009

5. Conclusions

The purpose of this study is to analyse the performances of imputation methods for multiple regression with missing heteroscedastic data via simulation studies. In this article, we compared four single and three proposed composite imputation methods for missing data on dependent variable. Our simulation results indicate that the regression coefficient estimators obtained by the stochastic regression method perform well in almost all situations in terms of bias. However, in terms of MSE, when the sample size is small to medium, the mean imputation method outperforms other methods. On the other hand, when the sample size is large and the missing percentage is high which occurs frequently in this era (30-40%), the estimator obtained by the MSEW imputation method gives the smallest MSE.

Moreover, we can see that the proposed composite methods always perform better than the single methods: mean and hot deck method in terms of bias. In terms of MSE, for small sample size, the HKEW method always results in smaller MSE than that of hot deck, stochastic regression, and KNN methods. For large sample size, the composite methods always give smaller MSE than that of the previously mentioned methods.

In real life data, when the sample size and missing percentage are large the composite methods

Table 8 Biases and MSEs obtained from different imputation methods with 40% missing values for wine dataset

Sample Size	Imputation methods	β_0		β_1		β_2	
		Bias	MSE	Bias	MSE	Bias	MSE
20	Mean	-0.031	0.001	-0.805	0.574	0.418	0.492
	Hot deck	-0.127	0.019	-0.614	0.334	0.251	0.178
	Stochastic regression	0.074	0.006	-0.153	0.021	0.050	0.007
	KNN	0.920	0.977	-0.671	0.398	0.965	2.624
	HKEW	0.396	0.181	-0.642	0.365	0.608	1.042
	HSEW	-0.027	0.001	-0.384	0.130	0.151	0.064
	MSEW	0.022	0.000	-0.479	0.203	0.234	0.154
30	Mean	-1.101	1.401	-0.716	0.454	-0.164	0.076
	Hot deck	-0.679	0.533	-1.087	1.047	0.348	0.342
	Stochastic regression	0.673	0.524	1.497	1.983	-0.640	1.154
	KNN	-0.784	0.710	-0.262	0.061	-0.265	0.197
	HKEW	-0.732	0.619	-0.675	0.403	0.042	0.005
	HSEW	-0.003	0.000	-0.205	0.037	-0.146	0.060
	MSEW	-0.214	0.053	0.390	0.135	-0.402	0.456
50	Mean	-0.947	1.035	-0.850	0.639	-0.139	0.054
	Hot deck	-0.677	0.530	-0.843	0.629	0.064	0.012
	Stochastic regression	0.325	0.122	0.233	0.048	0.031	0.003
	KNN	-0.664	0.510	-0.526	0.248	-0.138	0.054
	HKEW	-0.671	0.520	-0.686	0.417	-0.037	0.004
	HSEW	-0.176	0.036	-0.306	0.082	0.048	0.006
	MSEW	-0.311	0.117	-0.308	0.084	-0.054	0.008
100	Mean	-1.064	1.309	-0.815	0.588	-0.256	0.185
	Hot deck	-0.797	0.738	-0.745	0.491	-0.148	0.062
	Stochastic regression	0.156	0.028	-0.214	0.041	0.241	0.164
	KNN	-0.477	0.231	-0.627	0.348	0.076	0.016
	HKEW	-0.622	0.448	-0.686	0.417	-0.036	0.004
	HSEW	-0.321	0.119	-0.480	0.024	0.046	0.006
	MSEW	-0.454	0.238	-0.514	0.234	-0.007	0.000

perform well in terms of both bias and MSEs compared with the mean and hot deck imputation methods. These results are in agreement with the those in the simulation study. However, when the sample size is small to medium there is no exact method that performs well in all situations. Some further studies such as a composite imputation method with more than two single imputation methods or the inequivalent weight method could be considered for better performances.

Acknowledgements

We would like to thank the referees for their comments and suggestions on the manuscript. We would also like to acknowledge Lasbela University of Agriculture, Water and Marine Sciences, Pakistan for awarding the scholarship under the project of "Development of Infrastructure" to the first author and Prince of Songkla University for supporting the facility.

References

Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. Int. Stat. Rev. 2010; 78(1): 40-64.

Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC. Med. Inform. Decis. Mak. 2016; 16(3): 74.

Buhi ER, Goodson P, Neilands TB. Out of sight, not out of mind: strategies for handling missing data. Am. J. Health. Behav. 2008; 32(1): 83-92.

Dettori JR, Norvell DC, Chapman JR. The sin of missing data: Is all forgiven by way of imputation?. Global. Spine. J. 2018; 8(8): 892-894.

Donders ART, Van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. 2006; 59(10): 1087-1091.

Elliott P, Hawthorne G. Imputing missing repeated measures data: How should we proceed? Aust. N. Z. J. Psychiatry. 2005; 39(7): 575-582.

Enders CK. Applied missing data analysis. New York: The Guilford Press; 2010.

Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. CMAJ. 2012; 184(11): 1265-1269.

Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials a practical guide with flowcharts. BMC. Med. Res. Methodol. 2017; 17(1): 162.

Jerez JM, Molina I, Garca-Laencina PJ, Alba E, Ribelles N, Martn M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 2010; 50(2): 105-115.

Lamjaisue R, Thongteeraparp A, Sinsomboonthong J. Comparison of missing data estimation methods for the multiple regression analysis with missing at random dependent variable. Science and Technology Journal. 2017; 25(5): 766-777.

Lichman M. UCI Machine Learning Repository. 2013. Available from: http://archive.ics.uci.edu/ml

Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley & Sons; 1987.

Lodder P. To Impute or not Impute, Thats the Question. In: Mellenbergh GJ, Adr HJ, editors. Advising on research methods: Selected topics 2013. Huizen: Johannes van Kessel Publishing; 2014.

Munguía T, Armando J. Comparison of imputation methods for handling missing categorical data with univariate pattern. Journal of Quantitative Methods for Economics and Business Administration. 2014; 17: 101-120.

Myers TA. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. Commun. Methods. Meas. 2011; 5(4): 297-310.

R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018.

Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R. Imputing missing data: a comparison of methods for social work researchers. Soc. Work. Res. 2006; 30(1): 19-31.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338(b2393): 157-160.

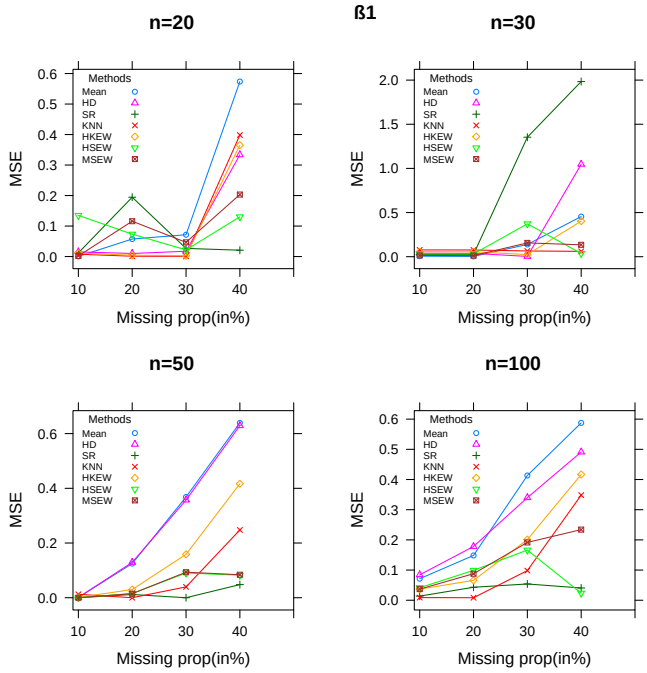

Figure 3 MSEs of $\hat{\beta}_1$ in different sample sizes, imputation methods, and missingness percentages for wine dataset

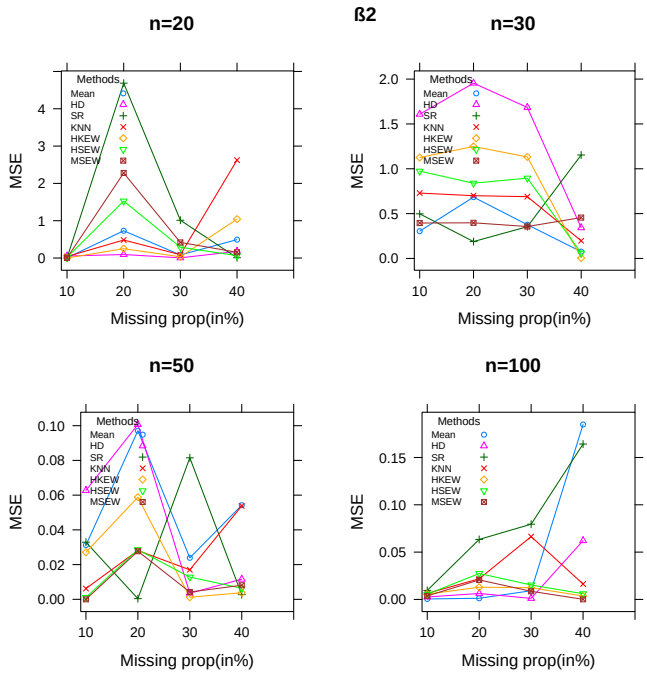

Figure 4 MSEs of $\hat{\beta}_2$ in different sample sizes, imputation methods, and missingness percentages for wine dataset