مجلة البحوث المالية والتجارية
المجلد (24) – العدد الثاني – إبريل 2023

Imputation for Missing Data in Statistical Matching Using Goal Programming

Abeer M. M. Elrefaey

Assistant Lecturer of Statistics, Faculty of Commerce, Al-Azhar University, Girls' Branch, Cairo, Egypt

Prof. Ramadan Hamed
Professor of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt.
Research Professor Social Research Center, AUC

Prof. Elham A. Ismail
Professor of Statistics, Faculty of Commerce, Al-Azhar University, Girls' Branch, Cairo, Egypt

Dr. Safia M. Ezzat
Lecturer of Statistics, Faculty of Commerce, Al-Azhar University, Girls' Branch, Cairo, Egypt

تاريخ الإرسال	2023-04-28
تاريخ القبول	2023-05-29
رابط المجلة:	https://jsst.journals.ekb.eg/

Abstract

Nearly all common statistical approaches assume complete information for all variables involved in the analysis, which making missing data problematic. Imputation is the process of substituting a missing value with a specific value, and it is most likely the most popular method for compensating for missing item values in a survey. This study suggests use of mathematical goal programming approach to impute missing data in statistical matching. The suggested approach adopts the regression method in imputation of the missing values. The regression coefficients are estimated using an estimated mathematical goal programming approach. The paper studies the cases when having variables with different skewed probability distributions (lognormal, Cauchy, chi square). The results of the simulation study indicate a good performance of the suggested approach in cases of skewed probability distribution. Using goal programming in regression is based on the minimizing the sum of absolute errors which is less affected by outliers compared to sum of squares of errors.

Keywords: Missing data, Imputation, Mathematical goal programming, Statistical matching, Missing completely at random, Missing at random.

1. Introduction

In statistics, imputation is the process of replacing missing data with substituted values. Missing data, also known as missing values, arise when data for certain variables or individuals is not collected, Enders (2022). Data can be lost for several causes including incomplete data entry, systems break down, and missing files. The reason for the missing data is important to consider because it helps to determine the type of missing data and what the researcher needs to do about it, Little and Rubin (2019).

Statistical matching methods combine two or more data sources (commonly of sample surveys data) on the same target population. Many of the techniques proposed for statistical matching at micro level are based on methods developed for the imputation of missing values: parametric (e.g., regression imputation), nonparametric (hot deck imputation) or mixed methods (e.g., methods based on predictive mean matching), D'Orazio et al. (2006).

Two data files, A and B, having common variables X are used in statistical matching. Files A and B have variables Y and Z, respectively. A single source cannot provide a file containing variables Y, Z, and X. One must then merge the two files such that the distributions of the variables of interest stay as unaltered as feasible, Kum and Masterson (2008).

The process of statistical matching for file merging can be viewed as a process of imputing Z values for the candidate records (X, Y) in file A using (X, Z) records from file B. It is assumed that the Z values are missing at random in the combined file. However, it differs from the usual imputation procedures because there are no files containing the complete set of values (X, Y, Z). Therefore, some additional techniques are required to estimate the conditional distribution $f(Z|X,Y)$ from records which in turn could be used for drawing imputed values, Rubin (1986). Two situations arise: Case I: Y Ignorable This corresponds to the assumption of conditional independence of Y and Z given X i.e., $f(Z|X,Y) = f(Z|X)$. Thus, the information in Y can be ignored and the problem of completing records with missing Z values in file A reduces to the usual imputation problem in a single file. Case II: Y Non-ignorable in this case, the Y information is not ignored in the process of statistical matching.

This study presented a mathematical goal programming technique to estimate the regression parameters to impute the missing data in statistical matching in single file.

2. Overview of Statistical Matching and Missing data

2.1. Statistical matching

Suppose there are two sample files, File A and File B, taken from two different surveys. Suppose further that File A contains potentially vector-valued variables (X, Y), while File B contains potentially vector-valued (X, Z). The objective of statistical matching is to combine these two files to obtain at least one file to obtain at least one file containing (X, Y, Z). "Statistical matching" as the technique began to be called, has been widely practiced since the advent of public use files in the 1960's. Arguably, the desire to employ statistical matching was even an impetus for the release of several of the early public use files, Moriarity and Scheuren (2001).

Wiest et al. (2019) described how they matched two data sets (the German Ageing Survey and the Study of Educational Attainment and Interests of Older Adults) in order to check the impact of educational involvement on later-life wellbeing. They focused on the matching proceedings and how to find the best-matched dataset. The impacts of educational activities on life satisfaction in later life are investigated using matched data. In quantitative research, the topic focuses on future data needs and methodologies for examining the broader advantages of adult learning. This paper showed that a real-world application of statistical matching allows us to deal with restricted secondary data in an efficient, inventive, and resourceful way. This paper showed that statistical matching with panel studies significantly enhance their analytic power. Small variation in measures may significantly enhance the suitable of matching variables and then overall quality of statistical matching. Hence, they presented cross-sectional studies take in consideration potential recipient panel-studies.

Conta et al. (2021) proposed the use of graphical models to deal with the statistical matching uncertainty for multivariate categorical variables. They gave the basics on Bayesian Networks (BN) and the concept of uncertainty in statistical matching when BNs are used is illustrated. They evaluated the performance of the proposed approach with and without auxiliary information and compared it with the saturated multinomial model in terms of uncertainty reduction. Finally,

the proposed methodology reveals a good performance in terms of uncertainty measure reduction, due to the role of qualitative auxiliary information.

D'Alberto et al. (2021) proposed the statistical matching benefit transfer approach (SMBT). This approach used the non-parametric micro statistical matching for benefit transfer. SMBT improved benefit transfer (BT) in its ability to reproduce the heterogeneity of the individuals' preferences at the policy site and thus to properly reproduce the true willingness to pay (WTP) distribution. They validated SMBT and compared it with both value and function transfer in accordance with BT. They transferred both the mean and the median WTP with value transfer while we apply the function transfer with both linear and Tobit model specifications.

Ahfock et al. (2022) established conditions where the maximum entropy model is minimax optimal in the file-matching problem. Their result can be used to motivate a conservative choice of an imputation model. Maximum likelihood estimation of the minimax optimal model can be carried out using data augmentation and the EM algorithm. Also, they found that minimax optimal strategy outperformed off-the-shelf imputation algorithms in the real data analysis due to the violation of the conditional independence assumption.

2.2. Missing data

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data is a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Missing data can occur because of nonresponse or no information is provided for one or more items or for a whole unit, Graham (2012).

Khan and Hoque (2020) proposed an algorithm SICE (Single Center Imputation from Multiple Chained Equation) for missing data imputation. Their approach is an extension of Multivariate Imputation by Chained Equation (MICE) algorithm in two different ways to impute categorical and numeric data. They used the UCI Machine Learning Repository, ETH Zurich, and kaggle to collected three public datasets. These datasets were used to compare their algorithms against with existing ones. SICE algorithm outperformed with the four datasets for binary and numeric data imputation. The results indicated that (SICE) algorithm is a perfect missing data imputation method, particularly for big datasets where (MICE) algorithm is too hard to utilize because of its

complication. But it fails to capture the essence of correlation and thus limits the scope of accurate prediction of missing values.

Yu et al. (2020) proposed a new regression multiple imputation (RMI) method. Their new method connects the multiple imputation (MI) method with the expectation maximization (EM) method. They used simulated studies and actual data to assess the effectiveness of the new suggested approach. As the iteration k of the iterative multiple imputation closes to infinity, the estimators are asymptotically efficient and converge point-wise for small m values.

Mahdy et al. (2021) discussed 10 imputation methods in the binary logistic regression model and then analysed the performance of these methods based on a medical application. They applied missing data in three cases: in x's only, in y only, and x's and y together. Akaki Information criterion (AIC), Bayesian information criterion (BIC), and R^2 criteria are used. The results showed that expectation-maximization (EM) and k-nearest neighbour imputation approaches are suitable for estimating missing values in this model, whether data are missing in dependent variables, independent variables, or both.

Thongsri and Samart (2022) developed a method for handling missing data in multiple linear regressions at random on both response and independent variables. They compared five techniques for missing data with the proposed composite imputation method: stochastic regression random forest with equivalent weight (SREW). Monte Carlo simulations were performed with sample sizes of 30, 60, 90, 120, and 150, missing percentages of 10%, 20%, 30%, and 40%, and standard deviations of error of 1, 3, and 5. The average mean square error is used to compare efficiency (AMSE). The results display that the SREW is the most efficient in all cases, while the hot deck has the greatest AMSE in virtually all cases, particularly when the missing percentage is significant.

Marcelinoa et al. (2022) suggested an experimental framework to assess impact of missing data. furthermore, the basis regression models - Decision Tree, Random Forests, Adaboost, K-Nearest Neighbours, Support Vector Machines, and Neural Networks - have been performance analysed. The result indicated that KNN outperformed others in regression models that contain missing data.

3. Suggested mathematical Goal Programming Approach

3.1. Linear regression model:

Let $x_{i0} = 1$ for i=1, 2, …, n. let $X_1, X_2, \dots, X_k$ be k independent random variables and let Y be a dependent random variable. Then linear relationship of the form:

$$y_i = \sum_{j=0}^{k} \beta_j \, x_{ij} + e_i \, , i = 1, 2, \dots \dots, n \qquad (1)$$

is assumed, where $\beta_0, \beta_1, \dots, \beta_k$ are the parameters to be estimated and $e_i (i = 1, 2, \dots, n)$ are the error components which are assumed to have skewed probability distribution. The linear absolute residuals method requires us to estimate the values of unknown parameters $\beta_0, \beta_1, \dots, \beta_k$ so as to

$$minimize \ \sum_{i=1}^{n} |y_i - \hat{y}_i|, \qquad (2)$$

Where:

$$\hat{y}_i = \sum_{j=0}^{k} \beta_j \, x_{ij} \, , \ i = 1, 2, \dots \dots, n \qquad (3)$$

$$\beta_j, \qquad j = 0, 1, \dots \dots, k$$

3.2. Linear goal programming approach for minimizing linear absolute deviations (LAD) in regression:

Let y_i be the ith goal, d_i^+ be positive deviation from the ith goal and d_i^- be negative deviation from the ith goal. Then the problem of minimizing

$$\sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} y_{diff} \qquad (4)$$

The problem may be reformulated as

$$min \ \sum_{i=1}^{n} (d_i^+ + d_i^-) \qquad (5)$$

Subject to:

$$\sum_{j=0}^{k} x_{ij} \beta_j + d_i^+ - d_i^- = y_i \, , i = 1, 2, \dots \dots, n \qquad (6)$$

$$\left| \frac{\bar{y}_{diff}}{s_{\bar{y}_{diff}}} \right| \leq t_{\left(\frac{\alpha}{2}, n-1\right)} \qquad (7)$$

$$d_i^+ \geq 0, i = 1, 2, \dots \dots, n$$

$$d_i^- \geq 0, i = 1, 2, \dots \dots, n$$

$$\beta_j \ unrestricted \ in \ sign, j = 0, 1, \dots \dots, k$$

It can prove that the objective function (5) is equivalent to (4) as follows:

From (6) $d_i^- = max(0, \hat{y}_i - y_i) = 0.5\left((\hat{y}_i - y_i) - |\hat{y}_i - y_i|\right)$

$$d_i^+ = max(0, y_i - \hat{y}_i) = 0.5\left((y_i - \hat{y}_i) - |y_i - \hat{y}_i|\right)$$

$$\sum_{i=1}^{n}(d_i^+ + d_i^-) = \sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Constraint 7 is necessary to have significant estimates of the regression parameters were.

4. Experimental design to statistical matching in one file:

In this study, the effect of missing value imputation in any data set problem is evaluated experimentally according to the following steps. First, incomplete data are filled with estimated values using the suggested goal programming in (5)–(7). Second, the performance of the suggested mathematical goal programming approach is measured using common statistical tests. The study used the suggested approach to impute or complete the missing data when data are missing in dependent variable in two designs:

4.1. The first design (using t test):

The first design is based on generating complete data according to the specified suggested sample sizes, then deleting several items equal to the suggested percentage of missing data. The suggested method is applied, and a complete set of data is obtained. To evaluate the approach, the two completed data sets before and after imputation are compared.

4.1.1. Steps of the first design:

1- **The study considers different samples sizes; 100, 300, and 500.**

2- **The study generates the data from three distributions: Lognormal (1.2,1), Chi square (2) and Cauchy (5,1) which are heavy tailed or peaked tailed distribution.**

3- **For each sample size and distribution, the 4 different percentages of missing data are considered (10%, 20%,30%, and 50%).**

4- **The number of observations is randomly deleted according to the specified percentages of missing data.**

5- **This simulation is repeated 10 replications.**

6- **The total number of replications is 3 sample sizes X 3 distributions X 4 percentages of missing values X 10 replications = 360 replications.**

7- **The suggested mathematical goal programming approach is used to estimate the regression coefficient and impute or complete the missing data for each replication.**

8- **For the two sets of samples (before and after the estimation), we determine mean, standard deviation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), then we are compared the values for all measured. Where MAE and RMSE are two criteria used to evaluate metric used with regression and the lower values of them are better:**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where

$n : sample\ size.$ $y_i : original\ value.$ $\hat{y}_i : imputed\ value.$

9- **We used the paired samples t test to determine whether the mean difference between two samples is zero for all combinations, and to evaluate the effectiveness of our suggested mathematical goal programming approach.**

where:

$$H_0: \mu_{org} = \mu_{imp} \qquad H_1: \mu_{org} \neq \mu_{imp}$$

Where

$$\mu_{org}\ is\ mean\ of\ the\ original\ sample,$$

$$\mu_{imp}\ is\ mean\ of\ the\ imputed\ sample.$$

4.1.2. Results for the first design:

Table (1) The mean and standard deviation of the dependent variable (y) for the two samples for Lognormal distribution (1.2,1)

	Replications	n = 100					n = 300					n = 500				
		Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample			
			10%	20%	30%	50%		10%	20%	30%	50%		10%	20%	30%	50%
Mean	replication 1	5.124	5.393	5.577	5.803	5.889	6.038	5.965	6.085	6.176	6.113	5.272	5.226	5.219	5.147	5.297
	replication 2	6.216	6.077	6.174	6.450	6.852	5.344	5.475	5.585	5.654	5.374	5.493	5.468	5.604	5.635	5.750
	replication 3	5.404	5.425	5.501	5.469	5.555	6.086	6.046	6.091	6.216	6.195	5.532	5.531	5.512	5.566	5.306
	replication 4	7.110	7.320	6.947	6.139	6.125	4.886	5.093	5.285	5.157	5.459	5.107	5.193	5.247	5.231	5.483
	replication 5	7.380	7.451	7.591	7.716	7.611	5.355	5.357	5.282	5.526	5.566	5.763	5.889	5.762	5.574	5.599
	replication 6	4.585	4.749	4.883	4.741	5.272	6.153	6.219	6.252	6.327	6.274	5.085	5.259	5.424	5.452	5.511
	replication 7	6.025	6.314	6.406	6.158	6.519	4.705	4.922	5.012	5.160	5.271	5.754	5.597	5.675	5.782	5.771
	replication 8	5.234	5.361	4.839	4.831	5.170	5.643	5.752	5.748	5.810	5.823	5.401	5.462	5.618	5.575	5.657
	replication 9	5.454	5.820	5.979	5.997	5.310	5.374	5.390	5.509	5.556	5.648	5.199	5.311	5.364	5.516	5.789
	replication 10	4.689	4.868	5.030	5.421	5.413	5.232	5.412	5.388	5.441	5.691	5.380	5.340	5.454	5.534	5.635
S.D.	replication 1	5.736	5.647	5.436	5.179	4.120	6.716	6.244	6.074	5.911	4.859	6.233	5.707	5.297	4.501	3.604
	replication 2	9.346	9.196	9.132	9.016	8.743	7.883	7.739	7.510	7.137	5.945	6.461	6.025	5.832	5.349	4.684
	replication 3	5.602	5.522	5.408	4.998	4.304	9.913	9.669	9.402	9.227	8.411	8.842	8.222	7.886	7.593	3.557
	replication 4	11.534	11.444	10.417	5.259	4.062	5.795	5.654	5.569	4.635	4.004	6.242	5.982	5.728	4.868	4.322
	replication 5	13.343	13.242	13.173	13.034	12.491	6.364	5.926	5.533	5.422	4.593	8.281	8.068	7.719	5.919	4.778
	replication 6	4.963	4.879	4.718	4.141	3.932	8.683	8.565	8.378	8.120	7.478	6.140	6.025	5.875	5.508	4.483
	replication 7	7.611	7.514	7.132	6.611	6.308	5.217	5.033	4.771	4.551	3.756	7.333	6.137	5.876	5.650	5.035
	replication 8	5.997	5.767	3.247	3.107	2.598	6.634	6.509	6.171	6.006	5.430	6.088	5.849	5.742	4.374	3.709
	replication 9	9.806	9.746	9.598	9.494	2.951	6.563	6.310	6.152	5.980	4.982	5.877	5.626	5.198	5.031	4.656
	replication 10	4.340	4.282	3.958	3.846	3.055	5.838	5.644	5.370	5.008	4.186	6.661	6.097	5.887	5.685	4.845

Table (2) The mean and standard deviation of the dependent variable (y) for the two samples for Chi-Square distribution (2)

	Replications	n = 100					n = 300					n = 500				
		Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample			
			10%	20%	10%	50%		10%	20%	30%	50%		10%	20%	30%	50%
Mean	replication 1	1.873	1.857	2.017	2.060	2.223	1.860	1.828	1.918	1.952	2.016	2.038	2.048	2.068	2.062	2.075
	replication 2	1.972	1.992	2.072	2.114	2.218	1.923	1.933	1.950	1.986	2.065	1.985	1.998	1.999	2.031	2.107
	replication 3	2.106	2.125	2.196	2.041	2.022	2.042	2.069	2.086	2.103	2.119	2.147	2.188	2.194	2.221	2.251
	replication 4	2.077	2.047	1.950	1.996	2.009	2.059	2.081	2.135	2.146	2.111	1.973	1.951	1.995	2.007	2.050
	replication 5	2.500	2.568	2.675	2.579	2.476	1.938	1.971	1.964	2.009	2.121	1.990	1.967	1.951	1.954	1.979
	replication 6	2.354	2.494	2.477	2.416	2.383	2.058	2.109	2.105	2.110	2.205	1.902	1.928	1.976	2.018	2.010
	replication 7	1.857	1.998	2.064	2.155	2.324	2.061	2.067	2.096	2.172	2.048	2.052	2.065	2.114	2.122	2.216
	replication 8	2.011	2.087	2.181	2.223	2.119	1.752	1.779	1.845	1.870	1.976	2.123	2.163	2.179	2.165	2.204
	replication 9	2.177	2.206	2.287	2.262	2.341	2.037	2.032	2.052	2.036	2.033	1.968	2.003	2.059	2.030	2.082
	replication 10	2.216	2.178	2.137	2.047	2.060	2.010	2.008	2.035	2.073	2.106	2.010	2.061	2.099	2.113	2.190
S.D.	replication 1	1.624	1.517	1.452	1.368	1.226	1.890	1.791	1.749	1.692	1.364	1.987	1.891	1.791	1.679	1.320
	replication 2	2.007	1.986	1.970	1.894	1.561	1.915	1.810	1.682	1.581	1.397	1.896	1.826	1.672	1.604	1.426
	replication 3	2.046	1.884	1.824	1.596	1.164	1.852	1.737	1.591	1.527	1.258	1.947	1.885	1.736	1.656	1.432
	replication 4	1.854	1.722	1.393	1.208	1.045	2.099	2.024	1.918	1.744	1.309	2.031	1.913	1.840	1.688	1.479
	replication 5	2.621	2.564	2.506	2.102	1.782	1.828	1.704	1.577	1.519	1.361	1.942	1.815	1.637	1.524	1.294
	replication 6	2.429	2.374	2.303	2.215	1.948	1.903	1.851	1.764	1.666	1.509	1.876	1.774	1.729	1.664	1.196
	replication 7	1.895	1.852	1.802	1.720	1.571	2.177	2.008	1.942	1.855	1.316	2.300	2.155	2.070	1.849	1.667
	replication 8	1.924	1.887	1.829	1.586	1.218	1.697	1.619	1.526	1.367	1.214	2.237	2.141	2.063	1.933	1.541
	replication 9	1.812	1.762	1.686	1.496	1.201	1.988	1.906	1.818	1.724	1.457	1.980	1.848	1.777	1.597	1.362
	replication 10	2.104	1.931	1.727	1.439	1.112	1.955	1.843	1.765	1.688	1.376	2.020	1.952	1.893	1.788	1.537

Table (3) The mean and standard deviation of the dependent variable (y) for the two samples for Cauchy distribution (5,1)

	Replications	n = 100					n = 300					n = 500				
		Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample				Y Original sample	Y Imputed sample			
			10%	20%	30%	50%		10%	20%	30%	50%		10%	20%	30%	50%
Mean	replication 1	5.738	5.279	4.909	4.629	4.838	3.959	3.933	4.019	3.838	3.691	5.953	5.710	5.454	5.310	4.942
	replication 2	3.162	3.451	3.351	3.243	2.912	4.392	4.447	4.526	4.573	4.496	6.309	6.215	6.045	6.002	5.366
	replication 3	5.365	4.980	4.757	4.635	3.994	4.257	4.219	4.281	4.336	4.101	3.366	3.114	2.804	2.481	3.795
	replication 4	4.703	4.612	4.582	3.672	3.664	4.452	4.727	4.754	4.763	4.808	5.190	5.042	4.825	4.691	4.269
	replication 5	4.205	4.161	4.209	4.148	3.894	6.638	6.231	6.225	6.000	5.817	5.923	5.759	5.599	5.390	4.929
	replication 6	4.943	4.731	4.548	4.413	4.193	5.797	5.648	5.812	5.972	5.132	-11.417	-11.608	-11.587	-11.853	-12.128
	replication 7	1.625	2.923	2.576	2.491	2.197	6.466	6.531	6.448	6.247	5.790	4.655	4.444	4.434	4.107	4.121
	replication 8	1.120	1.019	1.867	1.605	1.164	2.229	2.294	2.028	1.871	2.237	4.811	4.672	4.866	4.643	4.031
	replication 9	5.330	5.238	5.134	4.897	4.649	6.982	6.741	6.775	6.652	4.904	5.019	4.834	4.822	4.513	4.390
	replication 10	4.725	4.580	4.568	4.653	4.412	3.392	3.320	3.489	3.493	3.487	5.587	5.078	4.939	4.767	4.182
S.D.	replication 1	7.842	7.645	7.283	7.208	6.178	21.697	21.685	21.655	14.770	21.226	20.228	19.905	19.887	19.769	19.573
	replication 2	8.531	7.838	7.823	7.807	7.330	15.037	14.958	14.855	16.817	14.659	20.806	20.659	20.407	20.280	20.096
	replication 3	6.294	5.992	5.045	4.987	3.173	16.975	16.941	16.861	8.182	15.359	30.819	30.775	30.517	30.444	14.044
	replication 4	8.722	8.707	8.589	6.448	6.352	10.744	8.645	8.362	24.145	7.504	24.911	24.839	24.838	24.646	24.543
	replication 5	5.290	5.266	5.091	5.028	4.966	24.685	24.426	24.403	18.207	23.917	49.509	49.507	49.479	49.478	49.356
	replication 6	3.777	3.724	3.501	3.459	3.313	18.880	18.778	18.444	20.539	8.341	328.331	328.31	328.29	328.28	328.24
	replication 7	27.821	23.591	23.494	23.479	23.448	21.175	21.125	20.983	35.772	20.112	10.677	10.646	10.094	9.979	8.682
	replication 8	25.601	25.583	23.862	23.807	23.697	36.157	36.033	35.835	25.055	35.610	13.133	13.046	9.654	9.570	9.203
	replication 9	5.793	5.800	5.804	5.775	5.762	27.117	25.159	25.123	21.384	10.334	17.053	17.035	16.552	16.512	13.240
	replication 10	4.701	4.607	4.580	4.242	3.912	22.016	21.747	21.562	21.549	21.285	11.249	9.056	8.916	8.719	7.081

Tables (1-3) present the values of mean and standard deviation for the dependent variable Y before and after imputation for all cases considered. The results indicate that the values of mean, and the standard deviation are approximately the same.

Table (4) The RMSE and MAE for Lognormal distribution

	Replications	n = 100				n = 300				n = 500			
		percentages of missing				percentages of missing				percentages of missing			
		10%	20%	30%	50%	10%	20%	30%	50%	10%	20%	30%	50%
RMSE	replication 1	1.210	2.007	2.636	4.094	2.483	2.874	3.192	4.641	2.493	3.276	4.287	5.081
	replication 2	1.660	1.985	2.442	3.217	1.534	2.435	3.383	5.177	2.330	2.792	3.631	4.457
	replication 3	0.961	1.504	2.543	3.599	2.180	3.131	3.607	5.228	3.250	3.993	4.529	8.075
	replication 4	1.268	4.964	10.259	10.783	1.420	1.800	3.541	4.282	1.817	2.518	3.924	4.544
	replication 5	1.581	1.975	2.670	4.602	2.318	3.125	3.350	4.421	1.868	2.999	5.785	6.758
	replication 6	1.143	1.790	2.809	3.282	1.403	2.251	3.038	4.383	1.271	1.882	2.779	4.247
	replication 7	1.204	2.631	3.751	4.224	1.535	2.261	2.721	3.775	4.007	4.384	4.672	5.325
	replication 8	1.692	4.929	5.016	5.367	1.291	2.438	2.821	3.808	1.699	2.054	4.247	4.842
	replication 9	1.247	2.111	2.539	9.294	1.805	2.311	2.725	4.285	1.730	2.767	3.079	3.631
	replication 10	0.991	2.004	2.369	3.317	1.554	2.328	3.035	4.112	2.677	3.128	3.488	4.584
MAE	replication 1	0.321	0.782	1.239	2.167	0.465	0.819	1.149	2.195	0.494	0.967	1.474	2.277
	replication 2	0.369	0.663	1.063	1.912	0.396	0.823	1.258	2.366	0.506	0.904	1.382	2.262
	replication 3	0.247	0.557	1.088	1.889	0.485	0.918	1.277	2.312	0.510	1.011	1.471	2.607
	replication 4	0.374	1.432	2.743	3.724	0.400	0.691	1.266	1.994	0.403	0.809	1.273	2.047
	replication 5	0.346	0.686	1.110	2.114	0.486	0.949	1.286	2.144	0.408	0.885	1.445	2.486
	replication 6	0.334	0.712	1.235	1.916	0.375	0.814	1.231	2.104	0.348	0.717	1.115	2.045
	replication 7	0.318	0.822	1.489	2.254	0.402	0.803	1.228	2.088	0.609	1.063	1.462	2.309
	replication 8	0.452	1.441	1.664	2.424	0.332	0.781	1.157	1.990	0.399	0.715	1.234	2.005
	replication 9	0.369	0.809	1.192	2.713	0.404	0.805	1.170	2.089	0.400	0.823	1.173	1.883
	replication 10	0.297	0.738	1.129	1.931	0.443	0.859	1.312	2.075	0.451	0.851	1.247	2.144

Table (5) The RMSE and MAE for Chi-Square distribution

	Replications	n = 100				n = 300				n = 500			
		percentages of missing				percentages of missing				percentages of missing			
		10%	20%	30%	50%	10%	20%	30%	50%	10%	20%	30%	50%
RMSE	replication 1	0.557	0.778	0.920	1.127	0.588	0.746	0.882	1.339	0.610	0.867	1.066	1.488
	replication 2	0.329	0.469	0.751	1.333	0.623	0.917	1.086	1.326	0.544	0.730	0.959	1.330
	replication 3	0.811	0.959	1.272	1.676	0.648	0.952	1.053	1.364	0.517	0.896	1.020	1.267
	replication 4	0.664	1.184	1.384	1.513	0.560	0.856	1.169	1.643	0.494	0.887	1.029	1.321
	replication 5	0.520	0.706	1.559	1.918	0.669	0.929	1.032	1.246	0.670	0.865	1.136	1.404
	replication 6	0.481	0.757	0.984	1.447	0.447	0.710	0.916	1.160	0.681	1.034	1.194	1.445
	replication 7	0.521	0.712	0.914	1.174	0.844	0.988	1.149	1.736	0.620	0.755	0.901	1.464
	replication 8	0.434	0.661	1.118	1.491	0.533	0.791	1.053	1.250	0.807	1.013	1.373	1.595
	replication 9	0.436	0.684	1.031	1.352	0.555	0.800	0.984	1.351	0.653	0.870	1.130	1.626
	replication 10	0.842	1.193	1.519	1.762	0.653	0.842	0.995	1.395	0.722	0.897	1.183	1.453
MAE	replication 1	0.163	0.327	0.472	0.736	0.156	0.283	0.406	0.700	0.150	0.312	0.463	0.812
	replication 2	0.086	0.183	0.346	0.768	0.157	0.317	0.464	0.732	0.134	0.313	0.443	0.716
	replication 3	0.225	0.367	0.594	0.978	0.149	0.320	0.438	0.765	0.132	0.274	0.409	0.717
	replication 4	0.173	0.405	0.576	0.806	0.144	0.296	0.443	0.833	0.143	0.286	0.456	0.766
	replication 5	0.144	0.277	0.491	0.858	0.190	0.364	0.482	0.758	0.168	0.345	0.489	0.796
	replication 6	0.140	0.277	0.447	0.812	0.124	0.263	0.420	0.697	0.157	0.269	0.400	0.777
	replication 7	0.153	0.280	0.444	0.748	0.187	0.332	0.500	0.890	0.178	0.340	0.533	0.820
	replication 8	0.119	0.265	0.473	0.798	0.132	0.281	0.452	0.705	0.159	0.310	0.489	0.801
	replication 9	0.116	0.264	0.452	0.797	0.132	0.273	0.416	0.731	0.157	0.300	0.486	0.782
	replication 10	0.208	0.402	0.649	0.975	0.164	0.309	0.453	0.799	0.148	0.279	0.439	0.756

Table (6) The RMSE and MAE for Cauchy distribution

	Replications	n = 100				n = 300				n = 500			
		percentages of missing				percentages of missing				percentages of missing			
		10%	20%	30%	50%	10%	20%	30%	50%	10%	20%	30%	50%
RMSE	replication 1	2.167	3.315	3.570	5.124	0.725	1.491	2.592	4.576	3.824	4.091	4.711	5.613
	replication 2	3.365	3.392	3.425	4.305	1.572	2.297	2.819	3.367	2.581	4.260	4.855	5.875
	replication 3	2.204	3.964	4.067	5.657	1.176	2.027	2.424	7.311	1.709	4.363	4.845	27.386
	replication 4	0.724	1.601	5.935	6.037	6.459	6.817	7.033	7.769	2.094	2.299	3.932	4.694
	replication 5	0.562	1.426	1.638	1.861	3.863	4.032	5.395	6.375	1.088	2.172	2.457	4.466
	replication 6	0.961	1.700	1.849	2.145	2.126	3.981	4.944	16.917	2.319	4.370	4.510	6.498
	replication 7	14.776	14.903	14.920	14.945	1.429	2.982	5.313	6.888	1.195	3.552	3.976	6.299
	replication 8	0.678	9.457	9.541	9.709	3.018	4.663	5.047	6.244	1.681	8.836	8.980	9.448
	replication 9	0.472	0.722	1.228	1.482	10.157	10.251	10.438	25.087	1.190	4.211	4.480	10.860
	replication 10	1.066	1.188	2.061	2.697	3.498	4.604	5.406	5.790	6.847	7.063	7.358	9.080
MAE	replication 1	0.459	0.896	1.178	2.003	0.177	0.424	0.831	1.419	0.602	0.951	1.332	2.104
	replication 2	0.455	0.583	0.705	1.276	0.263	0.579	0.836	1.400	0.397	0.863	1.247	2.125
	replication 3	0.420	0.944	1.171	1.927	0.244	0.499	0.755	1.830	0.347	0.983	1.339	3.909
	replication 4	0.185	0.485	1.422	1.852	0.628	0.920	1.216	1.759	0.404	0.651	1.154	1.803
	replication 5	0.155	0.391	0.585	0.915	0.563	0.763	1.208	1.775	0.275	0.597	0.885	1.789
	replication 6	0.247	0.533	0.708	1.109	0.326	0.788	1.192	2.639	0.468	0.964	1.245	2.231
	replication 7	1.817	2.184	2.359	2.711	0.274	0.649	1.187	1.939	0.301	0.815	1.177	2.072
	replication 8	0.165	1.264	1.548	2.125	0.343	0.767	1.034	1.755	0.349	1.092	1.447	2.214
	replication 9	0.123	0.245	0.482	0.818	1.044	1.299	1.569	3.659	0.256	0.832	1.149	2.228
	replication 10	0.257	0.377	0.759	1.366	0.451	0.838	1.177	1.702	0.698	1.058	1.425	2.274

Tables (4-6) present the values of MAE and RMSE for all cases considered. The results indicate for all combinations are likely very good value.

Table (7) The P – values for mean for all distribution to all samples size

	Replications	n = 100				n = 300				n = 500			
		percentages of missing				percentages of missing				percentages of missing			
		10%	20%	30%	50%	10%	20%	30%	50%	10%	20%	30%	50%
Lognormal	replication 1	.0253*	.0233*	.0092**	.0613	.6129	.7770	.4563	.7810	.6821	.7205	.5149	.9123
	replication 2	.4048	.8332	.3403	.0475*	.1387	.0862	.1120	.9201	.8155	.3742	.3808	.1968
	replication 3	.8271	.5226	.7993	.6776	.7487	.9804	.5331	.7185	.9911	.9104	.8674	.5316
	replication 4	.0983	.7440	.3466	.3634	.0115*	.0001**	.1867	.0203*	.2907	.2148	.4809	.0646
	replication 5	.6570	.2877	.2105	.6191	.9893	.6887	.3763	.4100	.1309	.9941	.4648	.5882
	replication 6	.1507	.0959	.5806	.0356*	.4201	.4458	.3238	.6332	.0022**	.0001**	.0247*	.3809
	replication 7	.0157*	.1491	.7261	.2447	.0145*	.0186*	.0036**	.0092**	.3809	.6881	.8930	.9431
	replication 8	.4568	.4255	.4240	.9051	.1408	.4540	.3036	.4134	.4265	.0184*	.3600	.2387
	replication 9	.0029**	.0122*	.0319*	.8780	.8795	.3142	.2481	.2694	.1483	.1845	.0211*	.0003**
	replication 10	.0706	.0883	.0017**	.0282*	.0451*	.2448	.2341	.0528	.7413	.5998	.3245	.2143
Chi-Square	replication 1	0.776	0.064	0.041*	0.002**	0.345	0.184	0.072	0.044*	0.724	0.437	0.621	0.577
	replication 2	0.540	0.031*	0.057	0.064	0.786	0.612	0.321	0.063	0.591	0.726	0.314	0.031*
	replication 3	0.813	0.352	0.611	0.617	0.466	0.425	0.313	0.326	0.061	0.232	0.108	0.077
	replication 4	0.649	0.284	0.562	0.655	0.480	0.120	0.197	0.579	0.469	0.559	0.495	0.220
	replication 5	0.190	0.012*	0.613	0.903	0.403	0.636	0.237	0.011*	0.452	0.395	0.505	0.866
	replication 6	0.003**	0.104	0.532	0.845	0.049*	0.258	0.330	0.029*	0.355	0.030*	0.004**	0.100
	replication 7	0.006**	0.003**	0.001**	0.000**	0.913	0.542	0.096	0.899	0.714	0.170	0.255	0.021*
	replication 8	0.077	0.009**	0.058	0.469	0.388	0.043*	0.053	0.002**	0.174	0.149	0.412	0.266
	replication 9	0.498	0.108	0.412	0.226	0.879	0.744	0.996	0.962	0.287	0.023*	0.243	0.080
	replication 10	0.660	0.511	0.269	0.378	0.968	0.608	0.275	0.231	0.035*	0.006**	0.016	0.002**
Cauchy	replication 1	.033*	.012*	.002**	.079	.532	.492	.418	.310	.156	.006**	.002**	.000**
	replication 2	.392	.581	.814	.565	.542	.312	.267	.592	.417	.165	.158	.000
	replication 3	.080	.126	.073	.015*	.579	.839	.571	.712	.001**	.004*	.000**	.727
	replication 4	.212	.450	.082	.085	.462	.445	.446	.429	.114	.000**	.004**	.000**
	replication 5	.436	.976	.729	.095	.068	.077	.040*	.025*	.001**	.001**	.000**	.000**
	replication 6	.027*	.019*	.004**	.000**	.226	.948	.541	.497	.066	.386	.031*	.014*
	replication 7	.382	.526	.564	.704	.435	.918	.476	.089	.000**	.164	.002**	.058
	replication 8	.137	.432	.614	.964	.711	.456	.219	.981	.064	.889	.676	.065
	replication 9	.050	.006**	.000**	.000**	.682	.727	.585	.152	.000**	.297	.012*	.195
	replication 10	.175	.189	.730	.248	.725	.715	.746	.775	.096	.040*	.013*	.001**
The percentage of accepting the null hypothesis													
significance level	$\alpha = 0.05$	77%	70%	73%	73%	87%	90%	93%	77%	80%	67%	60%	67%
	$\alpha = 0.01$	87%	87%	80%	83%	97%	93%	93%	90%	80%	77%	77%	73%

* Significant at the 0.05 level

**Significant at the 0.01 level

Table (7) displays p-values of the paired t test for both samples (the original sample and imputed sample) for all combinations. The results indicate that the mean for the two samples is equal in most cases for all 10 numbers of replications. The average ratio of accepting the null hypothesis is 76% at significance level 0.05 and 85% at significance level 0.01.

4.2. The second design (using coefficient of determination R^2):

This design is based on the same steps as in the first design, except that the samples are generated first without the missing values. For each case, the regression model is repeated 100 times and the one with the highest value of R^2 is selected for imputation.

4.2.1. Steps of the second design:

1- **The considered sample sizes are 100, 500, and 1000.**
2- **The study generates incomplete samples depending on the suggested percentage of missing data 10%, 30% and 50%.**
3- **The study generates the data distributed from three distributions: Lognormal (1.2,1), Chi square (2) and Cauchy (5,1).**
4- **The suggested mathematical goal programming approach is used to estimate the regression coefficients 100 times and compute R^2 each time. For each case, select the model with the highest level of R^2.**
5- **Use the selected model is to impute or complete the missing data to complete the samples sizes.**
6- **Then we obtained two unequal samples size: an original sample (incomplete) and other sample (after addition number of observation equal percentages of missing data) for all combinations.**
7- **For the two samples, we determine mean and standard deviation to evaluate the performance of the suggested mathematical goal programming approach.**
8- **The effectiveness of the suggested mathematical goal programming approach is measured using the t test and the F test to test the equality between the two means and two variances, respectively.**

4.2.2. Results for the second design:

Table (8) The mean and P – values for all combinations for second design

percentages of missing		10%			30%			50%		
Sample size	distribution	Mean before	Mean after	P-value	Mean before	Mean after	P-value	Mean before	Mean after	P-value
100	Lognormal	10.715	10.602	0.922	10.208	10.718	0.602	9.845	8.808	0.428
	Chi - square	6.908	6.953	0.931	6.598	6.687	0.876	7.056	7.141	0.942
	Cauchy	8.690	8.945	0.800	10.397	7.002	0.124	11.456	8.168	0.257
500	Lognormal	10.157	10.250	0.833	9.814	9.946	0.796	9.962	10.326	0.465
	Chi - square	6.779	6.749	0.905	6.712	6.781	0.785	6.980	6.940	0.880
	Cauchy	9.066	9.210	0.918	14.847	13.106	0.750	11.659	11.148	0.777
1000	Lognormal	10.542	10.376	0.612	10.058	10.101	0.898	10.915	10.461	0.277
	Chi - square	7.224	7.149	0.645	6.897	6.882	0.934	7.030	6.941	0.646
	Cauchy	10.841	10.650	0.907	9.156	8.896	0.852	12.523	10.820	0.607

Table (9) The standard deviation and P – values for all combinations for second design

percentages of missing		10%			30%			50%		
Sample size	distribution	Variance before	Variance after	P-value	Variance before	Variance after	P-value	Variance before	Variance after	P-value
100	lognormal	8.038	7.846	0.812	5.895	6.789	0.214	5.800	10.206	0.000**
	Chi - square	3.552	3.651	0.794	3.755	3.448	0.433	4.065	9.990	0.000**
	cauchy	7.061	6.790	0.703	7.820	20.007	0.000**	13.411	21.925	0.000**
500	lognormal	6.740	6.873	0.672	7.339	7.321	0.956	6.280	6.741	0.205
	Chi - square	3.857	3.863	0.974	3.591	3.665	0.684	3.394	3.432	0.849
	cauchy	22.005	20.936	0.278	80.840	74.830	0.115	24.915	19.700	0.000**
1000	lognormal	7.155	7.090	0.778	6.929	6.701	0.334	7.788	7.276	0.076
	Chi - square	3.525	3.565	0.729	3.676	3.641	0.781	3.528	3.563	0.806
	cauchy	36.296	34.554	0.130	26.841	29.996	0.002**	65.856	47.810	0.000**

*** Significant at the 0.05 level**

****Significant at the 0.01 level**

Tables (8-9) show the means, variances, and the p-values for the complete and incomplete data sets lognormal distribution, Chi-Square, and Cauchy distributions. The results for p-values indicate that the mean for the two samples is equal at all cases in all percentages of missing and all distributions used, while the variance for the two samples is equal at most of cases in all combinations. The average ratio of accepting the null hypothesis for mean is 100 % and for the variance is 74 % at significance level 0.05 and 0.01.

5. Conclusions

The paper suggests using mathematical goal programming approach to impute the missing values by estimating the regression coefficients. The results of the simulation study indicate a good performance of the suggested approach in cases of skewed probability distribution. The differences between means of data before and after imputation are insignificant. The differences between the variances are almost insignificant. This approach is suggested to be used in statistical matching. Considering the equality of variances and other measures of skewness and kurtosis can be considered in further research.

References

Ahfock, D., Pyne, S. and McLachlan, G.J., 2022. Statistical file-matching of non-Gaussian data: A game theoretic approach. ***Computational Statistics & Data Analysis*****,** ***168*****, p.107387.**

Conti, P.L., Marella, D., Vicard, P. and Vitale, V., 2021. Multivariate statistical matching using graphical modeling. ***International Journal of Approximate Reasoning*****,** ***130*****, pp.150-169.**

D'Alberto, R., Zavalloni, M., Raggi, M. and Viaggi, D., 2021. A Statistical Matching approach to reproduce the heterogeneity of willingness to pay in benefit transfer. ***Socio-Economic Planning Sciences*****,** ***74*****, p.100935.**

D'Orazio, M., Di Zio, M. and Scanu, M., 2006. ***Statistical matching: Theory and practice*****. John Wiley & Sons.**

Enders, C.K., 2022. ***Applied missing data analysis*****. Guilford Publications.**

Graham, J.W., 2012. ***Missing data: Analysis and design*****. Springer Science & Business Media.**

Khan, S.I. and Hoque, A.S.M.L., 2020. SICE: an improved missing data imputation technique. ***Journal of big Data*****,** ***7*****(1), pp.1-21.**

Kum, H. and Masterson, T., 2008. Statistical matching using propensity scores: Theory and application to the levy institute measure of economic well-being.

Little, R.J. and Rubin, D.B., 2019. ***Statistical analysis with missing data*** **(Vol. 793). John Wiley & Sons.**

Mahdy, S.M., Abonazel, M.R. and Ghallab, M.G., 2021, October. A Review of Ten Imputation Methods for Handling Missing Values in Logistic Regression: A Medical Application. In ***Science Forum (Journal of Pure and Applied Sciences)*** **(Vol. 21, No. 3, pp. 434-434). Faculty of Science, Abubakar Tafawa Balewa University Bauchi.**

Marcelino, C.G., Leite, G.M.C., Celes, P. and Pedreira, C.E., 2022. Missing data analysis in regression. ***Applied Artificial Intelligence*****,** ***36*****(1), p.2032925.**

Moriarity, C. and Scheuren, F., 2001. Statistical matching: a paradigm for assessing the uncertainty in the procedure. ***Journal of Official Statistics*****,** ***17*****(3), p.407.**

Rubin, D.B., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. ***Journal of Business & Economic Statistics*****, *4*(1), pp.87-94.**

Thongsri, T. and Samart, K., 2022. Composite Imputation Method for the Multiple Linear Regression with Missing at Random Data. ***Computer Science*****, *17*(1), pp.51-62.**

Wiest, M., Kutscher, T., Willeke, J., Merkel, J., Hoffmann, M., Kaufmann-Kuchta, K. and Widany, S., 2019. The potential of statistical matching for the analysis of wider benefits of learning in later life. ***European journal for Research on the Education and Learning of Adults*****, *10*(3), pp.291-306.**

Yu, L., Liu, L. and Peace, K.E., 2020. Regression multiple imputation for missing data analysis. ***Statistical Methods in Medical Research*****, *29*(9), pp.2647-2664.**