Statistics
Innovation

ARTICLE

Open Access

https://doi.org/10.48130/stati-0024-0002
Statistics Innovation **2024**, 1: e002

SAS MACRO for forward model selection in a logistic regression model with missing values

Mingyue Han#, Xueqing Zhang# and ChangXing Ma*

Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, NY 14221, USA
# Authors contributed equally: Mingyue Han, Xueqing Zhang
* Corresponding author, E-mail: cxma@buffalo.edu

Abstract

Missing data often presents significant challenges in statistical modeling. Previous practices have demonstrated that the SAS LOGISTIC procedure for forward, backward, or stepwise selections may produce unreliable results when a substantial portion of data is missing in logistic regression models. To address this issue, a new Macro-based Forward Model Selection method (MFMS) was developed to overcome the limitations of the LOGISTIC procedure. The MFMS approach preserves information from incomplete observations by Macro, automated controlled addition of one significant effect at a time while retaining incomplete observations for other variables in the analysis. To evaluate the robustness of MFMS, simulations were conducted using independent multivariate normal data with a binary response. The complete data was randomly deleted at rates of 5%, 10%, 15%, and 20% for each predictor variable, generating datasets under three mechanisms. The performance of MFMS, PROC LOGISTIC, and MI procedures were compared for logistic modeling across different levels of missingness and mechanisms. The results demonstrated that MFMS consistently outperformed the other methods, proving to be the most reliable model selection method. The Macro implementation of MFMS is available for download as a supplementary file to this paper.

Citation: Han M, Zhang X, Ma C. 2024. SAS MACRO for forward model selection in a logistic regression model with missing values. *Statistics Innovation* 1: e002 https://doi.org/10.48130/stati-0024-0002

Introduction

Missing data have long been a serious impediment in statistical analysis and modeling in research and business. The logistic regression model is a standard statistical method used in analysis when the response variables are dichotomous or binary[1]. The present work aims to demonstrate the advantages or disadvantages of logistic regression modeling methods among SAS PROC LOGISTIC forward selection, SAS MI procedure, and the developed macro-based forward selection method, using the artificial binary response data with the presence of random missing values.

Missing data mechanisms

According to the generating mechanisms, missing data can be classified into three board types: missing completely at random (MCAR) if missing data are completely independent of any other factors; missing at random (MAR) describes the missingness based on other complete variables; not missing at random (NMAR) if the probability of missing data systematically depends on incomplete variables. According to missing data mechanisms, distribution assumptions are made in advance indicating the modeling methods. To specify, missing data does not represent the entire data. Thus, the modeling assumption could be MCAR, MAR, or NMAR depending on the variables of interest[2−5]. MCAR is possible to test in practice empirically.

Deletion and single imputation

Traditionally, missing data are deleted or replaced with single imputations[5]. Maximum likelihood and multiple imputation are advantageous[6]. The deletion method includes two approaches: listwise deletion and pairwise deletion. In SAS, the default effect selection of the LOGISTIC procedure listwise deletes any observation with missing values in independent and response variables. The analysis is restricted to the remaining complete data set using standard procedures[7]. Listwise deletion of missing values is under the assumption of MCAR and results in bias estimates since MCAR is rarely satisfied[5]. In addition, deleting missing values severely reduces the analyzable sample size, particularly for data with a high proportion of missing values or missing in many variables. This results in the loss of power of the significance test[8].

Single imputation is another traditional approach to processing missing data. The most widely used single imputation methods include mean imputation, regression imputation, and stochastic regression[9]. Mean imputation replaces the missing values with arithmetic mean based on the available data of the corresponding variables. The added mean values are equivalent to a group of uncorrelated data, which results in an underestimation of the overall correlations. On the other hand, regression imputation will strengthen the variable correlation. It predicts a regression equation from available values and replaces the missing values with predicted regression outcomes[6]. Mean imputation and regression imputation dramatically reduce the variability of data and result in bias. Stochastic regression imputation adds residual terms to regression imputation, which effectively restores the variability represented by the lost values[3].

Multiple imputation

Multiple imputation (MI) is the procedure that replaces the missing value with more than one imputation method. Each missing value is replaced by m (m > 1) plausible imputed values given the observed values. The imputation creates m complete data sets which are analyzed separately with the same procedure, resulting in m sets of parameter estimates and standard errors. The multiple sets of results are subsequently merged into one set of results. These are the three steps of multiple imputation: imputation phase, analysis phase, and pooling phase[10−13].

Imputation algorithms - Markov Chain Monte Carlo

Multiple imputation is regarded as a model-based imputation since the imputation model is constructed according to the distributional relationship between the missing values and the observed values[14]. Data augmentation is the most often used algorithm for normally distributed data which belongs to the family of Markov Chain Monte Carlo (MCMC) procedure. MCMC draws pseudo-random samples with the Monte Carlo method using Markov chains[10,11]. MCMC is a composite of two steps: The imputation step (I-step) and the Posterior step (P-step). Initial I-step is principally identical to stochastic imputation where model-based imputed values replace the missing values and the error terms are added to the estimated parameters to represent the uncertainty of observed data. In P-step, the I-step data perform as building blocks to generate new estimates of means and covariances under the Bayesian estimation principles. The new parameters randomly differ from those used in the previous I-step and are successively used to create imputed values in the preceding I-step. Repeating this two-step procedure for specific times yields multiple copies of complete data sets, each of which contains unique estimate parameters of the missing values[15,16]. Multiple imputation shares the advantage of a single imputation and restores the variability of imputed complete data.

SAS MI procedure

To apply the MI procedure, the type and distribution of the data need to satisfy certain assumptions. MI assumes that the data are from a continuous multivariate distribution and contain missing values that may occur on any of the variables. It also assumes that the missing data are under the MAR mechanism[17]. The next step is to analyze the existing missing data patterns, arbitrary or other specific patterns such as monotone. The PROC MI default MCMC method assumes that the multivariate normal data can be used. MCMC is an appropriate method for continuous arbitrary missing data[18].

Methods

Macro-based Forward Model Selection method

SAS LOGISTIC procedure forward selection provides an automated model selection for binary response. It computes and ranks the *p*-value of the Chi-square statistic for each effect not in the model. If the minimum *p*-value is smaller than α, this corresponding effect is added to the model and never removed. The above procedures are repeated until none of the remaining effects is significant at level α[17]. However, forward selection is not a good method to process missing data, since the observations with missing values are entirely excluded from analysis. The huge loss of information represented by incomplete observations impairs the accuracy of effect selection and power of significance test.

To keep the maximum information in incomplete observations, the Macro-based Forward Model Selection (MFMS) method was developed, in which the macro controls the automated effect selection using the forward selection algorithms. The biggest advantage of MFMS is to process one variable each time while keeping the incomplete observations in other variables undeleted. MFMS can be decomposed into steps: *p*-values of the Chi-square score for all one-variable models are calculated and ranked using the SAS LOGISTIC procedure. The smallest *p*-value (P-min) is considered as the candidate for the first cycle. If P-min is smaller than α, it indicates that the corresponding variable is significant at level α and eligible to enter the model. Only if the addition of the first effect happens, we consider all the two-variable combinations consisting of the first enrolled variable and one of the remained variables not in the model. Again, the smallest *p*-value is compared with α. If significant, the effect not in the previous model will enter as the second effect. Otherwise, the program ends up with the previous model. Macro controls the sequential entry of significant effects one by one until the termination point and returns the final effects list. The final model will contain the best significant effect added in each cycle.

Fan[19] reported the application of MFMS algorithms to analyze missing data in linear regression and concluded that MFMS is superior to traditional REG and MI procedures.

Simulation and model assumptions

Logistic regression is used to investigate the binary response, ordinal response, and nominal response. For binary response models, the response variable Y can take on one of two possible values, denoted for convenience by 1 and 2. Suppose vector X is the explanatory variable, and the response probability is:

$$\pi = \Pr(Y = 1|x)$$

The logic response model can be a linear form as:

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_i \mathrm{X}_i$$

A logistic regression model with 200 observations was simulated, each of which consists of a binary response Y and 20 independent variables X_1, X_2, …, X_{20}, as Table 1. In most cases, actual data usually have a small sample size and a complex set of variables. Additionally, the smaller the sample size, the larger the influence that the missingness could cause. So, the complete data was simulated with 200 samples and 20 X variables.

The model relies on several assumptions. Each X variable is independently drawn from a standard normal distribution as our complete explanatory variable matrix X. Y is the binary logistic function of X. Thus, our complete data set is a binary response with 20 multivariate normal-distributed explanatory variables. The assumption of our models, along with MCAR and MAR mechanisms, satisfy the requirements for the SA LOGISTIC and MI procedures.

The complete and missing data were simulated based on the following logistic function with descending coefficients:

$$logit(\pi) = \mathbf{ln}\left(\frac{\pi}{1-\pi}\right) = 1 - x_1 + 0.8x_2 + 0.6x_3$$

The descending order of coefficients was designed to examine how the coefficients can affect the model selection.

First, the 100 simulated complete data sets were processed by traditional logistic procedures with forward selection and each of the final effects list was recorded. By counting the effect selected times in the 100 procedures, the numbers of each X selected out of 100 times were acquired. Similarly, the 5%, 10%, 15%, and 20% randomly deleted datasets were processed by the proposed MFMS method, SAS traditional logistic procedure, and MI procedure respectively. Merging the results from the first and the second steps by SAS SQL procedure, an overall list of each X variable selected numbers were achieved for the complete data and at each missing percentage as shown in Tables 2–7.

Results

The development of the MFMS method aims to circumvent the shortcomings of LOGISTIC forward selection when processing missing data of logistic regression. MI procedure MCMC method is the

Table 1. The complete data matrix.

Observation	Y	X_1	X_2	…	X_{20}
1	Y_1	X_{11}	X_{12}	…	$X_{1,20}$
2	Y_2	X_{21}	X_{22}	…	$X_{2,20}$
3	Y_3	X_{31}	X_{32}	…	$X_{3,20}$
…	…	…	…	…	…
200	Y_{200}	$X_{200,1}$	$X_{200,2}$	…	$X_{200,20}$

Table 2. Model selection at 5% missing for 200 sample size, MCAR.

V	Complete data			SAS macro			SAS forward			MI		
	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE
x1	1.161	100	0.250	1.162	100	0.345	1.152	99	0.322	1.092	100	0.244
x2	−0.880	100	0.252	−0.938	99	0.334	−0.947	89	0.287	−0.837	100	0.266
x3	−0.679	96	0.186	−0.712	87	0.239	−0.750	61	0.233	−0.626	98	0.208
x4	−0.130	8	0.474	0.049	8	0.621	−0.013	15	0.452	−0.009	62	0.267
x5	−0.083	5	0.393	0.060	7	0.563	−0.015	15	0.445	0.007	58	0.240
x6	0.118	22	0.428	0.168	10	0.472	0.107	13	0.525	0.021	65	0.286
x7	0.229	10	0.337	−0.066	10	0.608	0.254	15	0.314	0.016	65	0.255
x8	0.446	9	0.057	0.012	15	0.543	−0.014	15	0.336	0.027	60	0.262
x9	0.187	8	0.406	0.110	16	0.478	−0.037	17	0.360	0.065	57	0.251
x10	0.095	14	0.403	−0.060	17	0.554	0.045	12	0.362	0.040	61	0.269
x11	−0.041	7	0.407	−0.059	9	0.567	0.079	7	0.348	0.007	57	0.232
x12	0.018	9	0.474	0.000	13	0.596	−0.091	8	0.376	0.009	58	0.243
x13	−0.085	13	0.443	0.101	20	0.482	−0.161	14	0.417	0.036	59	0.265
x14	0.071	6	0.578	0.212	11	0.544	0.010	17	0.334	−0.022	51	0.268
x15	−0.135	8	0.426	−0.022	12	0.553	−0.086	13	0.352	−0.017	63	0.253
x16	0.040	6	0.480	−0.090	11	0.455	−0.122	18	0.415	−0.078	59	0.240
x17	0.034	11	0.417	0.042	12	0.547	0.017	13	0.351	−0.034	57	0.248
x18	0.038	8	0.435	0.016	12	0.511	0.174	15	0.401	0.041	59	0.236
x19	−0.431	4	0.085	0.002	15	0.510	−0.228	10	0.311	0.005	55	0.239
x20	−0.029	7	0.482	−0.246	6	0.524	0.092	15	0.298	0.001	59	0.253

Table 3. Model selection at 10% missing for 200 sample size, MCAR.

V	Complete data			SAS macro			SAS forward			MI		
	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE
x1	1.161	100	0.250	1.197	100	0.576	1.045	54	0.312	0.985	100	0.221
x2	−0.880	100	0.252	−1.012	97	0.418	−0.867	35	0.346	−0.748	100	0.224
x3	−0.679	96	0.186	−0.867	79	0.365	−0.677	25	0.266	−0.571	98	0.193
x4	−0.130	8	0.474	0.132	16	0.884	0.020	13	0.343	−0.019	54	0.265
x5	−0.083	5	0.393	−0.233	13	0.746	0.011	14	0.246	−0.001	60	0.223
x6	0.118	22	0.428	−0.018	19	0.710	0.029	9	0.229	0.018	63	0.276
x7	0.229	10	0.337	−0.028	14	1.206	0.012	9	0.221	0.009	59	0.224
x8	0.446	9	0.057	0.335	9	0.746	0.077	14	0.210	0.049	66	0.228
x9	0.187	8	0.406	0.374	9	0.707	0.121	15	0.281	0.055	64	0.240
x10	0.095	14	0.403	0.14	17	0.725	0.091	15	0.224	0.025	62	0.256
x11	−0.041	7	0.407	−0.397	17	0.586	−0.083	9	0.281	−0.026	51	0.229
x12	0.018	9	0.474	−0.207	16	0.739	0.037	12	0.328	−0.010	52	0.238
x13	−0.085	13	0.443	0.018	17	0.727	0.028	15	0.297	0.015	55	0.261
x14	0.071	6	0.578	0.014	12	0.828	−0.012	15	0.241	0.004	60	0.238
x15	−0.135	8	0.426	−0.133	13	0.750	−0.060	14	0.308	0.007	60	0.240
x16	0.040	6	0.480	0.051	10	0.748	−0.079	14	0.224	−0.076	46	0.246
x17	0.034	11	0.417	−0.93	10	0.898	0.027	21	0.322	−0.028	64	0.229
x18	0.038	8	0.435	−0.102	10	1.034	0.124	16	0.298	0.018	58	0.254
x19	−0.431	4	0.085	−0.026	8	0.866	−0.024	15	0.258	−0.038	56	0.216
x20	−0.029	7	0.482	0.071	10	0.860	−0.021	14	0.184	0.009	59	0.234

most popular logistic modeling procedure. The MFMS, PROC LOGISTIC, and PROC MI methods were compared by testing the selection frequency of each X variable. The significant level of entry for all methods was 0.1. Detailed methodologies are described in the Methods section. Several rules for model selection accuracy were established: (1) the true model variables could be selected 50 times or more; (2) the true model variables could be enrolled into models more frequently than other variables; (3) the results of model selection and parameter estimates should be close to those of true models or complete data. (4) less false positive effects. The methods that satisfy those criteria are believed to be reliable for processing missing data. Tables 2–5 show the selecting frequency of X variables at different missing levels among all methods. Tables 6 & 7 present the selecting frequency of X variables under different missing mechanisms.

Columns 2–4 in each table show the results of complete data model selection by PROC LOGISTIC at significance level 0.1 where X1, X2, and X3 are in the true model and X4–X20 are not in the model. Tables 2–5 show that with the current model setting, almost all three model variables, except X3 for 96 times, were selected to enter into the model 100 times. The frequencies of other non-model effects were far behind. In addition, the coefficient estimates of X1–X3 matched the true model with minor differences. The above indicated the accuracy of the LOGISTIC procedure in model selection for complete data.

To demonstrate the model selection efficiency and accuracy, the performance of the three modeling methods were analyzed at each missing level under MCAR.

At a 5% missing level (Table 2), the model selection pattern and the coefficient estimates obtained from the proposed MFMS

Table 4. Model selection at 15% missing for 200 sample size, MCAR.

V	Complete data			SAS macro			SAS forward			MI		
	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE
x1	1.161	100	0.250	1.635	99	1.790	0.959	14	0.212	0.879	100	0.203
x2	−0.880	100	0.252	−1.608	97	3.742	−0.721	12	0.292	−0.685	100	0.201
x3	−0.679	96	0.186	−1.364	75	2.802	−0.656	4	0.223	−0.525	97	0.156
x4	−0.130	8	0.474	0.731	13	2.760	0.087	5	0.250	−0.015	59	0.235
x5	−0.083	5	0.393	0.847	9	1.421	0.023	4	0.129	0.002	52	0.239
x6	0.118	22	0.428	−0.065	15	1.023	−0.015	9	0.145	0.026	60	0.262
x7	0.229	10	0.337	1.715	17	6.451	−0.076	6	0.323	0.010	55	0.218
x8	0.446	9	0.057	−1.392	16	5.247	0.092	2	0.024	0.059	53	0.242
x9	0.187	8	0.406	0.252	12	1.769	−0.030	3	0.093	0.070	59	0.234
x10	0.095	14	0.403	−0.736	11	2.755	0.060	9	0.087	0.040	58	0.236
x11	−0.041	7	0.407	0.146	15	1.496	0.048	6	0.204	0.011	56	0.221
x12	0.018	9	0.474	0.127	10	0.969	0.094	4	0.214	0.027	59	0.218
x13	−0.085	13	0.443	0.493	15	1.020	0.009	7	0.144	0.033	56	0.259
x14	0.071	6	0.578	0.089	16	1.213	0.177	6	0.190	−0.021	58	0.223
x15	−0.135	8	0.426	−0.107	12	0.810	0.026	3	0.187	−0.017	60	0.228
x16	0.040	6	0.480	−0.318	18	0.769	−0.087	5	0.268	−0.064	54	0.203
x17	0.034	11	0.417	−0.126	15	1.107	−0.002	10	0.122	−0.047	65	0.219
x18	0.038	8	0.435	−0.342	13	1.420	0.021	5	0.285	0.000	58	0.231
x19	−0.431	4	0.085	1.089	14	4.293	−0.035	5	0.159	−0.032	56	0.223
x20	−0.029	7	0.482	2.323	19	9.146	−0.056	6	0.068	0.051	61	0.230

Table 5. Model selection at 20% missing for 200 sample size, MCAR.

V	Complete data			SAS macro			SAS forward			MI		
	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE
x1	1.161	100	0.250	3.586	99	8.270	0.916	10	0.147	0.795	100	0.191
x2	−0.880	100	0.252	−3.931	95	12.843	−0.882	6	0.204	−0.599	100	0.205
x3	−0.679	96	0.186	−3.412	65	9.724	.	.	.	−0.446	98	0.149
x4	−0.130	8	0.474	−3.784	8	12.557	0.084	3	0.101	0.040	53	0.207
x5	−0.083	5	0.393	4.551	14	12.84	0.013	6	0.145	0.041	58	0.226
x6	0.118	22	0.428	6.864	9	13.589	−0.065	6	0.115	0.025	56	0.236
x7	0.229	10	0.337	−1.193	12	5.979	−0.067	4	0.092	−0.020	57	0.229
x8	0.446	9	0.057	0.864	15	6.120	−0.075	2	0.370	0.053	59	0.224
x9	0.187	8	0.406	−0.884	13	20.753	.	.	.	0.063	54	0.225
x10	0.095	14	0.403	0.387	10	2.697	0.025	21	0.196	0.047	57	0.217
x11	−0.041	7	0.407	1.571	20	8.790	−0.014	1	.	−0.002	60	0.215
x12	0.018	9	0.474	−0.349	10	1.559	−0.069	8	0.111	0.000	60	0.188
x13	−0.085	13	0.443	0.512	16	2.378	0.013	6	0.240	0.054	61	0.221
x14	0.071	6	0.578	0.622	16	2.122	.	.	.	−0.011	55	0.236
x15	−0.135	8	0.426	2.194	18	15.942	0.040	4	0.234	−0.033	58	0.236
x16	0.040	6	0.480	−2.824	18	11.498	.	.	.	−0.028	60	0.195
x17	0.034	11	0.417	−2.208	13	7.217	−0.010	8	0.175	−0.031	53	0.205
x18	0.038	8	0.435	−0.337	9	1.457	0.115	3	0.113	−0.007	51	0.196
x19	−0.431	4	0.085	−1.937	11	11.302	.	.	.	−0.040	56	0.199
x20	−0.029	7	0.482	−1.990	13	5.940	−0.014	22	0.160	−0.026	55	0.212

method were highly consistent with the complete data. X1, X2, and X3 were found for 100, 99, and 87, respectively. This suggests the validity of the MFMS method at a 5% missing level. SAS PROC LOGISTIC did not perform as well as MFMS, which had a lower selection frequency and a larger difference compared to MFMS. However, it was still acceptable in the current setting. The MI also identified X1 and X2 100 times, while X3 was found 98 times. The coefficient estimates were also close to those of the complete data. However, after carefully examining the MI results, it was found that all the non-model variables were included in the model more than 50 times. This indicated that the MI procedure not only selected the true model variables but also incorrectly included the non-model variables at a high frequency. This increased the likelihood of achieving a biased conclusion based on the insignificant model. The issues inherent in the forward selection and the MI procedure worsened with increasing missingness.

To evaluate the performance of the three model selection methods under MAR data scenarios, datasets with a 5% missing level for MAR were established. The results, shown in Table 6, demonstrate consistent conclusions with those observed under MCAR. Since NMAR does not meet the assumptions required by the LOGISTIC and MI procedures, only the performance of the MFMS method under the NMAR mechanism were evaluated. The results, presented in Table 7, indicate that the MFMS method also performs well under NMAR conditions.

At a 10% missing level shown in Table 3, MFMS enrolled X1–X3 100, 97, and 79 times, respectively. The parameter estimates were close to the complete data with slight changes. This approved that MFMS was reliable at the 10% missing level. SAS LOGISTIC could not maintain its accuracy. It found X1 54 times, X2 35 times, and X3 25 times. The selection frequency dramatically decreased for true model variables. Only X1 entered the model more than 50 times,

Table 6. Model selection at 5% missing for 200 sample size, MAR.

V	Complete data			SAS macro			SAS forward			MI		
	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE	Mean	N	MSE
x1	1.122	100	0.226	1.090	100	0.307	1.171	92	0.349	1.051	100	0.234
x2	−0.884	98	0.223	−0.892	100	0.248	−0.967	82	0.322	−0.822	100	0.244
x3	−0.644	95	0.193	−0.733	91	0.226	−0.726	67	0.265	−0.583	100	0.210
x4	−0.029	13	0.448	0.147	7	0.482	0.186	11	0.468	0.024	65	0.271
x5	0.007	15	0.460	−0.057	13	0.468	−0.079	22	0.480	−0.001	60	0.297
x6	−0.132	12	0.386	0.156	11	0.487	−0.262	12	0.418	−0.028	60	0.268
x7	−0.209	11	0.383	0.110	12	0.551	0.015	13	0.481	−0.024	62	0.264
x8	0.114	15	0.415	0.034	14	0.561	0.095	17	0.488	0.026	59	0.295
x9	−0.001	12	0.415	0.072	8	0.495	−0.123	16	0.418	0.025	63	0.256
x10	−0.068	7	0.468	−0.091	11	0.475	−0.207	15	0.277	−0.049	54	0.259
x11	−0.186	6	0.403	−0.147	10	0.504	−0.054	8	0.422	−0.034	61	0.233
x12	−0.099	13	0.420	−0.056	16	0.555	−0.057	17	0.531	−0.035	64	0.254
x13	0.143	16	0.360	−0.198	7	0.464	0.002	16	0.369	0.012	65	0.249
x14	0.147	14	0.400	0.157	7	0.582	−0.081	15	0.534	0.048	70	0.259
x15	−0.080	12	0.465	0.187	14	0.483	−0.071	12	0.529	−0.006	58	0.274
x16	−0.177	11	0.435	−0.363	4	0.554	0.129	10	0.349	−0.050	51	0.254
x17	0.268	10	0.358	−0.211	17	0.466	0.014	15	0.416	0.053	51	0.256
x18	0.033	10	0.428	0.172	12	0.456	0.231	8	0.240	0.063	66	0.246
x19	0.242	9	0.386	0.160	10	0.514	0.049	11	0.472	0.018	62	0.265
x20	0.039	8	0.472	0.084	10	0.470	−0.132	10	0.333	0.031	56	0.246

Table 7. Model selection at 5% missing for 200 sample size, NMAR.

V	Complete data			SAS macro		
	Mean	N	MSE	Mean	N	MSE
x1	1.097	100	0.258	1.145	100	0.316
x2	−0.859	100	0.206	−0.949	100	0.271
x3	−0.656	95	0.205	−0.718	87	0.268
x4	0.009	9	0.486	0.098	19	0.535
x5	0.027	10	0.451	0.003	13	0.656
x6	0.106	9	0.463	0.185	12	0.432
x7	0.134	16	0.442	0.004	10	0.559
x8	0.001	16	0.409	0.012	5	0.481
x9	−0.171	8	0.463	−0.013	14	0.521
x10	0.057	11	0.419	0.200	15	0.410
x11	−0.147	11	0.448	0.000	15	0.562
x12	−0.154	10	0.419	0.049	10	0.583
x13	0.248	10	0.342	−0.042	18	0.558
x14	−0.032	12	0.409	0.197	16	0.559
x15	0.053	9	0.409	0.009	13	0.607
x16	−0.091	12	0.378	0.048	12	0.641
x17	0.193	12	0.402	0.140	13	0.552
x18	0.197	8	0.410	0.043	10	0.561
x19	0.416	7	0.371	−0.125	14	0.660
x20	−0.001	11	0.486	0.085	10	0.668

while X2 and X3 could not stand out from other non-model variables. It indicated that PROC LOGISTIC was no longer able to build up a correct logit model. This was because missingness at this level caused a huge loss of usable data. Although it was only 10% missing in each X variable. When considering 20 X variables, the listwise deletion algorithms resulted in a much higher percentage of overall missing. The complete subset for SAS LOGISTIC was much smaller than the 200 sample size. As seen with a 10% missing level, MI enrolled too many non-significant variables and was not a reliable method.

The results are shown in Table 4 for 15% missingness. MFMS sturdily selected the true model variables with a similar frequency of 10%. However, the estimates of coefficients moderately deviated from the true model. SAS LOGISTIC failed to select any significant variable. All three true model variables were found under 50 times. The coefficient estimates by MI started to deviate from the true model while the model selection was almost the same as previous missing levels.

As seen in Table 5, the 20% missing level where MFMS still found the true model was examined, but the parameter estimates largely deviated from the complete data. SAS LOGISTIC completely lost its accuracy of model selection. No significant variable was found. Parts of the X variables, including X3, X9, X14, X16, and X19, were not found even once. No parameter was estimated for those variables. Those demonstrated that SAS LOGISTIC was no longer able to analyze the data at a high missing level. The frequency of variable selection by MI seldom changed with increasing of missing level. The parameter estimates deviated further from the true model.

In summary, the MFMS satisfied the criteria to be a reliable logistic modeling method at 5% and 10% missing levels. MFMS exhibited a high consistency of model selection pattern compared to the complete data set. The three true model variables were successively selected at all missing levels. Additionally, the happening of false positive selection was efficiently forbidden, since MFMS only added the best significant effect with the smallest *p*-value each cycle. MFMS also extensively reduced the exclusion of incomplete observations. Only the missing values in the current analyzed variable were excluded and the other variable was untouched. It maintained the maximum information and uncertainty of the missing data. As proof, the parameter estimates at a 5% missing level were very close to those of complete sets. With the increase of missingness to 15%, the parameter estimates slightly deviated from complete data. However, compared to the SAS LOGISTIC and MI procedures, MFMS was still the best one. At a 20% missing level, MFMS was able to select the right model, but with incorrect parameter estimates due to the low events per variable.

SAS LOGISTIC excluded any observation that contained missing values. When missingness happens in multiple variables, the listwise deletion causes severe loss of information. The selecting frequency of true model variables dramatically decreased at 10% and higher missing levels. Finally, it completely failed to select the accurate model and estimate the parameters at the 15% and 20% missing levels.

The MI procedure could choose the true model variables at all missing levels 100 times. However, the selection frequency of non-significant variables was also around 50 times. MI automatically sampled the missing data with model-based imputation, which led to an overestimation of the logistic association and a narrower confidence interval. Those resulted in insignificant model selection and biased parameter estimates.

In summary, it was concluded that MFMS was the best reliable selection method for binary logistic models with independently normal-distributed multivariate for low-moderate level missing under MCAR, MAR, or NMAR assumption.

Real example

The World Health Organization estimates 12 million annual deaths from heart diseases globally, with half of all deaths in the United States and other developed countries attributed to cardiovascular diseases. Early detection enables high-risk individuals to adopt preventive lifestyle changes. A publicly available cardiovascular dataset from a study in Framingham, Massachusetts (USA), with over 4,000 records and 15 attributes, was analyzed to identify key risk factors for heart disease. The dataset is publicly available on the Kaggle website. Given its missing values, this dataset was used to compare SAS forward selection with the Macro-based selection method. Results are shown in Tables 8 & 9.

The results indicate that both methods identify key variables — age, systolic blood pressure, cigarettes smoked per day, gender, glucose level, and total cholesterol level — as significant predictors of cardiovascular disease risk. However, the macro-based selection method also includes an additional variable, the history of stroke, with a significant *p*-value (0.016), suggesting it may have identified an additional important predictor overlooked by the forward selection method.

Table 8. SAS forward selection results.

Parameter	Estimate	Standard error	Pr > ChiSq
Intercept	9.130	0.476	< 0.0001
Male	−0.561	0.107	< 0.0001
Age	−0.066	0.006	< 0.0001
cigsPerDay	−0.019	0.004	< 0.0001
totChol	−0.002	0.001	0.043
sysBP	−0.018	0.002	< 0.0001
Glucose	−0.007	0.002	< 0.0001

Male: gender, male or female (Nominal); Age: Age of the patient (Continuous); cigsPerDay: the number of cigarettes that the person smoked on average in one day (continuous); totChol: total cholesterol level (Continuous); sysBP: systolic blood pressure (Continuous); Glucose: glucose level (Continuous).

Table 9. SAS macro-based selection results.

Parameter	Estimate	Standard error	Pr > ChiSq
Intercept	9.125	0.463	< 0.0001
Age	−0.066	0.006	< 0.0001
sysBP	−0.017	0.002	< 0.0001
cigsPerDay	−0.020	0.004	< 0.0001
Male	−0.549	0.104	< 0.0001
Glucose	−0.008	0.002	< 0.0001
PrevalentStroke	**−1.064**	**0.442**	**0.016**
totChol	−0.003	0.001	0.018

Age: Age of the patient (Continuous); sysBP: systolic blood pressure (Continuous); cigsPerDay: the number of cigarettes that the person smoked on average in one day (continuous); Male: gender, male or female (Nominal); Glucose: glucose level (Continuous); PrevalentStroke: whether or not the patient had previously had a stroke (Nominal); totChol: total cholesterol level (Continuous).

Discussion

Logit function parameters are estimated using the method of maximum likelihood[17]. While sample size does not directly influence parameter estimates in logistic regression, the number of events is critical. However, unless the two levels of the binary response are highly imbalanced, such as 10% vs 90%, a reduction in sample size indirectly impacts parameter estimates by decreasing the number of events. Research indicates that biases in parameter estimates decrease as sample size increases in binary logistic models[20]. Insufficient events can lead to infinite maximum likelihood estimates, resulting in parameter estimation failure. The minimum requirement for accurate estimation of logistic parameters is 10 events per explanatory variable (EPV)[21].

The present simulation included 20 explanatory variables (X) and 200 samples. Under the assumptions of the model, the probability of each binary response was not extremely imbalanced. However, the EPV in the complete dataset was lower than 10, which explains the minor bias observed in parameter estimates. At low levels of missingness, the MFMS method was reliable for selecting significant variables and estimating parameters since the EPV did not drop substantially. As missingness increased to 20%, the EPV decreased dramatically. Although MFMS could still identify significant variables, parameter estimates were biased due to the lower EPV, which resulted in increased variance and wider confidence intervals.

When the sample size increased to 500, the accuracy of parameter estimates improved significantly compared to the estimates derived from 200 samples, as shown in Appendix I. However, given funding and time constraints, researchers often work with smaller sample sizes of around 200. To address parameter estimate bias at high levels of missingness, MFMS can first be applied to identify significant variables, such as X1–X3. A logistic model can then be fitted with only these significant variables, substantially increasing the EPV. Parameters can subsequently be re-estimated using PROC LOGISTIC or PROC MI. This approach was tested in the present model, and the re-estimated parameters, presented in Appendix II, were much closer to the true model values.

In the simulated logistic model, it was assumed that the X variables were independently normally distributed and considered only the main effects in the true model. Interactions between X variables were not accounted for. Independence among variables is rare in practical scenarios. Interactions can provide valuable information about relationships and associations between variables. Future research will explore the effects of interactions on model performance.

Another limitation of the present approach is its focus on continuous explanatory variables. Extending the procedure to accommodate mixed categorical and continuous variables would make it more applicable to a broader range of problems.

Additionally, the simulated data used in this study have the same sample size and the same percentage of missingness across all variables. In real-world data, the percentage of missingness and the sample size of the complete subset often vary between variables, leading to differences in the degrees of freedom (DOF) for each variable. MFMS probably could not find the criteria for statistics. Despite this, MFMS demonstrated robustness by successfully selecting significant variables one-by-one using the Chi-square-associated *p*-value. In this study, the impact of DOF differences was negligible.

The model currently focuses mainly on MCAR mechanisms. However, MAR scenarios are more prevalent in practice. Under MAR, the MI procedure often performs better. Future research could explore using MI to preprocess MAR data by identifying a subset of variables that fit the model. This subset could then undergo a

LOGISTIC backward selection step to re-evaluate and remove false positives[6,7,10]. However, MI is not always the optimal method for MAR or other complex data scenarios. Since MI involves random draws, the parameter estimates and test statistics can vary between runs on the same dataset, potentially leading to different conclusions from the same data and procedures[22].

The MACRO code is provided in Appendix III. This MACRO can also handle categorical predictors and interaction terms. Additionally, the methodology in the MACRO can be adapted for model selection in other SAS procedures, such as GENMOD and MIXED, broadening its utility.

Author contributions

The authors confirm contribution to the paper as follows: conceptualization: Ma C; methodology: Han M, Ma C; writing — original draft preparation: Han M; revision, writing — review & editing: Zhang X, Ma C. All authors have read and agreed to the published version of the manuscript.

Data availability

The data presented in this study are openly available in the Kaggle website www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper at (https://www.maxapress.com/article/doi/10.48130/stati-0024-0002)

Dates

Received 27 September 2024; Revised 29 November 2024; Accepted 17 December 2024; Published online 26 December 2024

References

1. Hosmer D, Lemeshow S. 2000. *Applied logistic regression*. 2nd Edition. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0471722146
2. Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581−92
3. Little RJA, Rubin DB. 2002. *Statistical analysis with missing data*. 2nd Edition. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781119013563
4. Ibrahim JG, Zhu H, Tang N. 2008. Model selection criteria for missing data problems using the EM algorithm. *Journal of the American Statistical Association* 103(484):1648−58
5. Peugh JL, Enders CK. 2004. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research* 74:525−56
6. Baraldi AN, Enders CK. 2010. An introduction to modern missing data analyses. *Journal of School Psychology* 48:5−37
7. SAS Institute Inc. 1999. *SAS/STAT User's Guide*. Version 8. Cary, NC: SAS Institute Inc. www.sfu.ca/sasdoc/sashtml/hrddoc/indfiles/57388.htm
8. Allison PD. 2001. *Missing data. Series: Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: SAGE Publications.
9. Pigott TD. 2001. A review of methods for missing data. *Educational Research and Evaluation* 7(4):353−83
10. Rubin DB. 1987. *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470316696
11. Allison PD. 2002. *Missing data*. Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781412985079
12. Enders CK. 2010. *Applied missing data analysis*. New York: Guilford Press
13. Schafer JL, Olsen MK. 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* 33:545−71
14. Liu M, Taylor JMG, Belin TR. 2000. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* 56(4):1157−63
15. White IR, Royston P, Wood AM. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 30(4):377−99
16. Young R, Johnson DR. 2011. Imputing the missing Y's: implications for survey producers and survey users. *Proceedings of the AAPOR conference abstracts* pp. 6242−48
17. SAS Institute Inc. 2010. What's New in SAS/STAT 9.22. In *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/cdl/en/statugwhatsnew/63792/PDF/default/statugwhatsnew.pdf
18. Berglund PA. 2010. *An introduction to multiple imputation of complex sample data using SAS® v9.2*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings10/265-2010.pdf
19. Fan Y. 2014. *Comparison of model selection methods in multiple linear regression model with missing value*. Master's Thesis. University of Buffalo, USA.
20. Bergtold JS, Yeager EA, Featherstone AM. 2011. Sample size and robustness of inferences from logistic regression in the presence of nonlinearity and multicollinearity. *2011 Annual Meeting, 24−26 July 2011, Pittsburgh, Pennsylvania*. Pennsylvania: Agricultural and Applied Economics Association. pp. 227−41. doi: 10.22004/AG.ECON.103771
21. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49(12):1373−79
22. Allison PD. 2012. Handling missing data by maximum likelihood. *SAS Global Forum 2012. Stat Horizons*. Haverford, PA: Statistical Horizons. https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf