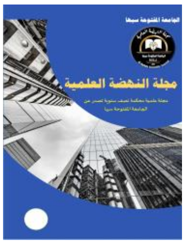
مجلة النهضة العلمية

Al-Nahda Scientific Journal

Journal homepage :https://ous.edu.ly/Journal/index.php/OUSJ



Assessing the Performance of EM Algorithm and Multiple Imputation in Beta Regression with Missing Data

Aissa O Asserhai[1] & Alsaidi M. Altaher[2]

[1, 2] Department of Statistics, Faculty of Sciences, Sebha University, Libya

Email: [1] ais.assrhani@sebhau.edu.ly , [2] als.altaher@sebhau.edu.ly

Abstract

Handling missing data remains a fundamental challenge in statistical modeling, particularly within regression models. This study evaluates and contrasts two widely used imputation techniques, the Expectation-Maximization (EM) algorithm and Multiple Imputation (MI), in the context of beta regression. The EM algorithm iteratively estimates missing values by maximizing the likelihood function, while MI generates multiple plausible datasets to account for the uncertainty of missing data. Using the gasoline yield dataset with artificially induced missingness at 5%, 10%, and 15%, we assessed the performance of both methods across various link functions and likelihood estimators. Findings suggest that while both methods are effective at lower missingness levels, EM consistently yields more robust parameter estimates at moderate levels of missingness (around 10%), and maintains strong performance as it increases, especially when coupled with the log-log link function. These findings may offer valuable insights for researchers and practitioners dealing with incomplete data in beta regression models.

Keywords: Missing Data, EM algorithm, Multiple Imputation (MI), Beta Regression, Link function.

تقييم أداء خوارزمية تعظيم التوقع و طريقة التعويض المتعدد في إنحدار بيتا

عيسى السرحي [1] والسعيدي محمد الطاهر [2]

[1،2]قسم الإحصاء، كلية العلوم، جامعة سبها، ليبيا

المراسلة: [1] ais.assrhani@sebhau.edu.ly , [2] als.altaher@sebhau.edu.ly

الملخص

معالجة البيانات المفقودة، تظل تحديًا أساسيًا في النمذجة الإحصائية، لا سيما في نماذج الانحدار . تُقيّم هذه الدراسة وتُقارن بين تقنيتين شائعتي الاستخدام في المعالجة، وهما خوارزمية تعظيم التوقع (EM) والتعويض المتعدد (MI)، وذلك في سياق انحدار بيتا. تُقدّر خوارزمية تعظيم التوقع القيم المفقودة تكراريًا من خلال تعظيم دالة الاحتمال، بينما تُولّد خوارزمية التعويض المتعدد مجموعة بيانات متعددة معقولة لمراعاة عدم اليقين في البيانات المفقودة. باستخدام مجموعة بيانات إنتاج البنزين مع فقد مُستحثّ اصطناعيًا عند نسب 5% و10% و15%، قمنا بتقييم أداء كلتا الطريقتين عبر دوال ربط ومُقدّرات امكان مُختلفة. تُشير النتائج إلى أنه على الرغم من فعالية كلتا الطريقتين عند مستويات فقد منخفضة، إلا أن خوارزمية تعظيم التوقع تُنتج باستمرار تقديرات للمعالم أكثر دقة عند مستويات فقد معتدلة (حوالي 10%)، مع الحفاظ على

Received 21 April 2025 - Accepted 30 June 2025

أداء قوي مع ازدياد مستوى الفقد، خاصةً عند اقترانها بدالة ربط log–log. قد تُقدّم هذه النتائج رؤى قيّمة للباحثين والممارسين الذين يتعاملون مع البيانات غير المكتملة في نماذج انحدار بيتا.

الكلمات المفتاحية: البيانات المفقودة، خوارزمية تعظيم التوقع (EM)، التعويض المتعدد (MI)، انحدار بيتا، دالة الربط.

1 Introduction

Beta regression is tailored for modeling dependent variables bounded within the (0, 1) interval, such as rates, ratios, and proportions. Its flexibility makes it suitable for a variety of applications across disciplines like: medicine, environmental research, finance, social sciences, and natural sciences [1]. Despite its advantages, one of the common and critical challenges faced when applying regression models, including beta regression, is handling missing data [2]. If left unaddressed or treated inappropriately, they can introduce bias and weaken statistical inference. Various strategies exist for handling missing data. Traditional approaches, such as listwise deletion (excluding all cases with any missing values), is considered among the least effective methods in practical applications [3], single imputation approaches, such as mean or median, substitution. Although these methods are simple and easy to implement, they suffer from substantial drawbacks. For instance, excluding incomplete cases can lead to a loss of valuable information and potential bias if data are not missing completely at random. Similarly, imputing missing values with single estimate tends to underestimate variability, distort the data distribution, and ignore the inherent uncertainty associated with missingness [4]. To overcome these limitations, more advanced techniques have been developed, notably the Multiple Imputation (MI) and the Expectation-Maximization (EM) algorithm, which provide more principled ways to account for uncertainty in missing data [5][6]. While both EM and MI are widely used in regression contexts [7], there is limited research evaluating their performance within the beta regression framework.

Therefore, this paper aims to compare EM and MI methods within the context of beta regression models under varying levels of missingness. Specifically, the study evaluates the effectiveness of both techniques using different link functions and maximum likelihood estimators to provide practical insights into the optimal handling of missing data in beta regression models. Section 2 outlines theoretical foundations, Section 3 describes imputation methods, Section 4 presents empirical results, and Section 5 concludes the study.

2 Methodology

2.1 Beta Regression

Beta regression model was proposed by Ferrari and Cribari [8], particularly useful for modeling variables constrained within the (0, 1), interval, such as proportions, rates, and percentages. In this framework the response variable y is assumed to follow a beta distribution, commonly reparametrized in terms of the mean (μ) and a precision parameter (ϕ). The probability density can be expressed as

$$f\left(y,\mu,\phi\right)=\begin{cases}\dfrac{\Gamma\left(\phi\right)}{\Gamma\left(\mu\phi\right)\Gamma\left((1-\mu)\phi\right)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} & y\in\left(0,1\right)\\ & \phi>0\\ 0 & o.w\end{cases} \tag{1}$$

where $\Gamma(.)$ denotes the gamma function, μ, $\mu\in(0,1)$, represents the mean of the distribution. The variance of y is given by $Var(Y)=\frac{\mu(1-\mu)}{\phi+1}$ [9]. To related the mean response to a set of linear predictors, a link function $g(.)$ employed, yielding the model formulation

$$g(\mu_i)=\eta_i=X^T\beta \tag{2}$$

where X^T denotes the vector of explanatory variables and β is a vector of regression coefficients, η_i is the linear predictor, $g(.)$ is a strictly monotonic and twice differentiable [10] Common choices link function include the logit, the probit function, and the log-log link

The log-likelihood function for the model is

$$\begin{aligned} l(\beta)=\sum_{i=1}^{n}[\ln L(\mu_i,\phi)]=\sum_{i=1}^{n}[\ln\Gamma(\phi)-\ln\Gamma(\mu_i,\phi)-\ln\Gamma((1-\mu_i),\phi) \\ +((1-\mu_i),\phi)\ln y_i+((1-\mu_i)\phi-1)\ln(1-y_i)] \end{aligned} \tag{3}$$

Filling the maximum likelihood estimator (MLE) for β vector at fixed value for ϕ we take the partial derivative with respect to β and equaling (4) to the zero and solve by a iterative procedure. For details see [11]

$$\begin{aligned} \frac{\partial l(\beta)}{\partial\beta_j} &= \sum_{i=1}^{n}\left(\frac{dl(\beta)}{dl(\theta_i)}\frac{dl(\theta_i)}{dl(\mu_i)}\frac{dl(\mu_i)}{dl(\eta_i)}\frac{\partial l(\eta_i)}{\partial l(\beta_i)}\right) \\ &= \sum_{i=1}^{n}\left(\frac{y_i-\mu_i d\mu_i}{\varphi\mu_i(1-\mu_i)}\cdot\frac{d(\mu_i)}{d(\eta_i)}x_{ij}\right) \end{aligned} \tag{4}$$

2.2 Missing Data Mechanism

Understanding the underlying mechanism of missing data is critical when choosing appropriate imputation methods. According to Enders [12], three primary mechanisms are defined:

Missing Not at Random (MNAR):

The probability of missing data on a variable is related to the value of the variable itself, even after controlling for other variables. This means the missingness is directly related to the missing values of the variables Y_{mis} themselves

$$p(R|Y_{obs},Y_{mis})=p(R|Y_{mis}) \tag{5}$$

Where is the indicator of missing data, Y_{obs} and Y_{mis} are the observed and missing parts of the data

Missing at Random (MAR):

Here, the probability of missing data on a variable is related to some of the observed data, but not the missing data itself. This means that it depends only on the observed values of the variable Y_{obs} in the dataset:

$$p\left(R \middle| Y_{obs}, Y_{mis}\right) = p\left(R \middle| Y_{obs}\right) \tag{6}$$

Missing Completely at Random (MCAR):

Missingness occurs entirely at random and is unrelated to either observed or unobserved data. This means that R is purely random and does not depend on the values of any variables in the dataset:

$$p\left(R \middle| Y_{obs}, Y_{mis}\right) = p\left(R\right) \tag{7}$$

These mechanisms help in understanding the nature of the missing data and guide the choice of methods for imputation or analysis. For instance, methods like Multiple Imputation and the EM algorithm are often used under the assumption of MAR [5]. In this study were simulated under the MCAR which is often considered the least problematic for unbiased analysis when handled appropriately.

3 Missing Data Imputation Methods

Single imputation methods, such as mean imputation, median imputation, and mode imputation, are commonly used to handle missing data. However, these methods have several drawbacks and limitations: Bias Introduction: single imputation methods can introduce bias into the dataset [13]. For example, mean imputation can reduce the variability of the data, leading to underestimated standard errors and inflated test statistics. [14]. Ignoring Data Relationships: These methods do not account for the relationships between variables. For instance, mean imputation replaces missing values with the mean of the observed values, ignoring any potential correlation between the missing variable and other variables in the dataset. Distortion of Data Distribution: Imputing missing values with the mean can distort the original distribution of the data [15]. This can be particularly problematic for skewed distributions or when the proportion of missing data is high [16]. Underestimation of Variability: Single imputation methods tend to underestimate the variability in the data [4]. This can lead to overly optimistic confidence intervals and p-values, which can affect the validity of statistical inferences [17]

Due to the above drawbacks mentioned for the classical imputation methods, we give attention to the two widely used methods for dealing with missing data is the Expectation-Maximization (EM) algorithm and Multiple Imputation (MI)

3.1 The Expectation-Maximization (EM) algorithm

The EM algorithm, proposed by Dempster et al [18], is an iterative method for maximum likelihood estimation in the presence of incomplete data. It operates through two main steps: the Expectation (E) step and the Maximization (M) step.

EM Algorithm Steps

i. Start with initial guesses for the parameters.
ii. E-step: calculate the expected value of the log-likelihood function, with respect to the conditional distribution of the missing data given the observed data and the current parameter estimates

iii. M-step (Maximization): Maximize the expected log-likelihood found in the Estep to update the parameter estimates:

iv. Iteration: Repeat the E-step and M-step until convergence, i.e., until the parameter estimates change by less than a pre-specified threshold.

3. 2 MI imputation

Multiple Imputation is a statistical technique proposed by Rubin (1987)[19], and used to handle missing data by creating multiple complete datasets. Each dataset is analyzed separately, and the results are combined to produce estimates and inferences that account for the uncertainty due to missing data

Steps in the MI Algorithm

Step 1: Replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. This is done multiple times to create several complete datasets.

Step 2: Generate (m) complete datasets by repeating the imputation process (m) times. Each dataset will have different imputed values

Step 3: Perform the desired statistical analysis on each of the (m) complete datasets separately.

Step 4: Combine the results from the (m) analyses to produce a single set of estimates. This involves averaging the estimates and adjusting the standard errors to reflect the variability between the imputed datasets.

3.3 Model Evaluation criteria

To assess model performance and select the optimal imputation methods, the following statistical criteria were utilized.

Akaike Information Criterion (AIC)

Balances model fit and complexity, with lower values indicating a better trade-off. It is defined as:

$$AIC = -2\ln\left(l\right) + 2k \tag{8}$$

where: (l) is the likelihood of the model and (k) is the number of parameters in the model.

Bayesian Information Criterion (BIC)

Similar to AIC but imposes stronger penalty for model complexity, favoring more parsimonious models with lower values. It is defined as:

$$BIC = -2\ln\left(l\right) + k\ln\left(n\right) \tag{9}$$

where (n) is the sample size.

Pseudo R-squared

Measures the proportion of variability that is explained by the model, with higher values indicating a better fit. Pseudo R-squared is used in the context of models where traditional R-squared is not applicable, such as beta regression. One common form is McFadden’s R-squared:

$$\text{Pseudo R-squared} = 1 - \left(\frac{\log\left(L_{null}\right)}{\log\left(L_{fit}\right)} \right) \tag{10}$$

(L_{fit}) is the likelihood of the fitted model. And L_{null} is the likelihood of the null model (a model with only an intercept).

Log-Likelihood

The log-likelihood (loglink) is a measure of how well the model explains the observed data, where higher values indicate better explanatory power.

4 Results and Discussion

4.1 Experimental Design

To empirically compare the EM and MI imputation techniques, we employed the gasoline yield dataset originally presented by Prater (1956) [20], This dataset contains 32 observations and several explanatory variables related to the efficiency of crude oil conversion into gasoline, a typical context where beta regression is appropriate due to the bounded nature of the response variable. The basic descriptive statistics are shown in Table 1

Table 1: The basic descriptive statistics for Gasoline Yield dataset

Variable	Description	Mean	Median	Sd	Skewness	Koutosis	min	max
Yield (Y)	Proportion of crude oil converted to gasoline	0.1966	0.1780	0.1072	0.3870	2.3440	0.0280	0.457
Temp	Temperature (in Fahrenheit) at which 100% gasoline has vaporized	332.10	349.00	37.541	0.5170	2.2970	205.00	444.0
Gravity	API gravity of the crude oil	39.250	40.000	5.6350	0.5890	3.0670	31.800	50.80
Pressure	Vapor pressure of the crude oil	4.1810	4.8000	2.6200	0.1150	1.9530	0.2000	8.600
ASTM	ASTM distillation temperature	332.10	349.00	69.760	-0.2790	1.9240	205.00	444.0

Artificial missing values were introduced completely at random (MCAR) into the explanatory variables at rates of 5%, 10%, and 15%. Both EM and MI methods were applied under Five different link functions: the logit link function $\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$; the probit link function $\eta_i = \Phi^{-1}\left(\mu_i\right)$, where $\Phi^{-1}(.)$ is the cumulative normal distribution function; the clog-log link function $\eta_i = \log\left(-\log\left(1-\mu_i\right)\right)$; the Log $\eta_i = \log\left(\mu_i\right)$; and the complementary log-log link function $\eta_i = -\log\left(-\log\left(\mu_i\right)\right)$ [8], the Cauchit link function $\eta_i = \tan\left(\pi\left(\mu_i - 0.5\right)\right)$. Furthermore, three variants of maximum likelihood estimators were used: standard ML with no correction (ML), ML with bias correction (BC), and ML with bias reduction (BR). The **betareg package** in R programming is utilized to apply beta regression. The models were assessed using AIC, BIC, Pseudo R², and Log-Likelihood as selection criteria.

4.2 Main Findings

Case1 : 5% Missing Data:

Both EM and MI demonstrated comparable performance, with EM slightly outperforming MI across most link functions. The differences in AIC and BIC values were minimal at this low missingness level, indicating that either method could be suitably applied under limited data loss conditions..

Case 2: 10% and 15% Missing Data:

For all link functions, EM consistently outperformed MI in terms of AIC, BIC, Pseudo R², and Log-Likelihood. EM's superiority became more pronounced as the missing data percentage increased. The gains tend to stabilize beyond the moderate missingness threshold. Across all scenarios the log-log link function resulted in the best model fit, whereas the cauchit link exhibited the weakest performance.

Table 2: criteria for comparison at 5% missing of data

Link function	Methods	ESTIMATOR	BIC	AIC	Pse R^2	Log-lik
Logit	EM	ML	-133.95	-125.15	0.9296	72.97
		BC	-133.04	-124.25	0.9296	72.52
		BR	-133.04	-124.24	0.9296	72.52
	MI	ML	-131.24	-122.44	0.9285	71.621
		BC	-130.34	-121.54	0.9284	71.17
		BR	-130.33	-121.54	0.9285	71.17
Probit	EM	ML	-139.69	-130.90	0.9454	75.85
		BC	-138.79	-129.99	0.9454	75.39
		BR	-138.79	-129.99	0.9454	75.39
	MI	ML	-136.50	-127.71	0.9434	74.25
		BC	-135.59	-126.80	0.9434	73.80
		BR	-135.59	-126.80	0.9434	73.80
Clog-log	EM	ML	-128.39	-119.60	0.9174	70.20
		BC	-127.48	-118.69	0.9174	69.74
		BR	-127.48	-118.69	0.9174	69.74
	MI	ML	-126.53	-117.74	0.9168	69.27
		BC	-125.62	-116.83	0.9168	68.81
		BR	-125.62	-116.83	0.9168	68.81
Cauchit	EM	ML	-102.963	-94.168	0.6648	57.48
		BC	-102.05	-93.252	0.6647	57.02
		BR	-102.04	-93.249	0.6647	57.02
	MI	ML	-103.14	-94.3443	0.6527	57.57
		BC	-102.22	-93.428	0.6524	57.11
		BR	-102.22	-93.425	0.6525	57.11
Log	EM	ML	-122.03	-113.23	0.9025	67.02
		BC	-121.12	-112.33	0.9025	66.56
		BR	-121.12	-112.32	0.9025	66.56
	MI	ML	-121.11	-112.32	0.9020	66.56
		BC	-120.20	-111.41	0.9020	66.10
		BR	-120.20	-111.41	0.9020	66.10
Log-log	EM	ML	-146.75	-137.95	0.9585	79.37
		BC	-145.84	-137.05	0.9585	78.920
		BR	-136.84	-145.60	0.9586	78.82
	MI	ML	-137.31	-146.10	0.9566	79.05
		BC	-136.40	-145.20	0.9566	78.60
		BR	-136.40	-145.20	0.9566	78.60

Table 3: criteria for comparison at 10% missing of data

Link function	Methods	ESTIMATOR	BIC	AIC	Pse R^2	Log-lik

Logit	EM	ML	-101.59	-110.38	0.8800	61.19
		BC	-100.67	-109.47	0.8800	60.73
		BR	-100.67	-109.47	0.8801	60.73
	MI	ML	-85.040	-93.840	0.8178	52.92
		BC	-84.120	-92.910	0.8178	52.46
		BR	-84.120	-92.910	0.8178	52.46
Probit	EM	ML	-103.75	-112.54	0.8886	62.27
		BC	-102.83	-111.63	0.8886	61.81
		BR	-102.83	-111.63	0.8886	61.81
	MI	ML	-86.570	-95.360	0.8141	53.68
		BC	-85.640	-94.440	0.8141	53.22
		BR	-85.640	-94.440	0.8141	53.22
Clog-log	EM	ML	-99.620	-108.42	0.8740	60.21
		BC	-98.710	-107.50	0.8740	59.75
		BR	-98.710	-107.50	0.8741	59.75
	MI	ML	-83.730	-92.530	0.8221	52.26
		BC	-82.810	-91.600	0.8221	51.80
		BR	-82.810	-91.600	0.8221	51.8
Cauchit	EM	ML	-86.720	-95.520	0.6219	53.76
		BC	-85.800	-94.600	0.6218	53.30
		BR	-85.800	-94.590	0.6224	53.30
	MI	ML	-74.410	-83.200	0.6523	47.60
		BC	-73.480	-82.270	0.6520	47.14
		BR	-73.470	-82.260	0.6522	47.13
Log	EM	ML	-97.250	-106.04	0.8642	59.02
		BC	-96.330	-105.12	0.8642	58.56
		BR	-96.330	-105.12	0.8643	58.56
	MI	ML	-82.170	-90.970	0.8239	51.48
		BC	-81.250	-90.040	0.8239	51.02
		BR	-81.240	-90.040	0.8239	51.02
Log-log	EM	ML	-106.320	-115.12	0.8854	63.56
		BC	-105.40	-114.20	0.8854	63.10
		BR	-105.40	-114.20	0.8854	63.10
	MI	ML	-88.440	-97.240	0.7934	54.62
		BC	-87.510	-96.300	0.7933	54.15
		BR	-87.510	-96.300	0.7933	54.15

Table 4: criteria for comparison at 15% missing of data

Link function	Methods	ESTIMATOR	BIC	AIC	Pse R^2	Log-lik
Logit	EM	ML	-99.700	-108.50	0.8562	60.25
		BC	-98.790	-107.58	0.8562	59.79
		BR	-98.790	-107.58	0.8562	59.79
	MI	ML	-94.610	-103.41	0.8443	57.70
		BC	-93.690	-102.49	0.8443	57.24
		BR	-93.690	-102.49	0.8443	57.24
Probit	EM	ML	-102.40	-111.19	0.8722	61.59
		BC	-101.48	-110.27	0.8722	61.14
		BR	-101.48	-110.27	0.8722	61.14
	MI	ML	-96.160	-104.96	0.8507	58.48
		BC	-95.240	-104.04	0.8507	58.02
		BR	-95.240	-104.03	0.8507	58.02
Clog-log	EM	ML	-97.190	-105.99	0.8443	58.99
		BC	-96.270	-105.07	0.8442	58.53
		BR	-96.270	-105.07	0.8442	58.53

	MI	ML	-93.280	-102.07	0.8376	57.04
		BC	-92.360	-101.16	0.8376	56.58
		BR	-92.360	-101.15	0.8376	56.58
Cauchit	EM	ML	-83.330	-92.120	0.5952	52.06
		BC	-82.410	-91.200	0.5945	51.60
		BR	-82.400	-91.200	0.5945	51.60
	MI	ML	-82.620	-91.410	0.6554	51.71
		BC	-81.690	-90.490	0.6551	51.24
		BR	-81.690	-90.480	0.6552	51.24
Log	EM	ML	-94.120	-102.90	0.8299	57.46
		BC	-93.200	-102.00	0.8299	57.00
		BR	-93.200	-101.99	0.8299	57.00
	MI	ML	-91.480	-100.27	0.8295	56.14
		BC	-90.560	-99.350	0.8294	55.68
		BR	-90.560	-99.350	0.8294	55.67
Log-log	EM	ML	-105.78	-114.58	0.8859	63.29
		BC	-104.86	-113.66	0.8859	62.83
		BR	-104.86	-113.66	0.8859	62.83
	MI	ML	-97.630	-106.43	0.8540	59.21
		BC	-96.710	-105.50	0.8540	58.75
		BR	-96.700	-105.50	0.8540	58.75

Figure 1: AIC values at 5%, 10%, and 15% levels of missing data

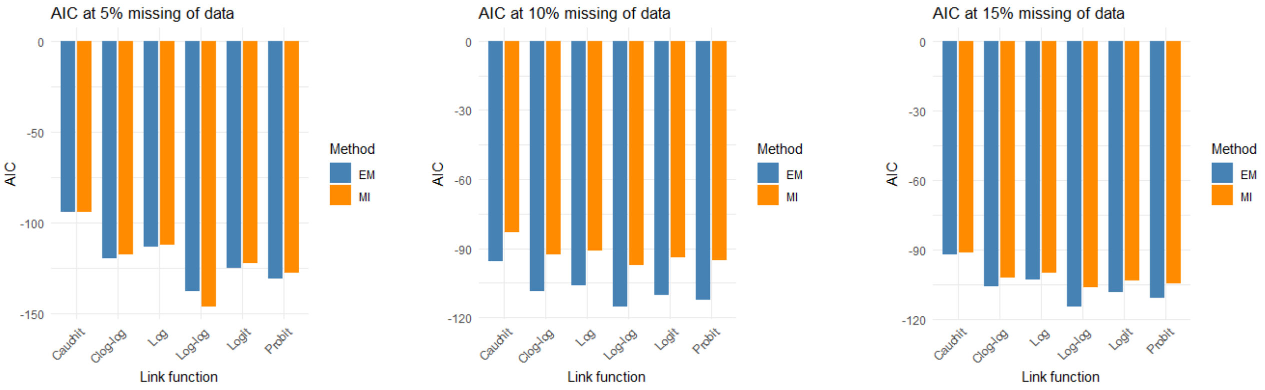


Figure 2: BIC values at 5%, 10%, and 15% levels of missing data

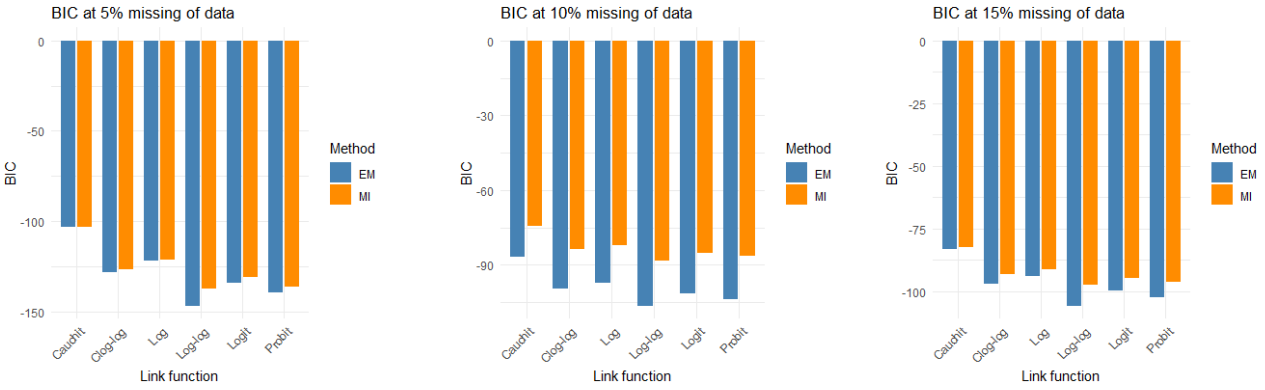


Figure 3: Pseudo R² values at 5%, 10%, and 15% levels of missing data

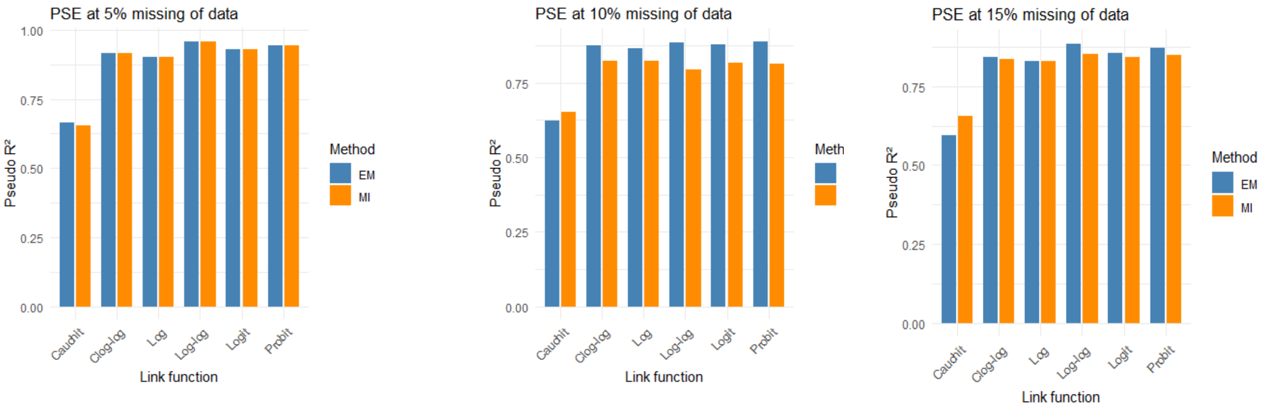
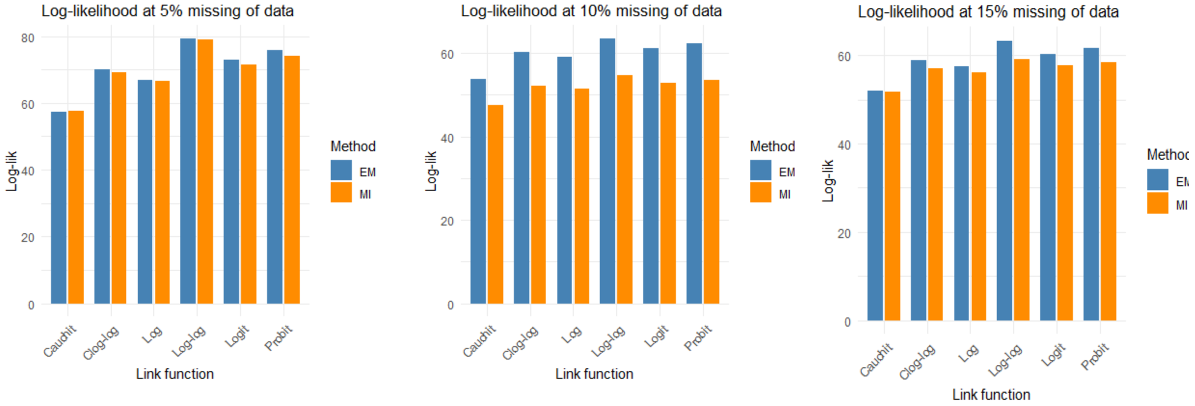


Figure 4: Log-Likelihood values at 5%, 10%, and 15% levels of missing data



5. Conclusion

This study compared two widely used imputation methods, the Expectation-Maximization (EM) algorithm and Multiple Imputation (MI), for handling missing data in beta regression models under varying percentages of missing data. Result indicate that although both methods preform adequately at lower missingness levels, EM accurate and robust parameter estimates compared to MI, Notably, EM's superiority was most evident at moderate missingness levels (around 10%) where is achieved substantial improvement in model fit and explanatory power across all model evaluation criteria. As missingness increased to 15% EM maintained its advantage over MI, but the performance gains stabilized, suggesting that EM is particularly effective under moderate to high missing data conditions without further pronounced improvement at extreme levels. Additionally, the choice of link function played a significant role in model performance, with the log-log link function, consistently producing the best fit, while the Cauchit link function resulted in the weakest outcomes across misssingness levels.

These findings provide valuable guidance for researchers and practitioners in selecting effective techniques for addressing missing data in beta regression models.

References

[1] Maluf, Y. S., Ferrari, S. L., &Queiroz, F. F. (2024). Robust beta regression through the logit transformation.*Metrika*, 1-21.

[2] Zhao LP, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equations. Biometrics. 1996 Dec;52(4):1165-82. PMID: 8962448.

[3] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations," *American psychologist*, vol. 54, p. 594, 1999.

[4] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts," *BMC Medical Research Methodology*, vol. 17, p. 162, 2017/12/06 2017.

[5] Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers.*SpringerPlus*, *2*, 1-17.

[6] H. Hasan, S. Ahmad, B. M. Osman, S. Sapri, and N. Othman, "A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: A simulation study," in *Mathematical Sciences Exploration for the Universal Preservation*, 2017, p. 060003.

[7] S. Avtar, G. Khuneswari, A. Abdullah, J. McColl, C. Wright, and G. Team, "Comparison between EM algorithm and multiple imputation on predicting children's weight at school entry," in *Journal of Physics: Conference Series*, 2019, p. 012124.

[8] S. Ferrari, F. Cribari-Neto, Beta regression for modeling rates and proportions, *J. Appl. Stat.*, **31** (2004), 799–815. doi: 10.1080/0266476042000214501.

[9] F. Bayer and F. Cribari-Neto, "Model Selection Criteria in Beta Regression with Varying Dispersion," *Communications in Statistics - Simulation and Computation*, vol. 46, p. 729, 2017.

[10] T. K. Ribeiro and S. L. Ferrari, "Robust estimation in beta regression via maximum L q-likelihood," *Statistical Papers*, vol. 64, pp. 321-353, 2023.

[11] M. Qasim, K. Månsson, and B. Golam Kibria, "On some beta ridge regression estimators: method, simulation and application," *Journal of Statistical Computation and Simulation*, vol. 91, pp. 1699-1712, 2021

[12] C. K. Enders, *Applied missing data analysis*: Guilford Publications, 2022.

[13] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured
datasets: a systematic review," *BMC Medical Research Methodology*, vol. 24, p. 188, 2024.

[14] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts," *BMC Medical Research Methodology*, vol. 17, p. 162, 2017/12/06 2017.

[15] A. Gelman, *Data analysis using regression and multilevel/hierarchical models*: Cambridge university press, 2007

[16] M. S. Santos, J. P. Soares, P. Henriques Abreu, H. Araújo, and J. Santos, "Influence of data distribution in missing data imputation," in *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, 2017, pp. 285-294.

[17] Biering K, Hjollund NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. ClinEpidemiol. 2015 Jan 16;7:91-106. doi: 10.2147/CLEP.S72247. PMID: 25653557; PMCID: PMC4303367

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, pp. 1-22, 1977.

[19] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley

[20] N. Prater, "Estimate gasoline yields from crudes," *Petroleum Refiner*, vol. 35, pp. 236-238, 1956.