A Study of Multicollinearity Detection and Rectification under Missing Values

Alhassan Umar Ahmad, U.V. Balakrishnan, Prem Shankar Jha
1Ph.D. Research Scholar Department of Mathematics Sharda University, Greater Noida UP Delhi NCR. Nigeria.
2,3Professor Department of Mathematics Sharda University, Greater Noida UP Delhi NCR. India.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstracts
In this paper, the consequences of missing observations on data-based multicollinearity were analyzed. Different missing values has a different effect on multicollinearity in the system of multiple regression model. Therefore, to ascertain the clear relationship between both multicollinearity and skipping values on monotone and arbitrary missing values, the collinear effects were potentially studied on two types of missing values. Similarly, the comparison was done to investigate each response of multicollinearity on each pattern of the missing values with the same informatics data. It was found that tolerance and variance inflation factor fluctuates due to the missing of information from the sample analyzed at a different percentages of the missing values.It was observed that the more missing values available in the sample obtain from either population statistics or survey than multicollinearity will be found in the system of multiple regression, this is because as the number of Missingness increase it shows a drastic decrease from the tolerance level on both monotone and arbitrary types as observed from the analysis.

Keywords; Missing Value, Monotone, Arbitrary, Imputation, and Multicollinearity.

1. Introduction

Today missing data is becoming more challenging than ever before, this is due to the rapid advancement in technology of computations couple with the current statistical techniques enhancing the analysis of variance. The higher demand for essential accuracy and efficient reliability from Governments, industries, and non-Governmental organizations to obtain better achievements during implementation and executions of their policies make a special topic that requests attention. This demand calls for the need to discover more on missingness and multicollinearity to minimized errors (Peugh, J.L and Enders, C.K. 2004). Absent of complete or partial information due to nonresponse or from any experimental research is problematic in statistics, this is because it will affect the result and render it invalid. It was well known that nearly most of the standard statistical procedures and techniques need complete data to process at higher accuracy but in most cases, it has no provision to handle the missing information, therefore missing observations will essentially reduce and shrunk the sample size which in return will directly increase the level of the standard error during analysis Marina, Soley-Bori (2013). Similarly, it will affect the precision of the confidence intervals which normally lead to type I error after the analysis (SAS Institute, 2005). Missing information is a big problem affecting the manual database, the electronic database it is sometimes responsible for making most of the statistical packages inactive (Catia M. ET, al. 2016).

Missing value occurs at any time and in any given experiments, surveys, or population study no matter how it is well designed and implemented. This missing data always undermines the efficiency and precision of research investigations. It brings much of instability and no reliability in processing and making of a final inference on any statistical investigation Marina, Soley-Bori (2013) and (Graham, J.W. 2009). The statistical power and ability of a model are compromised due to the absence of sensible information which is necessary to complement all the realities involves in any statistical analysis. Similarly, missing value brings a biased estimate that provides invalid interpretations due to the error originated from a lack of complete information during the recording of data collection in the survey field (Korean, J. Anesthesiol 2013). It is well known that missing information occurs when there is no data value or variable recorded for an observation. Missing data is a common phenomenon in a statistic. It has a significant effect on the conclusion of analysis and can affect it is final precision Acock AC (2005).Imputation procedures are normally used to correct the missing information but it requires a careful study of the data pattern, nature of the missingness and sequence at which the available information come into being, even though it is important if possible to trace the reasons of the missingness if is available before engaging to corrects the missingness (Pourahmadi, M. 1989). researchers in the empirical research field are now putting much serious

effort on how to treat the problem arising from the missing date, this is because it is now clear during statistical survey some selected respondents may voluntarily refuse to give out any information more especially when it comes to private information despite missing of such participation from the respondents always appear to affect the survey negatively Graham, J.W. (2009). One of the best solutions of the missingness of data a surveyor shall not allow anyone to happen and to achieve this one has to be careful in making designed and ensure good execution of the entire research procedures in the field because all the statistical methods and adjustments that will be applied to make corrections or imputations when missingness occurs will never be as appropriate as the original observations (Paul D. Allison, 2001).

1.1 Multicollinearity

Multicollinearity is a state of having higher inter-correlation among the dependent response and independent explanatory variables in multiple regression equations due to the existence of linear relationships among variables in the model (Farrar, D. E., and Glauber, R. R. 1981). To provide more reliable inferences at the time of the dissemination of result, the effect of missing values that reduce the representativeness of the desired sample to be studied from the main population which has a direct influence on multicollinearity do to shrinkage of size from the sample was carefully studied (Jamal, I. Daoud 2017) and (O'Hagan and Brendan 1975).

Farrar and Glauber 1969 study about the severity of multicollinearity in which it was categorized into non-harmful, medium, and severe multicollinearity in linear relationships among explanatory variables. Yoel Haitovsky also in his paper title Multicollinearity in Regression Analysis (1969) explained the existence or non-existence of multicollinearity in a system of multiple regression analysis. And In 1975 Farrar and Glauber proposed a criteria for detecting multicollinearity presence, which repressor variables are collinear and the nature of multicollinearity by the use of chi-square, F-test and T-test respectively. And John O'Hagan and Brendan McCab in 1975 study the tests for the severity of multicollinearity in Regression analysis while in this paper we have studied the implication of multicollinearity with missing values base on the parentages of the missing information. It was observed that large missing values are associated with higher multicollinearity found in the system of multiple regression analysis.

Besides,in section 1. This paper explains the concept of multicollinearity and missing value altogether, section 2. Talk about missing value mechanism, types, and a class of the missing values and reason for missing data either from any of dependents or independents variables. Section 3. It deals with the principals of obtaining dependent, explanatory variables, co-efficient of independent variables in a complete system of multiple regression analysis. In section 4. The paper provides insight into the statement of the problem, pattern, class of the missing value, and imputation technique for correction of the missing information. Section 5. Detects multicollinearity based on Monotone and Arbitrary types of missing data at a different level and percentages of the missing information and also present graphs which visualize the effects and consequences of both multicollinearity and missing data together by percentages. Section 6. Concluded the finding of the study where it discovered that both multicollinearity and missing values has negatives effect or bauble tragedy on the system of multiple regression analysis.

1.2 Missing-Value Mechanisms

It is on records data is been missed due to different reasons at the time of investigation or during data mining and processing. Of whatever reasons it needs to be handled scientifically to avoid the adverse negative effect of any missing information that occurs along the way due to one reason or another Marina, Soley-Bori (2013). Scientists who are working in the data visualization process are working hard to see that the issue of missing value is always properly addressed to avoid a shortage from the sample to be visualized or analyze. There are many methods in existence to replace the missing value but always it depends on the nature and pattern of the missingness.

2.2Missing value completely at random

Missing any of the explanatory variables completely at random occurs only if all the explanatory and response variables in the model have the same and equal chance of been missed along the process. In this case, the deletion or ignoring of any missing information base on either of the row or column affect the final inference drawn after the analysis. If the sample size is enhancing than the least square coefficient will be more consistent and unbiased (Graham, 2009). Even though efficiency measure the optimality of the estimator in the recovery of missing information essentially (Allison, 2001) and (Briggs et al., 2003).

2.3Missing Value at Random

The probability that a response or a predictor variable is missing from a set of samples is depended only on the availability of the obtained information from an investigation. It depends on the accessibility of the available recorded information. This can be linked with the process of logistic regression dealing with either 1 or 0 in place of availability or missing of a value a variable respectively Pourahmadi, M. (1989). When an explanatory variable is missed at random is acceptable to exclude the missing cases from either raw or any column from the investigation if the multiple regression model controls all the response that affect the probability of the missing observation. Missing value at random is a much more realistic assumption to study the performance and accuracy of the recovery procedure because missing information at random is term as the probability that response and predictor variables are missing base on the set of the observed responses (Schafer, 1997).

2.4Missing Value Due to Unobserved Predictors

This process is no longer at random but only depends on the accessibility of the information which has not been recorded and it can predict the missing observation from an investigation Hyun Kang (2013). The process must be modeled explicitly if the missing information is not at random or else acceptance of bias in the interpretation will be non-avoidable (Graham, J.W. 2009).

2.5Missing Value depends on the Missingness itself

His is where the missing of information is leading to another missing of inflation on some response and predictor variables Paul T. von Hippel (2004). A difficult situation normally arises where the probability of missingness depends on potentially missing information from the variable that is supposed to be in the sample (Schafer, 1997).

2.6Reasons for Missingness of a Data

Taking into consideration the main reasons for missing data always pre-empty a preview on the best mechanism to adopt for the systematic recovery of the missing information. Numerous cases have different reasons for the Missingness ranging from the time of design of a survey to the process of data mining activities, recoding phase, analysis, and interpretation. The missing value may occur due to the escape of one's memory or lost, non-applicability of the value at the instance, lack of interest at the point of recording, the variable was measured but not recorded due to identify or unidentified technical errors from the database (i.e. Disconnection of sensors, errors in communicating the value, accidental human omission, electricity failures), (Young W, Weckman G, Holland W 2011).

It was established and well distinguished between data missing that has identifiable or no identifiable reasons. This is redefining the nature and status of the missing information as either recoverable or not recoverable. Whenever information was missed for unidentified reasons it normally terms with the assumption that the missing is at random and unintended causes such type of Missingness is always classified as the recoverable one and otherwise, if the reasons for the Missingness are identified and is for reasons such is always no recoverable. Many a time the nature of the missingness and it is assumption are used to illustrate which type of methods shall be employed to recover the missing infarction.

3.0 Materials and Methods

From the principal of standard linear regression analysis that involve the algebraic formula in which dependent variable (Y), independent variables (Xi), coefficient of the explanatory variables (βi) and the statistical error term (α)

From the predictions process below we have

$Y = a + \beta i\ Xi$……..(1)

Such that…

Y = A predicted value of Y base on the available explanatory variables

a = Y value whenever X value is equivalent Zero (Y Intercept)

βi = Change in Y value for a small change in Xi

Xi = independent variables that are used to predict a value of the dependent variable Y

For Multiple Regression equation

$Y = a + \beta_1 X_2 + \beta_2 X_2 + \ldots\ldots$ (2)

Y = A predicted value of Y (which is your dependent variable)

a = the Y Intercept

β_1= the change in Y for each 1 increment change in X_1values

β_2= the change in Y for each unit increment change in the X_2 value

X =value of X (X is the Independent Variable) which can predict a value of Y as dependent variable)

Calculating the Regression Coefficients β_1 and β_2 the formulae are given below

$$\beta_1=\left[\frac{r_{y,x1}-r_{y,x2}r_{x1,x2}}{1-\left(r_{x1,x2}\right)^2}\right]\left[\frac{SD_y}{SD_{x1}}\right]\ldots\ldots\ldots\ldots\ldots \quad (3)$$

$$\beta_2=\left[\frac{r_{y,x2}-r_{y,x1}r_{x1,x2}}{1-\left(r_{x1,x2}\right)^2}\right]\left[\frac{SD_y}{SD_{x1}}\right]\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Whereas;

ry,x1= Correlation between blood pressure and age x1.

ry,x2 = Correlation between blood pressure and weight x2.

rx1,x2 = Correlation between age x1 and weight x2

(rx1, x2)2 = the coefficient of determination (r squared) between age x1 and weight x2

SDy = Standard Deviation for our Y (dependent) variable.

SDx_1= Standard Deviation for age X1

SDx_2= Standard Deviation for weight x2.

To find the coefficient of determination we have the following below;

$$R=\sqrt{\frac{\left[\left(r_{y,x1}\right)^2+\left(r_{y,x2}\right)^2\right]-\left(2r_{y,x1}r_{y,x2}r_{x1,x2}\right)}{1-\left(r_{x1,x2}\right)^2}}\ldots. \quad (5)$$

Whereas;

ry,x1= Correlation between blood pressure and age x1.

ry,x2 = Correlation between blood pressure and weight x2.

rx1,x2 = Correlation between age x1 and weight x2

(rx1, x2)2 = the coefficient of determination (r squared) between age x1 and weight x2

3.1 Variance Inflation Factor (VIF)

Variance inflation factor is determine to explore the level of multi-linear relationships that always exist in between variables (Johnston 1972; Green 1990; Kroll et al. 2004). Each explanatory variable is regressed against the other remaining explanatory variables and response variable.And the VIF is calculated as:

$VIF = R/ (1-R^2)$…….. (6)

Where R^2 is the regression model coefficient of determination (Rawlings et al. 1998). A VIF greater than 10 is a common threshold in detecting severe multicollinearity.Variance inflation factor inflate sample variance and the dependence properties of the variables involved in the system

Tolerance level is one unit minus the potion of variance an explanatory variable shares with the other independent variable which is not map to by other predictor variables and in the same vein tolerance is a measure of multicollinearity obtain from statistical analysis like SPSS, the variable's tolerance is given below;

Tolerance = $(1-R_2)$….. (7)

In recent days investigators and scientific researchers more especially those from a statistical point of views dose not doubt the challenges and fear due to the influence of both phenomenon's multicollinearity and missing value presence in any statistical project ranging from a survey, population study, statistical analysis and other areas where statistical application on demand for evaluation activities, this is because every day accuracy and precision is on high demand by Governments, Industries and other non-governmental organization from National, international and global perspective for the purpose of good achievement and higher delivery of their aims and objectives in order to reaffirm the execution of their policies but unfortunately those two phenomena's bring instability and rises the chances of higher standard error which affects good estimation and prediction process of which finally causes high non-reliability to the final inference drawn after all the statistical procedures.

3.2 Statement of the problem and the data to be analyzed

The study is aimed at investigating how multicollinearity is related either directly or indirectly with missing values as both phenomena have simple to complex implications on the accuracy of the final result of most statistical studies. It also discusses the fluctuation of variance inflation factors and tolerance level in every phase by percentages of the missing values involved from analysis.

The data obtained officially for this study only from the records Department of Sharda University Hospital in greater Noida, UP. Delhi NCR. The data has the following classifications with all what the variables stands for in which Blood Pressure (X_3), stands as a dependent variable while Age AG (X_1), Body Weight BW (X_2), Random Blood Sugar RBS (X_4), Body Temperature BT (X_5), Pulse Rate PR (X_6), Blood Oxygen Saturation SPO2 (X_7) as the independent variables respectively.

3.3Patterns of missingness

The nature of the missing patterns influences the stability of the analysis, this is because some missing values can be recoverable and no recoverable base on the type of the sample available. Predictions and estimations are found to be much more essential and stable if there is no missing value at all and the power of the prediction mechanism remains unaltered. The missing value pattern has it is own influence on the size of a sample and multicollinearity also has it is the effect on the missingness, this is also because missingness has a direct effect on sample sizes of each observation.

3.4Monotone type of missing values

This type of missing pattern can be generated due to a specification of a sequencing method that is unilabiate and occurs normally in a Colum wise making a section of some Colum to be incomplete especially toward the end of the columns. This method has a series of synthetic observations in which the missing information is happening always (Ruben 1987b) and this type of missing data occurs in a longitudinal study with drop-out of a section of information (SAS Institute, 2005).

S/N	X_1	X_2	X_3	X_4
1	85	79	140	140
2	45	62	120	103
3	60	54	190	94
4	92	88	104	127

5	44	76	235	125
6	85	79	140	-
7	45	62	-	-
8	60	54	-	-
9	92	88	-	-
10	44	-	-	-
11	45	-	-	-
12	-	-	-	-

Sample of the Monotone messiness of data

3.5Missing Data Arbitrarily

This type of missing pattern occurs at random irrespective of row or column-wise. The data is missing without abeyance to any order or a pattern and type. (SAS Institute, 2005) and (Enders & Bandalos, 2001).

S/N	X_1	X_2	X_3	X_4
1	85	79	140	-
2	45	-	-	103
3	60	54	190	94
4	-	88	-	127
5	44	76	235	125
6	45	59	-	-
7	-	-	140	-
8	45	62	120	103
9	60		190	
10	-	88	104	127
11	-	76	-	125
12	45	-	180	140
13	60	75	238	-

Sample of the arbitrary messiness of data

4.0Remedies for Missing Values

There are many procedures use to handle the issue of Missingness such as deletion, ignoring, imputation and Model-Based Methods (regression, multiple imputation, k-nearest neighbors), (Catia M. Salgado, et al, .2016) in May cases if the sample of the Missingness is small from the data to be analyzed very roughly, less than 5% of the total number of from either respond or explanatory variables and all the missing values occurs at random that is, whether the missing information is not depends up another values then the typical method of leastwise deletion is relatively "safe" than directly use delete and ignore procedures to get rid of missing value involved in the sample, delete and ignore process involve deletion or ignoring of the whole raw or column with defect of missing information but if the number of missing value is large from a big data sample the best way to resolve is that of imputation principle which is based on prediction and estimation from the existing information available from the original data (Steffi Pauli Susanti and Fazat Nur Aziza 2017). Single Imputation Methods involves the technique and of using mean/mode substitution, linear interpolation, hot deck, and cold deck (Marina, Soley-Bori 2013). And. This process will enhance the data and will lead to having complete information on each of the variables involves and because the sample size is improved therefore it will automatically reduce the effect of multicollinearity on the data (Pourahmadi, M. 1989).

A unit of cell or cells that happened to be missing data, where particular information of a variable is not available at all, then excluding or removal of such unit from the entire analysis is the most paramount to avoid the negative consequences of the missing variable in the analysis. This is usually considered as the default of the statistical packages and procedures (Briggs et al., 2003).

4.1Two Main Imputation Techniques

Imputation criteria include the use of neural network methods, Bayesian network, regression process it always depends on the nature of the missing value and sample data that is available (Nakai and Weiming, 2011) and Marina, Soley-Bori (2013).

4.2Marginal mean imputation criteria

This always involves the use of marginal variables in between the missing value and calculate the arithmetic mean that can be used to cover up the missing information (Schafer, 1997). It normally led to the eventual biased estimate of the variances and covariance which is generally is not needed because it affects the inference during analysis (Graham, 2009).

4.3Conditional mean imputation;- This procedure is used if explanatory variables have missing of particular information from among the variables to be analyzed, then the available variables can be used to estimate the missing value using multiple regression model to obtain the missing value with higher accuracy and precision (Briggs et al., 2003). This procedure will maintain higher reliability from the source of the data to minimize the chance of type ii error in the system (Allison, 2001).

5.0Detecting multicollinearity

Among the measures use for detecting the existence of multicollinearity from the model as a result of the missing values here we consider tolerance and variance inflation both indicators used here were computed for the regression parameters of all response and predictor variables present in the system of the regression model. Collinear relationships were revealed by different level of the missing values from both monotone and arbitrary types in which we closely monitor the level of deteriorations as the percentage of the missing values keep on increasing.

It was observed from different tables and percentages of the missing values variance inflation factors suffer from constant inflation from the variance of their parameters, the tolerance level also goes down as the percentages of the missing values keep on accelerating, and as such the standard error of all estimate relatively increased. It was discussed earlier monotone type of missing information it occurs normally at the end of the table and as a particular pattern unlike the arbitrary type of the missing information which occurs at random and it has no specific pattern at all

Table 01. Monotone Missingness at 0%

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	172.111	132.984		1.294	.197		
	X1	-.050	.143	-.028	-.350	.727	.805	1.243
	X2	.110	.154	.055	.712	.477	.831	1.204
	X4	.152	.137	.082	1.107	.270	.910	1.099
	X5	-.134	1.289	-.008	-.104	.917	.960	1.042
	X6	-.370	.150	-.188	-2.466	.015	.855	1.169
	X7	.151	.302	.036	.502	.616	.979	1.022

a. Dependent Variable: X3

From the above table 01, it indicates where the redundancy of an explanatory variable is relied more upon than any other among the explanatory variables more precisely X1 because of lower value of tolerance which is the measure or account of variability in the independent variable x1 which is never accounted for by other predictor variables present in the system and henceforth it is affected by multicollinearity more than all other explanatory variables in the model due to the low tolerance of 0.805 and equally having higher value of 1.243 as variance inflation factor which indicated how much the variance was inflated.

Multicollinearity as a measure of linear relationships among response and explanatory variables that are moderately or highly correlated either from a database or structural sources and from table 01 base on t-statistics and history of the correlation then we shall eliminate x1 to get rid of the existence multicollinearity in the system.

Table 02. Monotone Missingness at 5%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	195.599	146.873		1.332	.185		
	X1	-.061	.148	-.034	-.414	.679	.793	1.261
	X2	.144	.162	.071	.888	.375	.821	1.218
	X4	.151	.140	.082	1.073	.285	.908	1.101
	X5	-.359	1.425	-.019	-.252	.801	.958	1.043
	X6	-.381	.154	-.193	-2.468	.015	.854	1.172
	X7	.129	.312	.030	.415	.679	.983	1.017

Dependent Variable: X3

In table 02 above we have introduced 5% monotone type missing of Missingness and it shows that due to the sudden loss of about 5% of the data the level of tolerance practically fluctuate and deteriorate among all of the predictor variables where the tolerance of X1 changer from 0.805 to 0.793 which means about 1.5% of the tolerance was deteriorated due to 5% monotone Missingness from the data. While VIF which changes from 1.243 to 1.261 shows that there is 1.4% of the increase in the variability or the variance of an estimated regression coefficient is increased by 1.4% to raise the moderately discover collinear effect among the predictor variable.

Table 03. Monotone Missingness at 10%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	220.480	149.171		1.478	.141		
	X1	-.035	.153	-.019	-.231	.818	.777	1.287
	X2	.104	.171	.050	.611	.542	.807	1.239
	X4	.157	.143	.085	1.097	.274	.908	1.101
	X5	-.531	1.443	-.028	-.368	.713	.956	1.047

	X6	-.429	.158	-.219	-2.710	.007	.842	1.187
	X7	.098	.318	.023	.308	.759	.980	1.021

From the above table 03, it was 10% missing of data against what is on table01 it has to indicate a sudden change from tolerance level where it changes drastically changed from 0.805, 0.793 on table 01 and table 02 respectively but now change to 0.777 on table 03 indicating decreases intolerance as data is missed along the line due to missing values, this means there are additional deteriorations of 2.02% of the tolerance level due to the monotone missing values up to 10% of the total observations. Meanwhile for the IVF which is now changed from 1.261 to 1.287 shows that there is 2.0% of the increase in the variance inflation factors of an estimated regression coefficient.

Table 04. Monotone Missingness at 15%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	221.930	152.281		1.457	.147		
	X1	-.042	.158	-.023	-.269	.789	.779	1.284
	X2	.078	.178	.037	.440	.660	.807	1.238
	X4	.139	.147	.076	.948	.344	.905	1.105
	X5	-.487	1.472	-.026	-.331	.741	.956	1.046
	X6	-.434	.163	-.221	-2.659	.009	.842	1.187
	X7	.093	.326	.022	.285	.776	.978	1.023

a. Dependent Variable: X3

Equally the result continuer to change from one step to another in which and from table 04 there is at least 1% change as a result of the difference of the missing values of 10% to 15%. From the values of variance inflation, it was nearly less than a 1% increase in the variability.

Table 05. Monotone Missingness at 20%

Coefficientsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	210.312	155.050		1.356	.177		
	X1	-.064	.161	-.035	-.397	.692	.788	1.270
	X2	.051	.186	.024	.277	.782	.817	1.224
	X4	.157	.157	.085	1.002	.318	.867	1.153
	X5	-.404	1.499	-.022	-.269	.788	.950	1.052
	X6	-.446	.177	-.222	-2.521	.013	.806	1.241
	X7	.170	.334	.040	.509	.611	.984	1.016

Dependent Variable: X3

In the above table 05 from the variables x6 at 0% missing value, there is a change of tolerance from 0.855 which falls under low multicollinearity against 0.806 tolerance which indicates lower tolerance value that gives way to higher multicollinearity than before due to up to 20% missing values. And for the variance inflation factor also follow the same trend.

Table 06. Monotone Missingness at 25%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	219.223	158.826		1.380	.170		
	X1	-.069	.173	-.038	-.400	.690	.739	1.353
	X2	.015	.202	.007	.073	.942	.762	1.312
	X4	.130	.162	.071	.800	.425	.843	1.186
	X5	-.461	1.533	-.025	-.301	.764	.945	1.058
	X6	-.420	.186	-.208	-2.257	.026	.789	1.267
	X7	.184	.340	.045	.541	.589	.982	1.018

a. Dependent Variable: X3

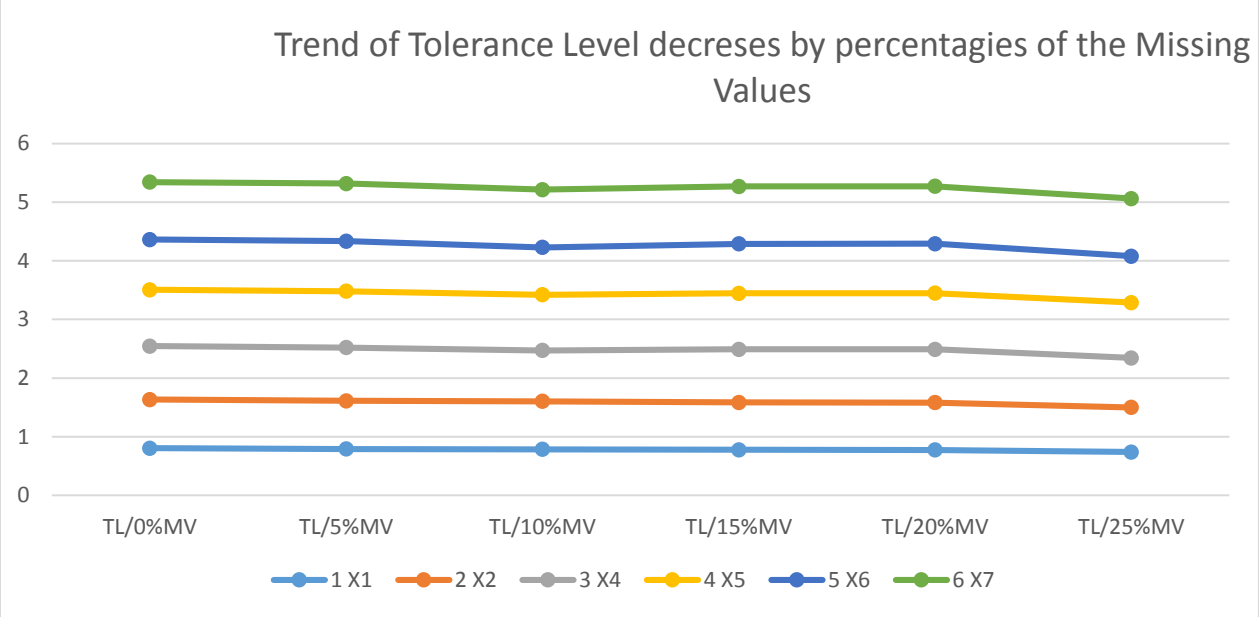
From table 06, X6 explanatory variable there is a change of tolerance level from 0.855 to 0.789 which accounts for about 7% variation due to 25% missing values by monotone pattern. Variance inflation factor change from 1.169 to 1.267 showing an increase in variation due to missing values effect in the model.

Table 07 Tolerance on different percentages by Monotone type of missing values

S/N	X_i	TL/0%MV	TL/5%MV	TL/10%MV	TL/15%MV	TL/20%MV	TL/25%MV
1	X_1	.805	.793	.788	.779	.777	.739
2	X_2	.831	.821	.817	.807	.807	.762
3	X_4	.910	.908	.867	.905	.908	.843
4	X_5	.960	.958	.950	.956	.956	.945
5	X_6	.855	.854	.806	.842	.842	.789
6	X_7	.979	.983	.984	.978	.980	.982

The above table 08 summarized all trend of variation which occurs due to change of the missing values at different percentages where it indicates decreasing values of tolerance by percentages of the missing information from the explanatory variables.

Fig. 01



From the above figure as the percentages of the missing values increases the tolerance values also decrease which indicates an increase in multicollinearity.

Table 09 VIF on different percentages by Monotone type of missing value

S/N	X_i	VIF/0%MV	VIF/5%MV	VIF/10%MV	VIF/15%MV	VIF/20%MV	VIF/25%MV
1	X_1	1.243	1.261	1.270	1.284	1.287	1.353
2	X_2	1.204	1.218	1.224	1.238	1.239	1.312
3	X_4	1.099	1.101	1.153	1.105	1.101	1.186
4	X_5	1.042	1.043	1.052	1.046	1.047	1.058
5	X_6	1.169	1.172	1.241	1.187	1.187	1.267
6	X_7	1.022	1.017	1.016	1.023	1.021	1.018

Above from table 09 it shows the amount of variability change based on percentage increase of the missing values.

Fig. 02

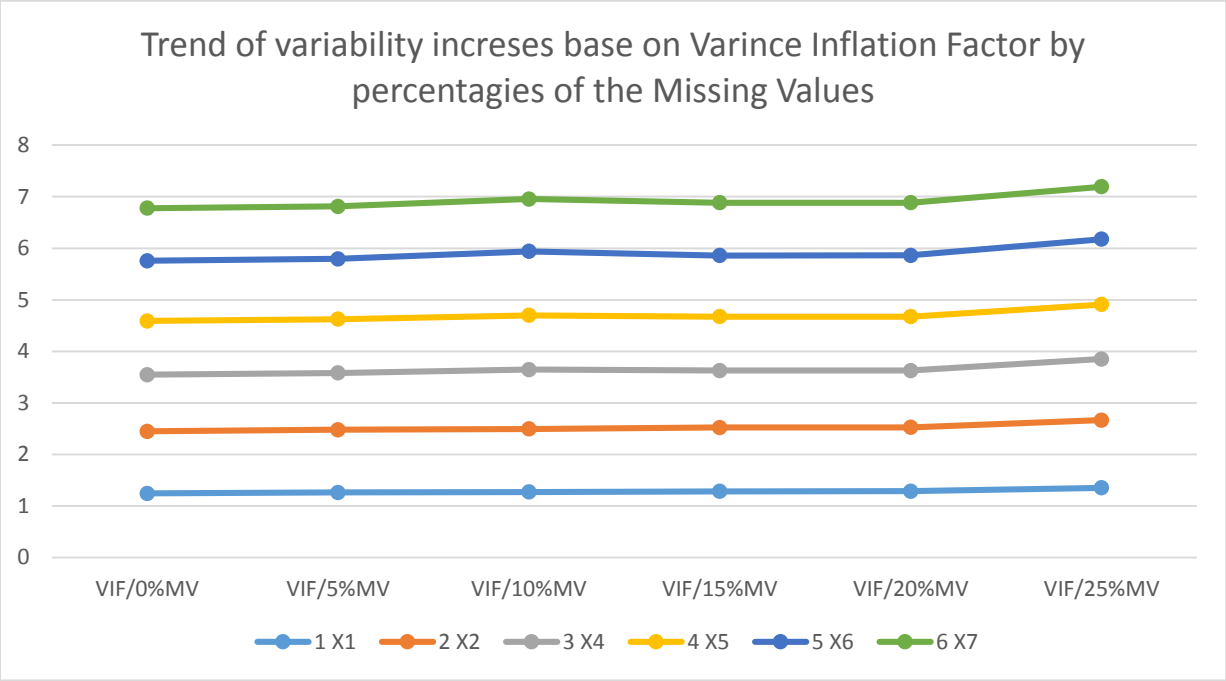


Table 07. Arbitrary Missingness at 0%

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	172.111	132.984		1.294	.197		
	X1	-.050	.143	-.028	-.350	.727	.805	1.243
	X2	.110	.154	.055	.712	.477	.831	1.204
	X4	.152	.137	.082	1.107	.270	.910	1.099
	X5	-.134	1.289	-.008	-.104	.917	.960	1.042
	X6	-.370	.150	-.188	-2.466	.015	.855	1.169
	X7	.151	.302	.036	.502	.616	.979	1.022

a. Dependent Variable: X3

From table 07 it shows that multicollinearity is relatively more contributes by the explanatory variable X1 than other predictor variables in the system, this is due to the small value of tolerance which is found with the explanatory variable X1 of about 0.805 and in the same vein the variance inflation factors of X1 was higher than predictors from the model which is up to 1.243 indicating more variability than others. Generally, table 07 shows more promising values of high tolerance from 0.805 to 0.979 which felled under a higher level of tolerance and it allows only the chances of a small or slight presence of collinear issue because of 0% missing values in the system.

Table 08. Arbitrary Missingness at 05%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const)	244.435	152.593		1.602	.112		
	X1	.009	.175	.005	.051	.959	.830	1.205
	X2	.074	.194	.034	.379	.705	.869	1.151
	X4	.040	.165	.022	.244	.807	.881	1.135
	X5	-.928	1.472	-.054	-.630	.529	.951	1.051
	X6	-.415	.182	-.211	-2.282	.024	.821	1.219
	X7	.384	.389	.086	.987	.326	.920	1.087

a. Dependent Variable: X3

From table 08, it was introduced 5% arbitrary missing of values and because of this there was a sudden change from the statistics where about 3.0% occurs among the tolerance level which change from 0.830 to 0.805, this is more the 5% monotone type of the missing values explained from table02. In which the percentage change is lower than what we obtained here for the reasons of randomness preteen of the missing values. Along the line variance inflation factors also change from 1.243 to 1.205 which means the variance of an estimate change to about 3%.

Table 09. Arbitrary Missingness at 10%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	333.500	213.832		1.560	.122		
	X1	-.083	.221	-.042	-.376	.708	.773	1.294
	X2	.094	.247	.040	.379	.706	.847	1.181
	X4	-.093	.200	-.049	-.465	.643	.847	1.181
	X5	-1.952	2.055	-.095	-.950	.345	.938	1.067
	X6	-.329	.231	-.162	-1.424	.158	.729	1.373
	X7	.624	.541	.115	1.154	.251	.947	1.056

a. Dependent Variable: X3

From the above table 09 it was 10% missing of data arbitrarily against what is on table01 it has indicated a significant change from tolerance level where it changes drastically changed from 0.805, 0.773 on table 07 and table 09 respectively but now change to 0.773 on table 09 indicating decreases intolerance as about 10% of the data missed arbitrarily 3% along the line due to missing values, this means there are additional deteriorations of 2.02% of the tolerance level due to the monotone missing values up to 10% of the total observations. Meanwhile for the IVF which is now changed from 1.243 to 1.294 shows that there is up to 4.0% of the increase in the variance inflation factors of an estimated regression coefficient

Table 10. Arbitrary Missingness at 15%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const)	493.968	275.770		1.791	.077		
	X1	.002	.242	.001	.007	.994	.795	1.257
	X2	.096	.267	.042	.359	.720	.890	1.124
	X4	-.137	.212	-.076	-.648	.519	.878	1.139
	X5	-3.396	2.694	-.139	-1.260	.212	.982	1.018
	X6	-.446	.254	-.217	-1.756	.083	.789	1.267
	X7	.537	.588	.103	.913	.364	.951	1.051

a. Dependent Variable: X3

From table 10 the results of tolerance continue to change from 0.795 against what is present on table 07 of 0.805. Sitting about 1.2% difference, this means it has now come down against what is presented in tables 09 to indicate a random fluctuation as a result of randomness in the missing values as a result of the arbitrary Missingness of information. %. From the values of the variance inflation factor, it fluctuates from 1.243 to 1.257 which was nearly less than a 1.1% increase in the variation.

Table 11. Arbitrary Missingness at 20%

Coefficientsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	-67.097	377.139		-.178	.860		
	X1	-.096	.282	-.054	-.340	.735	.704	1.421
	X2	.494	.362	.205	1.367	.179	.789	1.267
	X4	.315	.275	.168	1.147	.257	.828	1.208
	X5	1.909	3.731	.069	.512	.612	.989	1.012
	X6	-.800	.310	-.418	-2.580	.013	.677	1.476
	X7	.559	.792	.097	.706	.484	.943	1.061

Dependent Variable: X3

In the above table 11 from the variables x6 at 0% missing value, there is a change of tolerance from 0.855 which falls under low multicollinearity against 0.677 tolerance which indicates lower tolerance value that gives way to higher multicollinearity than before due to up to 20% missing values arbitrarily. And for the variance inflation factor also follow the same trend.

Table 12. Arbitrary Missingness at 25%

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Const.)	647.438	2856.392		.227	.823		
	X1	-.108	.508	-.061	-.213	.833	.463	2.159
	X2	.129	.582	.055	.222	.826	.629	1.590
	X4	.335	.403	.174	.831	.415	.871	1.148
	X5	-4.624	29.472	-.035	-.157	.877	.744	1.343
	X6	-.789	.549	-.396	-1.438	.165	.502	1.992
	X7	-.003	1.333	.000	-.002	.998	.707	1.414

a. Dependent Variable: X3

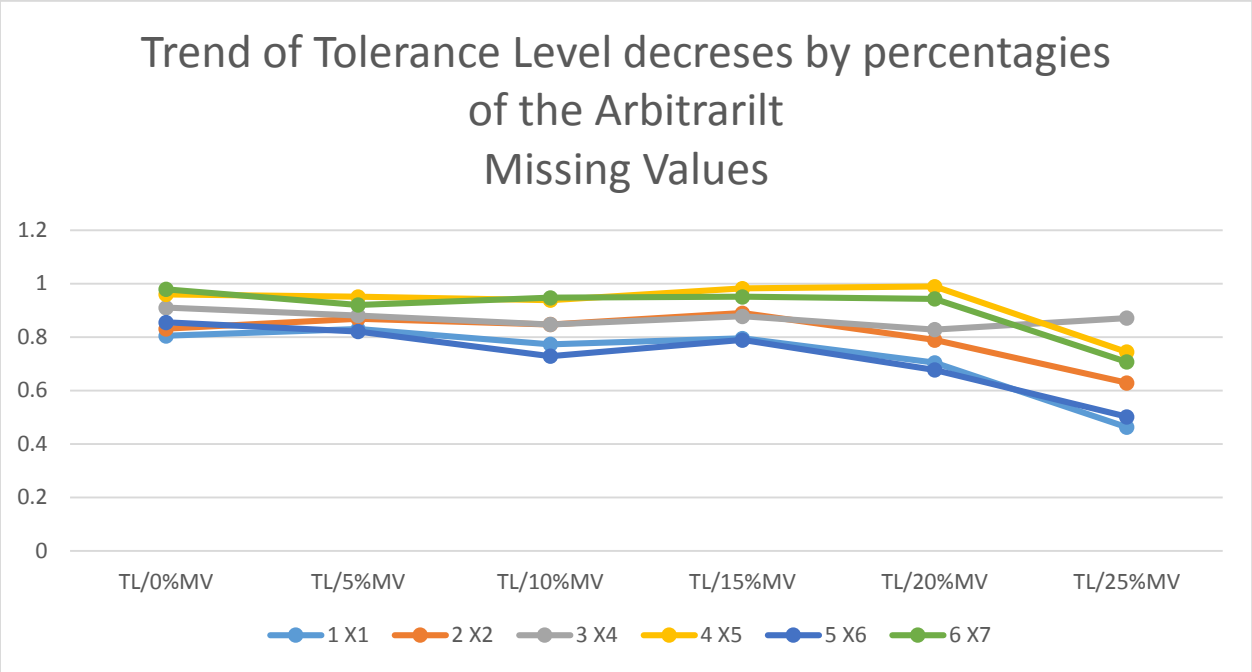
From table 06, X6 explanatory variable there is a change of tolerance level from 0.855 to 0.502 which accounts for about 40% variation due to 25% missing values by arbitrary pattern. Variance inflation factor change from 1.169 to 1.992 showing an increase in variation due to missing values effect in the model.

Table 13 Tolerance on Different Percentages by Arbitrary Type of Missing Value

S/N	X_i	TL/0%MV	TL/5%MV	TL/10%MV	TL/15%MV	TL/20%MV	TL/25%MV
1	X_1	0.805	0.83	0.773	0.795	0.704	0.463
2	X_2	0.831	0.869	0.847	0.89	0.789	0.629
3	X_4	0.91	0.881	0.847	0.878	0.828	0.871
4	X_5	0.96	0.951	0.938	0.982	0.989	0.744
5	X_6	0.855	0.821	0.729	0.789	0.677	0.502
6	X_7	0.979	0.920	0.947	0.951	0.943	0.707

Above table 13 summarized all trend of variation which occurs due to increasing change of the missing values at random from different percentages where it indicates decreasing values of tolerance by percentages of the missing information from the explanatory variables.

Fig. 03Table 14 Variance Inflation Factor on Different Percentages by Arbitrary Type of Missing Values



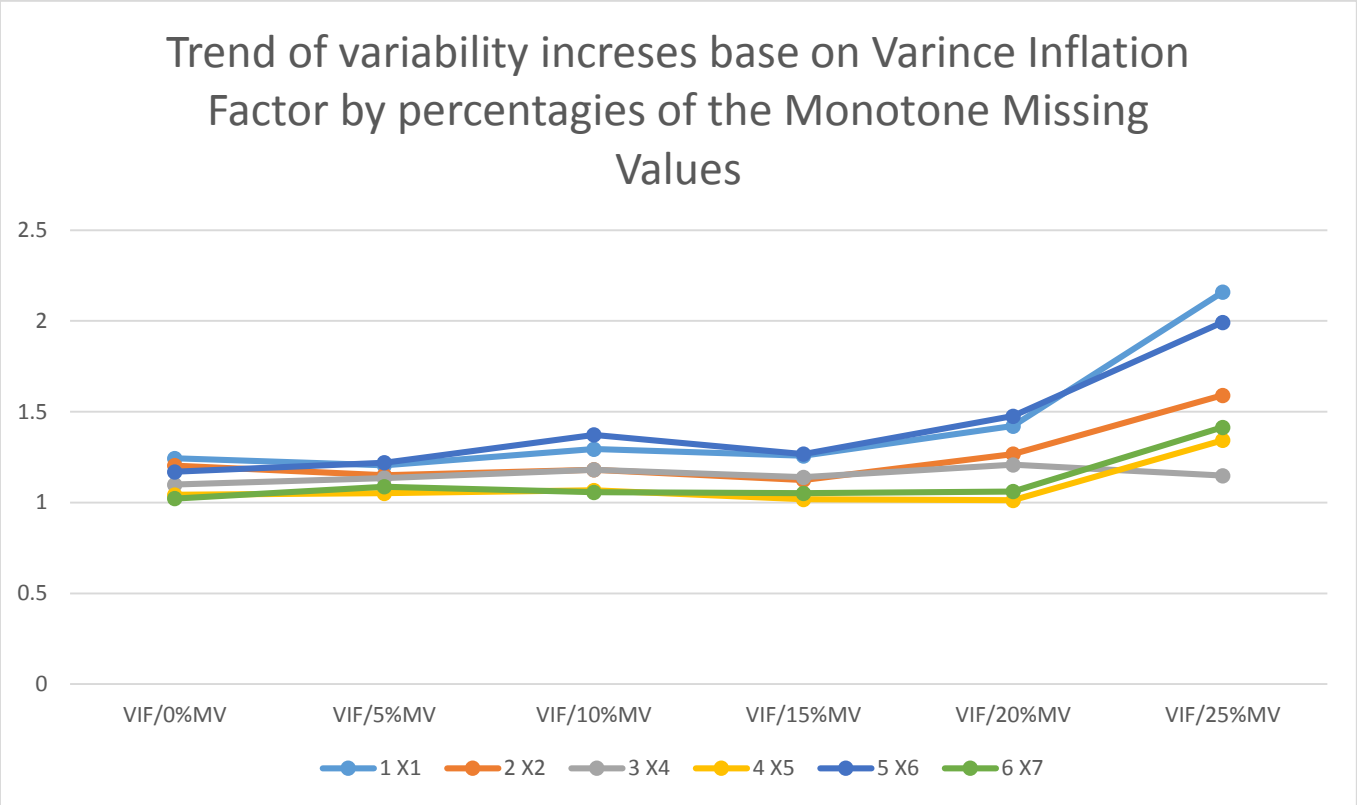
As it is showing vividly on the above figure 03. The higher the missing information the more deteriorating result on the graph.

Table 14 Variance Inflation Factor on Different Percentages by Arbitrary Type of Missing Values

S/N	X_i	VIF/0%MV	VIF/5%MV	VIF/10%MV	VIF/15%MV	VIF/20%MV	VIF/25%MV
1	X_1	1.243	1.205	1.294	1.257	1.421	2.159
2	X_2	1.204	1.151	1.181	1.124	1.267	1.59
3	X_4	1.099	1.135	1.181	1.139	1.208	1.148
4	X_5	1.042	1.051	1.067	1.018	1.012	1.343
5	X_6	1.169	1.219	1.373	1.267	1.476	1.992
6	X_7	1.022	1.087	1.056	1.051	1.061	1.414

Above table 14. It shows the number of variability changes based on the percentage increase of the missing values at random.

Fig. 04



It was observed that missing information at random has a higher effect on variance inflation factors then missing of the same information not at random even though both of them are now easy to handle due to the advancement in computation and roughens use of much statistical application. There are little difficulties in try to define the nature and potentiality of the missed information and so it is not possible to rule them out in totality, generally, there has to be an assumption by checking off proper references on other studies that were done practically. For instance an extensive follow-up in a particular survey done to investigate and ascertain the real earning of a respondent who was absent on the previous visit, this will cover up the shortcoming of the missing of information on that respondent. It is well to know that in such a survey nonresponse to the question of earning somehow depends on the characteristics such as education, race, religion, and gender and all this will not depend on the assumption that nonresponse probability is constant.

6.0 Conclusion and Discussion

Multicollinearity and missing values both have a great influence on the linear relationships always that exist in-between response and explanatory variables in the well-balanced system of the linear regression model, it is observed from the finding in this study that missing value affect the correlation and higher correlation indicate good presence of multicollinearity directly.

Both multicollinearity and missing values are always affected by the mode of the recording of the data, human error and the heterogeneity of the sample taking during a survey, therefore while dealing with such variables in many domains of the data mining and cleaning to effectively handle such a scenario a data scientist is always advised to use rebuts evaluation techniques while selecting an imputation method to take care of the missing information.

In this paper, it brings out categorically that no missing values are small no matter how it is, can change the nature of correlation, Tolerance, and variance inflation factors which finally in return will affect the linear relationships among the response and predictor variables and end up producing a severe multicollinearity. It was also established that both multicollinearity and missing values of what so ever types have a direct effect on the linear relationships between variables as such if they increase always add more values to the error of estimates statistics.

To take proper care of such short comes to bring about due to multicollinearity and missing values always imputation of the missing values is very essential by a linear combination of the existing values to predicts the Missingness rightfully. When analyzing data with missing values which is expected to have multicollinearity be it small, moderate or severe one has to study the pattern, Nature, and causes of the missing values to handle it effectively not have the double tragedy of both Collinearity and Missingness involves together it will affects the reliability of all statistics involved. Other obstacles including over and underestimation which bring about biased results shall always be taking proper care to ensure maximum accuracy and higher reliability of the estimates statistics.

It has been proved from above tables, that missing values lower the tolerance level and increase the level of multicollinearity in the system, the more data missed and more chance of having the increase in the level of the collinear relationships. Therefore it is now established that missing values courses multicollinearity and the larger the missing values and the higher is the presence of multicollinearity in the system, this is because from monotone missing value average tolerance level is 0.809 at 0% level of missingness while at 25% level of missing value the average tolerance level change to 0.84 and in the same vein from the arbitrary missing value at 0% it shows the average tolerance of 0.89 while at 25% level of missingness records the average tolerance level of 0.108. The change recoded from the average tolerances in both cases indicates how multicollinearity increases with increasing level of missing values at different percentages.

References

1. Abdelghani Hamaz and Mohamed Ibazizen (2009): Comparison of Two Estimation Methods of Missing Values Using Pitman-Closeness Criterion, Communications in Statistics - Theory and Methods, Vol. 38, Issure13, pp. 2210-2213.

2. Alhassan Umar Ahmad[1], U.V. Balakrishnan[2], Prem Shankar Jha[3], "Detection of collinearity effects on Explanatory Variables and Error Term in Multiple Regression" International Journal of Innovative Technology and Exploring Engineering (TM), IJITEE, 2019.

3. A.Kumar,PS.Rathore,A.Dubey,R,Agrawal (2020)"Low-Power Traffic Aware Emergency based Narrowband Protocol with holistic Ultra Wideband WBAN approach in biomedical application", Ad Hoc & Sensor Wireless Networks, ISSN: 1552-0633.

4. Allison, P., 2001. Missing data Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.

5. Bagya Lakshmi H., Gallo M., and Srinivasan, M.R. (2018): Comparison of regression models under Multicollinearity, Electronic Journal of Applied Statistical Analysis, Vol. 11, Issue 01, pp. 340-368.

6. Bernard J. Morzuch (1980): Principal Components and the Problem of Multicollinearity, Journal of the Northeastern Agr. Econ. Council Vol. 04, issue 1, pp. 81-84.

7. A.Dubey,A.Kumar,R.Agrawal. An efficient ACO-PSO based framework for data classification and pre-processing in big data", Evolutionary Intelligence Springer Electronic, ISSN 1864-5917.

8. Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., (2003). Missing presumed at random: cost-analysis of incomplete data. Health Economics Vol. 12, pp. 377–392.

9. Cismondi, F., Fialho A.S., Vieira S.M., Reti S.R., Sousa J.M.C., and Finkelstein S.N. (2013): Missing Data in Medical Databases: Impute, Delete, or Classify? Artif Intell Med Vol. 58, Issue 1, pp. 63-72.

10. Dempster, A.P., Laird, N.M., and Rubin D.B. (1997): Maximum likelihood from incomplete data via the EM algorithm. JRSSB, Vol. 39, issue 1, pp. 1-38.

11. Donald, E. Farrar, and Robert, R. Glauber (1967): Multicollinearity in Regression Analysis: The Problem Revisited, The MIT Press Stable. The Review of Economics and Statistics, Vol. 49, No. 1, pp. 92-107. Accessed: 13/09/2013 16:08.

12. Vishal Dutt, Sriramakrishnan Chandrasekaran, Vicente García-Díaz, (2020). "Quantum neural networks for disease treatment identification.", European Journal of Molecular & Clinical Medicine, 7(11), 57-67
13. file:///C:/Users/user/Desktop/Missing%20value/missing%20valu1.pdf: (Accessed 07-10-2019).
14. Graham, J.W. (2009): Missing data analysis: making it work in the real world. Annual Rev. Psychol, Vol. 60, pp. 549-576.
15. Graham, J.W. (2009): Missing Data Analysis: Making It Work in the Real World. Annu Rev Psychol; Vol. 60, Pp. 549-576.
16. Graham, J.W., (2009): Missing data analysis: making it work in the real world. Annu Rev Psychol. pp. 549-576.
17. Graham, J.W., (2009): Missing data analysis: making it work in the real world. Annu Rev Psychol. Vol. 60, pp. 549–576.
18. A.Kumar,T. Sairam And V.Dutt "Machine Learning Implementation For Smart Health Records: A Digital Carry Card", published in "Global Journal on Innovation, Opportunities and Challenges in Applied Artificial Intelligence and Machine Learning Vol. 3, Issue 1 – 2019.
19. Hyun Kang (2013): The Prevention and Handling of the Missing Data, Korean J Anesthesiol, Vol. 64, Issue 5, pp. 402-406.
20. Jamal, I. Daoud (2017): Multicollinearity and Regression Analysis, Journal of Physics, Conference Series 949. Department of Science and Engineering, IIUM, 53100, Jalan Gombak, Selangor Darul Ehsan, Malaysia.
21. John O'Hagan and Brendan McCab (1975): Tests for the Severity of Multicollinearity in Regression Analysis, the Review of Economics and Statistics, Vol. 57, No. 3 pp. 368-370.
22. S. Boyapati, S. R. Swarna, V. Dutt and N. Vyas, "Big Data Approach for Medical Data Classification: A Review Study," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 762-766, doi: 10.1109/ICISS49785.2020.9315870.
23. Marina, Soley-Bori (2013): Dealing with missing data: Key assumptions and methods for applied analysis; Technical Report No.4 pp. 1-20.
24. Nakai M and Weiming Key (2011): Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. Vol. 5, issue no.1, pp. 1-13.
25. O'Neill, R.T., and Temple R. (2012): The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Theory, Vol. 91, Issue 34, pp. 550-574.
26. S. M. Sasubilli, A. Kumar and V. Dutt, "Machine Learning Implementation on Medical Domain to Identify Disease Insights using TMS," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, 2020, pp. 1-4, doi: 10.1109/ICACCE49060.2020.9154960.
27. Paul T. von Hippel (2004): Biases in SPSS 12.0 Missing Value Analysis, The American Statistician, Vol. 58, Issue 2, pp.160-164.
28. Peng, C.Y., Harwell M.R., Liou, S.M., and Ehman, L.H. (2006): Advances in missing data methods and implications for educational research.
29. Peugh, J.L and Enders, C.K. (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res Vol.74, Issue 14, pp. 525-556.
30. S.Chandrasekaran and A.Kumar Implementing Medical Data Processing with Ann with Hybrid Approach of Implementation Journal of Advanced Research in Dynamical and Control Systems – JARDCS issue 10, vol.10, page 45-52, ISSN-1943-023X. 2018/09/15.
31. Rubin DB. (1996): Multiple imputations after 18+ years (with discussion). J Am Stat Assoc; 91: 473-89.
17. Acock AC. Working with missing values. J Marriage Fam 2005; 67: 1012-28.
32. Rubin, D.B. (1976): Inference and missing data. Biometrical, Vol. 63, Insurer 13, pp. 581-592.
33. SAS Institute, (2005): Multiple Imputation for Missing Data: Concepts and New Approaches. (Accessed 08/10/2019).
34. Schafer, J. L. (1997): Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
35. Schafer, J. L., 1997. Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
36. Sinharay, S., Stern, H.S, and Russell D. (2001): The use of multiple imputations for the analysis of missing data. Psychol Methods, Vol. 6, pp. 317-329.
37. Steffi Pauli Susanti and Fazat Nur Azizah (2017): Imputation of Missing Value Using Dynamic Bayesian Network for Multivariate Time Series Data, International Conference on Data and Software Engineering, IEEE, Vol. 978, Issure1, pp. 1449-1455.
38. Young, W., Weckman, G., and Holland, W. (2011): A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. Theory Issues Ergon Sci. Vol.12, Issure1, pp.15-43.

39. Swarn Avinash Kumar, Harsh Kumar, Srinivasa Rao Swarna, Vishal Dutt, "Early Diagnosis and Prediction of Recurrent Cancer Occurrence in a Patient Using Machine Learning", European Journal of Molecular & Clinical Medicine, 2020, Volume 7, Issue 7, Pages 6785-6794
40. Yuan Yang C., (2011): Multiple imputations for Missing Data: Concepts and New Development (SAS Version 9.0). SAS Institute Inc., Rockville, MA).
41. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2013): Missing data in medical databases: impute, delete, or classify? Artif Intell Med 58(1):63–72.
42. Peng CY, Harwell MR, Liou SM, Ehman LH (2006) Advances in missing data methods and implications for educational research.
43. Peugh JL, Enders CK (2004): Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res 74(4):525–556.
44. Young W, Weckman G, Holland W (2011): A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. Theory Issues Ergon Sci 12(1):15–43.
45. Catia M. Salgado, Carlos Azevedo, Hugo Proenza, and Susana M. Vieira (2016): missing data, ResearchGate, MIT Critical Data, Secondary Analysis of Electronic Health Records, pp.143-174.